Cognitive Computation
https://doi.org/10.1007/s12559-020-09800-x

Distant Supervised Construction and Evaluation of a Novel Dataset of Emotion-Tagged Social Media Comments in Spanish

Juan Pablo Tessore[1,2] · Leonardo Martín Esnaola[1] · Laura Lanzarini[3] · Sandra Baldassarri[4,5]

Received: 17 June 2020 / Accepted: 30 November 2020

Abstract

Tagged language resources are an essential requirement for developing machine-learning text-based classifiers. However, manual tagging is extremely time consuming and the resulting datasets are rather small, containing only a few thousand samples. Basic emotion datasets are particularly difficult to classify manually because categorization is prone to subjectivity, and thus, redundant classification is required to validate the assigned tag. Even though, in recent years, the amount of emotion-tagged text datasets in Spanish has been growing, it cannot be compared with the number, size, and quality of the datasets in English. Quality is a particularly concerning issue, as not many datasets in Spanish included a validation step in the construction process. In this article, a dataset of social media comments in Spanish is compiled, selected, filtered, and presented. A sample of the dataset is reclassified by a group of psychologists and validated using the Fleiss Kappa interrater agreement measure. Error analysis is performed by using the Sentic Computing tool BabelSenticNet. Results indicate that the agreement between the human raters and the automatically acquired tag is moderate, similar to other manually tagged datasets, with the advantages that the presented dataset contains several hundreds of thousands of tagged comments and it does not require extensive manual tagging. The agreement measured between human raters is very similar to the one between human raters and the original tag. Every measure presented is in the moderate agreement zone and, as such, suitable for training classification algorithms in sentiment analysis field.

Keywords Sentiment analysis · Dataset construction · Dataset validation · Facebook · Text mining

Introduction

Understanding emotions is key for human intelligence emulation and, thus, for the advancement of artificial intelligence. In addition, the opportunity to capture sentiments has gained interest in both the scientific community and the business world, which has led to the emerging fields of affective computing and sentiment analysis [1].

Affective computing [2] is a field of cognitive computing and artificial intelligence, whose objective is to develop systems that are able to recognize, interpret, process, and simulate human emotions. Sentiment analysis (SA) is a suitcase research problem that requires tackling many natural language processing (NLP) tasks [3]. It contains three layers. The first one is a syntactic layer that aims at pre-processing texts and includes tasks such as part-of-speech tagging, lemmatization, and micro text normalization. The second one is a semantic layer that aims at deconstructing the normalized text from the previous layer into concepts, resolve entities, and filter neutral content to improve sentiment classification

✉ Juan Pablo Tessore
juanpablo.tessore@itt.unnoba.edu.ar

Leonardo Martín Esnaola
leonardo.esnaola@itt.unnoba.edu.ar

Laura Lanzarini
laural@lidi.info.unlp.edu.ar

Sandra Baldassarri
sandra@unizar.es

[1] Instituto de Investigación y Transferencia en Tecnología (ITT), (Centro CICPBA), Universidad Nacional del Noroeste de Buenos Aires, Junín, Buenos Aires, Argentina

[2] Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Ciudad Autónoma de Buenos Aires, Argentina

[3] Facultad de Informática, Instituto de Investigación en Informática LIDI (Centro CICPBA), Universidad Nacional de La Plata, La Plata, Buenos Aires, Argentina

[4] Departamento de Informática e Ingeniería de Sistemas, Universidad de Zaragoza, Aragon, Zaragoza, España

[5] Instituto de Investigación en Ingeniería (I3A), Universidad de Zaragoza, Zaragoza, Aragon, España

Published online: 18 January 2021

accuracy. The tasks in this layer are, among others, concept extraction, word sense disambiguation, and subjectivity detection [4]. The last one is the pragmatics layer, focused on extracting meaning from both sentence structure and semantics obtained from previous layers, and it includes tasks such as polarity detection, aspect recognition, sarcasm detection [5], and personality recognition [6]. Medhat et al. proposed another definition for SA [7], stating that SA can be considered as a classification process with three primary classification levels: document level, sentence level, and aspect level, in which the goal is to detect when expressing a positive or negative opinion or sentiment, generally known as polarity detection. Cambria et al. [8] introduced the concept of *Sentic Computing* as a multi-disciplinary approach to SA, in which both computer and social sciences are combined to better recognize, interpret, and process opinions and sentiments on the Web.

According to Medhat et al. [7], datasets used in SA represent an essential issue. In that study, the authors also point out that the main sources of data are from product reviews; however, SA is applied in other domains as well. The purpose of applying SA techniques may vary depending on the end-user. For example, companies are interested in better understanding their customers and competitors to improve their market share [9]. Buyers, on the other hand, would like to make better purchasing decisions by taking advantage of the opinions of other buyers [10]. Politicians usually are very interested in knowing the public opinion in a timely and accurate manner, enabling better decision-making [11]. Even in medicine, there are SA applications, for example, to identify mentions of personal intake of medicines in tweets [12].

With the advent of Web 2.0, the availability of data sources has increased considerably. Through social media, millions of users worldwide interact with others, sharing their comments and experiences. Moreover, some social media sites also allow users to input additional information along with the text, such as emoticons, thumbs up/down, scores, categories, or some raw emotions.

As stated by Wang et al. [13], there are a lot of social media tools for carrying out sentiment analysis, but they are focus on finding the aggregate-level sentiment, such as sentiment polarity. Nevertheless, authors propose that, if finer-grained sentiment analysis can be achieved, it will yield more specific and more actionable results with detailed negative emotion subcategories such as anger, sadness, and anxiety or positive emotion subcategories such as happiness and excitement.

The terms *sentiment* and *emotion* are widely used but usually confused or misinterpreted [14] and have often been used interchangeably; however, sentiments are differentiated from emotions by the duration in which they are experienced. Wang et al. [15] stated that, while sentiments reflect feelings and attitudes, emotions provide a more refined characterization of the sentiments involved. Emotion sensing drills deeper to reveal the exact emotions expressed in the text. In their study, the authors hold that whatever emotion-sensing methodology is used, having a proper categorization model for emotions is always very important. In this sense, the study reviews many of the existing emotion models by considering the view of psychologists, as well as perspectives from social science, computing science, and engineering. The different models surveyed in the study vary in the number of emotions they recognize—some consist of six primary emotions, while others identify up to 24. Additionally, models can be divided between categorical and dimensional.

Among categorical models, Ekman's model of emotions [16] stands out. This model is based on the argument that there are six distinctive facial expressions (plus neutral): anger, fear, disgust, joy, sadness, and surprise. On the other hand, two-, three-, and four-dimensional models can be identified. Two-dimensional models are characterized by valence/arousal. Three-dimensional models incorporate an additional dimension, which varies according to the model in question. Lastly, there are several four-dimensional models, such as *The Hourglass of Emotions* model [17], which considers sensitivity, aptitude, pleasantness, and attention as dimensions. This model is an affective categorization model, primarily inspired by Plutchik's studies on human emotions, and is a biologically inspired as well as psychologically motivated emotion categorization model.

For the present paper, Ekman categorical model [16] is used, since it is one of the most widely adopted models for affect recognition [17], and because Ekman's basic emotions are somehow related to the Facebook reactions (LOVE, SAD, ANGRY, WOW, and HAHA).

In this work, a comment and a reaction produced by the same user in response to a given post are linked, because the authors assume that a topic may trigger, but usually not express, an emotion, whereas a comment usually conveys the emotion felt by the reader of that topic. Even though, in this case, the link between some of the reactions and Ekman's basic emotions could seem straightforward, it should be noted that the tagging process is not performed in a controlled environment, and the people that tagged the content is not trained for this specific task. In addition, it is presumed that there may be a significant level of noise on the comment-tag association, produced mainly by the presence of trolls, interaction between users, and the edition or deletion of comments and reactions.

The usefulness of basic emotion datasets depends on the reliability of the emotions assigned to the content. The ultimate goal of the users of this kind of datasets is to predict basic emotions, not Facebook reactions. In this regard, it is necessary to establish the strength of the link between those reactions and basic emotions. The use of the reactions in this manner could be seen as a form of distant supervision (DS) [18], in which data are tagged automatically or semi-automatically, using some safe signals already present as proxies. This approach allows building a larger dataset by eliminating the need for extensive manual tagging. Some other studies [19–27] have already used Facebook reactions but, unlike this work, they linked the reaction to the topic from which it stemmed.

Furthermore, social network data are usually noisy. They contain many issues, such as casual language, spelling errors, and troll activity. The latter is particularly damaging for the construction of basic-emotion datasets, because trolls usually post and repeat their comments and reactions regardless of the topic discussed, and they usually interact ironically or provocatively. Because of these known issues, there is a need to establish the quality of the datasets constructed from social network data. One way to achieve this is to measure the agreement among raters, regarding the reactions, for a small sample of the dataset. In this study, Fleiss kappa [28] will be used, as it is one of the most widely adopted interrater agreement measures.

This work is focused on the Spanish language because, to the best of the authors' knowledge, there are no studies that build and measure the quality of a distantly supervised tagged dataset in this language by comparing it with full manual tagging. The goal is to provide a valuable dataset that can be used in future studies.

In summary, in this paper, a SA issue, which is the generation of emotion-tagged datasets for the Spanish language, is addressed. The dataset presented is built by applying DS on Facebook comments and reactions, and it is validated using the Fleiss kappa interrater agreement measure and the *Sentic Computing* tool BabelSenticNet.

The remainder of this paper is organized as follows. "Related Work" reviews the literature on DS, Spanish datasets, and interrater agreement measurement. "Dataset Compilation and Filtering Process" presents the dataset along with the compilation and filtering process. "Experimental Setup" describes the validation process performed. In "Results and Discussion", the results are shown. Lastly, "Conclusions and Future Work" are discussed.

Related Work

Tagged datasets are a key ingredient for developing machine-learning text-based classifiers. Mercado et al. [29] stated that in any automatic text analysis, it is essential that there are adequate datasets available so that the data mining and machine-learning approaches can obtain reliable and informative results. Moreover, according to Lo et al. [30], most of the effort has been made in creating resources for formal languages, used in official communication, while, with the popularity of social media, informal linguistic variants are becoming widespread, and those variants require different considerations for their analysis. The mentioned study focuses on multilingual SA; out of all the approaches, lexicons, tools, and corpora listed, only a few focus on variants of the Spanish language. One exception to the above is SenticNet [31], a concept level knowledge base for sentiment analysis that supports 40 languages, including Spanish, using the tool BabelSenticNet [32].

Nevertheless, most resources available are for the English language—Justo et al. [33] stated that the majority of research in disciplines like SA addresses English, even though 48% of Internet resources are written in other languages. This results in the need for creating resources in other languages as well [30, 34]. However, as mentioned before, manual tagging is one of the most time-consuming tasks in the creation of emotional datasets. To overcome this issue, many studies have constructed datasets by using DS [18]. In this model, an already existing noisy label is linked to the content to build a tagged dataset automatically.

According to Roth et al. [35], DS allows creating large amounts of training data at a low cost. As the data obtained are inherently noisy, the most challenging problem is improving their quality by reducing the amount of noise.

DS was applied in the work of Go et al. [36], in which emoticons were used as labels to automatically classify a dataset of tweets into one of three categories, which were positive, negative, and neutral. The latter was discarded, and then Support Vector Machines (SVM), Maximum Entropy (ME), and Naive Bayes (NB) classifiers were trained and tested only with positive and negative tweets. The best accuracy reported by this study, 82.7%, was achieved using a combination of unigram and bigram features with ME and NB.

Bandhakavi et al. [37] used labeled (blogs, news headlines) and weakly labeled (tweets) emotion text to generate an emotion lexicon that jointly modeled both the emotionality and neutrality of documents at word level. Pool and Nissim [19] used Facebook reactions in a DS fashion to train an SVM for emotion detection. Nevertheless, they linked reactions to the original post, which is the most widely adopted association in studies that use Facebook reactions [20–27], rather than associating them to the comment, which is proposed in the present paper. In addition, they did not measure the reliability of the automatic tags.

DS is also beneficial when working with low resource languages, as presented in the work of Refaee [38], in which several experiments with distantly supervised datasets in Arabic were conducted. The author concluded that, for subjectivity classification, DS (emoticon and lexicon-based) outperforms fully supervised methods in this language. However, for sentiment classification (positive vs negative), dataset size had to be expanded and hashtags should also be used as labels. The author also states that the results of DS can be language-dependent and that this type of experiments should be conducted for every language.

In the work of Suttles et al. [39], a dataset of tweets was collected, and then, hashtags, emoticons, and emojis in the tweets were used to tag the dataset in a DS fashion. Having multiple ways to tag the dataset allowed performing cross-validation of the tagging process using the χ^2 goodness of fit test. The authors concluded that, with minor exceptions, there was consensus between the tags. The tagged dataset was then used to train machine-learning classifiers that obtained accuracies between 75 and 91% (tested with manually tagged tweets).

Felbo et al. [40] collected a very large dataset of tweets and used the emoticons as noisy labels to train a Long Short-Term Memory (LSTM) model. Emojis were stripped from the text, and the model was trained to learn which emoji was removed.

Since the dataset compiled and analyzed in the present work is in Spanish, literature was reviewed for articles that work with this specific language using DS for automatic basic emotion tagging. However, to the best of the authors' knowledge, most papers use polar tags (or a similar variant including neutral and several degrees of positive and negative). In this way, in the research of Moctezuma et al. [41], a dataset of 18 million tweets in Spanish was classified into positive and negative using Spanish affective lexicons. Martín et al. [42] mapped the rating attached to the comments of a touristic website also into a polar classification. Sandoval-Almazan et al. [27] measured the impact of Facebook posts in political campaigns by collecting Spanish posts together with some statistics, like the number of comments, shares, and reactions. However, in that work, the analysis was carried out only based on the emoticons included in the text of the comments, without analyzing the emotion that the text itself might reflect.

As mentioned, most papers that carry out sentiment classification in Spanish rely on datasets built with manual tagging. Such is the case for most datasets provided or presented at the workshop *Taller de Análisis de Sentimientos en Español* (TASS) [43], and also for the dataset used at the IberLef 2019 competition [44] that compiled tweets with different variants of Spanish. In total, the latter dataset contains around 15,000 tweets in five Spanish variants. Even though the amount of samples is significant, as it was classified manually, it surely took a lot of time and resources to compile. Datasets of this size have been easily compiled using DS in other languages [36, 40, 45–49].

An important aspect when working with DS is the validation of the data sources. As mentioned before, since the requirement of manual tagging is removed, datasets that rely on noisy labels tend to be particularly large. This considerable size may be a problem when validating the data. To solve this, two main strategies are usually adopted.

The first strategy is the χ^2 goodness-of-fit test. To implement this test, at least two different kinds of labels are used to cross-validate data. As mentioned in [39], this has been done using emojis, emoticons, and hashtags. The agreement among the tags was quite above chance for the majority of the classes. That was also the case in [38] for validating sentiment labels obtained with different approaches.

The second strategy is the Fleiss kappa interrater agreement measure [28]. This is useful for measuring agreement among a fixed number of raters over categorical data. The formula to calculate this measure can be seen in equation (1), where $\overline{P}$ is the probability of agreement among raters, and $\overline{P_e}$ is the probability of agreement by chance.

$$\kappa = \frac{\overline{P} - \overline{P_\mathrm{e}}}{1 - \overline{P_\mathrm{e}}} \tag{1}$$

The value of κ moves between − 1 (perfect disagreement) and 1 (perfect agreement). Carletta [50] established $\kappa > 0.80$ as a good reliability, with $0.67 < \kappa < 0.80$ allowing tentative conclusions to be drawn. However, the author also hints that discourse and dialog phenomena may be more complicated than other types of analysis (such as subject classification on newspaper articles). Hearst [51] suggested that this hint implies that the reliability required for this kind of studies may be justified on being lower. Moreover, it should also be noted that these conclusions were drawn before the era of social networks, and subsequent studies were more permissive with reliability requirements.

Cohen's kappa metric [52], a predecessor to Fleiss kappa metric but limited to two raters, was used in [38] to measure the agreement in the annotations of two newly collected tweet datasets in Arabic. The average κ was 0.786, indicating substantial agreement. The classes used were positive, negative, neutral, mixed, uncertain, and skip. The last two classes, while useful, tend to improve agreement measure results, as the most challenging content usually falls into them.

In the work of Gambino and Calvo [53], a dataset of 3,572 twitter messages in Spanish was compiled. Then, each tweet of the dataset was classified into one of six basic emotions (love, joy, surprise, anger, sadness, and fear) by four annotators. After the annotation process was completed, the resulting agreement measure was 0.49, indicating moderate agreement.

In SemEval 2019 [54], a dataset of textual dialogs was built for one of the tasks. It consisted of 38,424 dialogs that were manually tagged into four different classes (angry, happy, sad, and others) by seven human raters each. The Fleiss kappa score obtained for the tagged data was 0.59, also indicating moderate agreement. In this case, the class “others” may have helped to improve the final agreement score.

As seen, although Carletta [50] established challenging agreement requirements, most researchers carrying out emotion classification on social networks textual data consider that a moderate agreement in the Fleiss kappa scale is acceptable. This will also be the case in the present study. In addition, although many studies measure the quality of the manual tagging using the Fleiss kappa metric, no studies, in the Spanish language at least, compare the reliability of the datasets tagged using DS versus the ones manually tagged.

Dataset Compilation and Filtering Process

Comments and reactions were collected from Facebook, since it is one of the most widely used social networks, with more than 2,449 million active users worldwide as of January 2020 [55].

Those comments and reactions were taken from the interactions of many different Facebook users that posted across 13 widely read news portals in Argentina, namely, *Clarín*, *La Nación*, *Página 12*, *El Cronista*, *Ámbito Financiero*, *Todo Noticias*, *Crónica*, *CNN en Español*, *C5N*, *Agencia Télam*, *Diario Deportivo Olé*, *Teleshow*, and *Infobae*. These news portals cover different types of news, and they were chosen due to the variety of topics they cover and because they are among the most widely consumed in Argentina [56]. Each comment-reaction tuple reflects the user interaction with a particular post, by reacting to the post and writing a comment.

Comments posted during a period of 4 years were compiled. This aspect is addressed in “Comment Compilation”. However, not all of the collected comments ended up in the final dataset—they went through a selection and filtering process.

Figure 1 shows the entire process, from compilation to the final filtering of comments considered as useful. Each step of the process, namely comment compilation, comment tokenization, filtering by token count, filtering by language, and troll filtering, is explained in detail in the following sections. The final structure of the dataset is described in “Dataset Description”.

Comment Compilation

The extraction process of comments, reactions, and posts was performed using the Facebook API Graph tool [57]. This tool allows setting the interval for data retrieval, so extraction dates were set to 1st January 2016 until the end of December 2019, i.e., a 4-year extraction period. Then, the results were stored in a relational database.

Using a database, the selection process was more straightforward, since not all of the collected comments were valid. All comments and reactions in posts made by the news portals mentioned above were collected, but not all of those are significant to this study. Comments associated with the reaction LIKE were deliberately excluded from the study, since people generally use it to indicate that they saw that post. Moreover, each Facebook user can interact multiple times with a particular post by writing more than one comment. Therefore, the first comment—the older one—is considered as the purest or most significant for the expressed emotion.

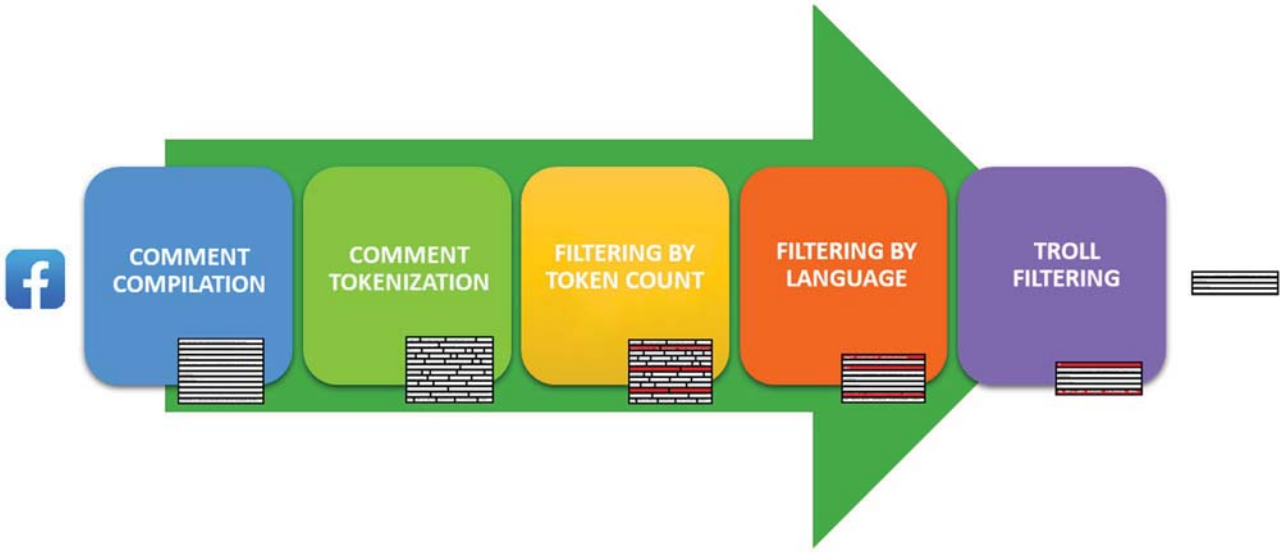


Fig. 1 Compilation and filtering process

Commonly, a user expresses a reaction and leaves a comment, and then continues to write new comments in response to other users' interactions in the same post.

Considering all collected comments, before selecting only those that are useful, the number of compiled comments was 20,996,169. However, after considering only those related to the emotions or reactions LOVE, HAHA, ANGRY, and SAD, and then selecting only the oldest contribution from each user (in the case of users posting more than one comment in the same post), the number of comments dropped to 1,716,413.

Comment Tokenization

Each comment consists of tokens, and each token is a sequence of characters. Special characters like white spaces and line breaks separate one token form another. Because of this, a token can be a well-formed word, an emoji, a link, or another kind of composition. Therefore, the first step is the tokenization of the comments. The TweetTokenizer class, from the NLTK library [58], was used for this step. For the rest of the filtering process, links, signs, non-printable characters, and Spanish stop words were not considered as a token when counting them on each comment. For example, consider the following comment: "*Señor Olé usted es diabolico*." Here, TweetTokenizer obtains six tokens: "*Señor*", "*Olé*", "*usted*", "*es*", "*diabolico*", and ".", but only four of those tokens are considered valid because the token "." is a punctuation mark and the token "*es*" is a Spanish stop word. Consequently, the comment in the example has only four valid tokens.

After tokenization, comments containing zero tokens were excluded as well by excluding links, signs, non-printable characters, and Spanish stop words. After this, the number of useful comments dropped to 1,674,912.

Filtering by Token Count

Not all collected comments are significant and can be linked to an emotion. For example, comments with only two valid tokens or less, in general, are just the name and surname of another Facebook user (when a user tags a friend, for instance). For this reason, a first filter was applied in order to remove all comments with only two valid tokens or less. For example, this filter removes comments like "*Mirá* http://www.eldestapeweb.com/le-entregaron-un-segundo-negocio-la-empresa-del-jefe-del-pami-n37687*", since it has only one valid token ("Mirá*"); or "*Claudia Cocco*", which consists of only two valid tokens, and is an example of a tagged user.

Consequently, comments tokenized as described in "Comment Tokenization", containing less than three tokens, were excluded from the dataset. After applying this filter, the number of useful comments dropped to 1,261,783.

Filtering by Language

Even though almost all users that interact with the selected news portals comment in Spanish, there are a few comments in other languages. Thus, another filter was applied in order to remove non-Spanish comments. For this process, "Python Bindings to CLD2" library [59], or simply CLD2, was used. The CLD2 language detection process was applied to all remaining comments after the application of the previously described filters, yielding a total of 1,035,045 comments that were written in Spanish.

Since language detection is a complex process that can present false positives, an extra validation step was made using Googletrans [60]. Since the use of Google Translate API is not free, and the number of daily requests for free is limited, a relatively small sample of 1,400 comments, randomly selected from the previous set of Spanish comments, was taken to perform a cross-validation process. All the comments analyzed with Googletrans were recognized as written in Spanish, which is an additional element to trust the results obtained with the CLD2 library.

Troll Filtering

Trolling is an interpersonal antisocial behavior prominent within Internet culture across the world, and Facebook, with more than two billion active users worldwide, has become the Internet's biggest playground for engaging in antisocial behaviors, mainly trolling. Trolling behavior includes starting aggressive comments and posting inflammatory, malicious messages in online comment sections to deliberately provoke, disrupt, and upset others [61].

Those comments and interactions are undesired for this study, since they do not necessarily reflect an emotion or a reaction to a particular topic. Trolls write comments in several posts, frequently the same comment, independently of the topic of the post. Hence, the troll filtering process consists in identifying all the comments that could potentially have been posted by trolls and exclude them from the dataset.

The process was made by first identifying all the comments that appear more than once, and then counting the number of appearances. The process revealed that there were 14,488 comments in the dataset that appeared at least twice. If the entirety of comments collected initially is considered (20,996,169), this number goes up to 237,309 repeated comments. These comments, which represent about 1.399% of the dataset, were excluded.

Table 1 Token and character level statistics for titles, subtitles, and comments

Level	Title			Subtitle			Comment		
	Max	Min	Avg	Max	Min	Avg	Max	Min	Avg
Token	41	0	12.65	388	0	22.57	1,218	3	19.36
Character	246	1	71.52	2,193	1	135.85	7,587	7	110.93

Below, there are some examples of contributions identified as troll comments, indicating the number of times they appear and the number of different posts in which they were made:

Comment "*por fin una buena noticia*", appears 102 times in 92 different posts.
Comment "PENSAMIENTO: TRUMP Y LOS REPUBLICANOS QUIEREN MANEJAR ESTE PAIS COMO MADURO, ORTEGA, CORREA Y CASTROS: TIENEN EL PODER EJECUTIVO Y EL JUDICIAL CON LA NOMINACION DE KAVANAUGH Y QUIEREN TENER EL TERCER PODER EL LEGISLATIVO (CONGRESO Y SENADO) SE IMAGINAN POR ESO ES IMPORTANTE SALIR A VOTAR E IMPEDIR ESA DICTADURA VOTEMOS DEMOCRATA Y ASI EVITAR ESE DICTADOR", appears 142 times in 142 different posts.
Comment "*PEOR ES SER DE riBer*", appears 171 times in 171 different posts.

After applying this filter, the number of useful comments dropped to 1,020,557.

Dataset Description

After the selection and filtering process, the dataset, included as Electronic supplementary material 1, was built. Its main characteristics are described below:

- Filename: "facebook_automatically_tagged_dataset.csv"
- Title: "Compilation of comments and reactions made by users to some Facebook public posts"
- Extracted from: public domain posts on Facebook
- Number of instances: 1,020,557
- Number of attributes per instance: 4 plus the class attribute
- Attribute information:

1. Sample code number (type: numeric)
2. Post title in Facebook (type: UTF8 encoded text)
3. Post subtitle in Facebook (type: UTF8 encoded text)
4. User comment to the post (type: UTF8 encoded text)
5. Class, which is the reaction of the user to the post (HAHA, LOVE, ANGRY, SAD)

- Missing attribute values: 0
- Class distribution: HAHA: 338,835 (33.20%), LOVE: 159,830 (15.66%), ANGRY 436,357 (42.75%), SAD: 85,535 (8.38%).

Dataset Statistics

Table 1 shows the maximum, minimum, and average token and character counts in post titles, subtitles, and comments. Table 2 shows the same information, but segmented by reaction.

Dataset Title, Subtitle and Comment Frequency

Other important information to consider about the dataset is the frequency of the instances in terms of the number of tokens and characters in them. Assuming they

Table 2 Token and character level statistics for titles, subtitles, and comments, segmented by reaction

Reaction	Level	Title			Subtitle			Comment		
		Max	Min	Avg	Max	Min	Avg	Max	Min	Avg
ANGRY	Character	246	3	73.12	2,193	1	137.26	7,505	7	119.03
	Token	41	1	12.92	388	0	22.84	1,185	3	20.86
HAHA	Character	241	1	70.36	2,193	1	134.32	7,587	10	100.39
	Token	35	0	12.38	388	0	22.14	1,218	3	17.49
LOVE	Character	231	3	69.30	2,193	2	133.26	7,347	9	103.01
	Token	41	1	12.23	388	0	22.12	1,195	3	17.69
SAD	Character	241	3	72.11	2,193	4	139.57	5,930	12	126.11
	Token	41	1	13.14	388	0	23.66	938	3	22.25

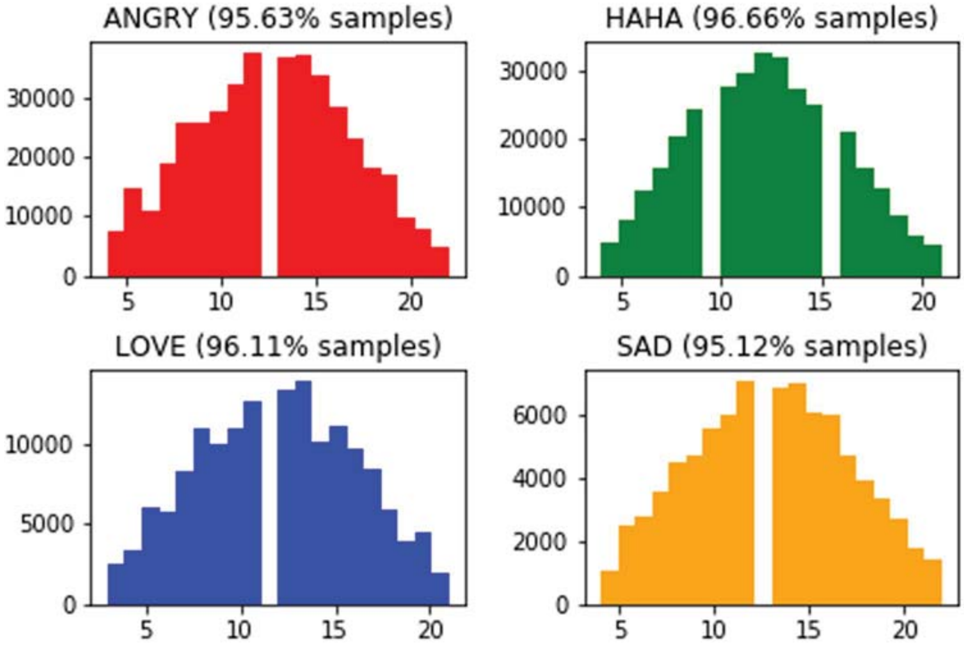


Fig. 2 Token level, post titles histogram

approximately follow a normal distribution, and by using the empirical rule, all instances around the mean value with a width of two standard deviations were considered for producing more comprehensible histograms. Figures 2, 3, and 4 show the frequency of instances in terms of tokens for titles, subtitles, and comments, respectively. Figures 5, 6, and 7, show character level instead.

Vocabulary Overlapping Level in Comments from Different Classes

Other relevant information about this dataset is how much overlap is among tokens from the different classes, i.e., the reactions related to the comments. Table 3 shows the overlapping level considering unique tokens for each reaction. For example, this table shows that 27% unique tokens of

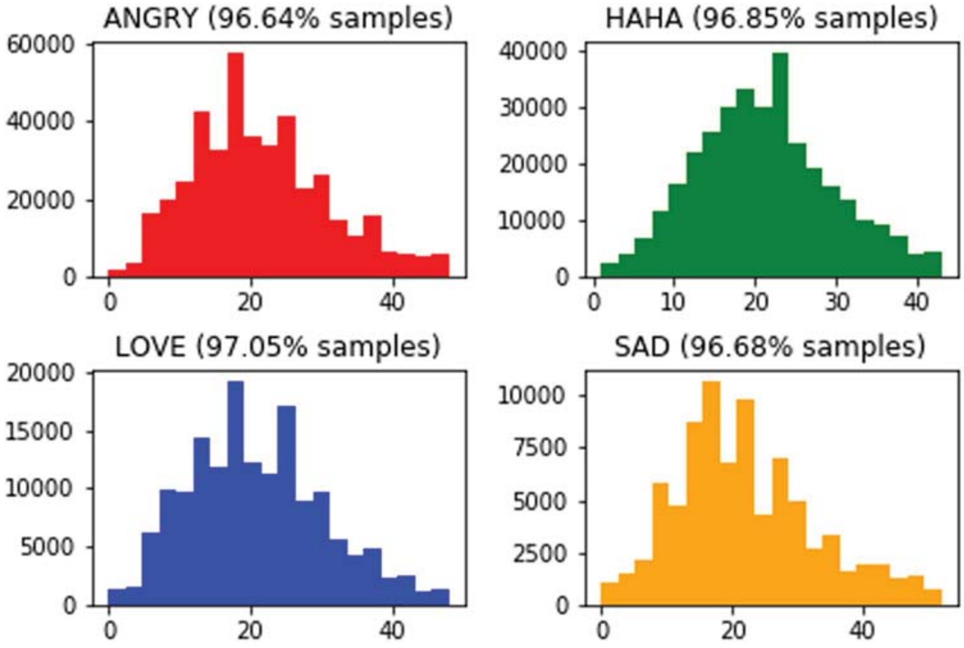


Fig. 3 Token level, post subtitles histogram

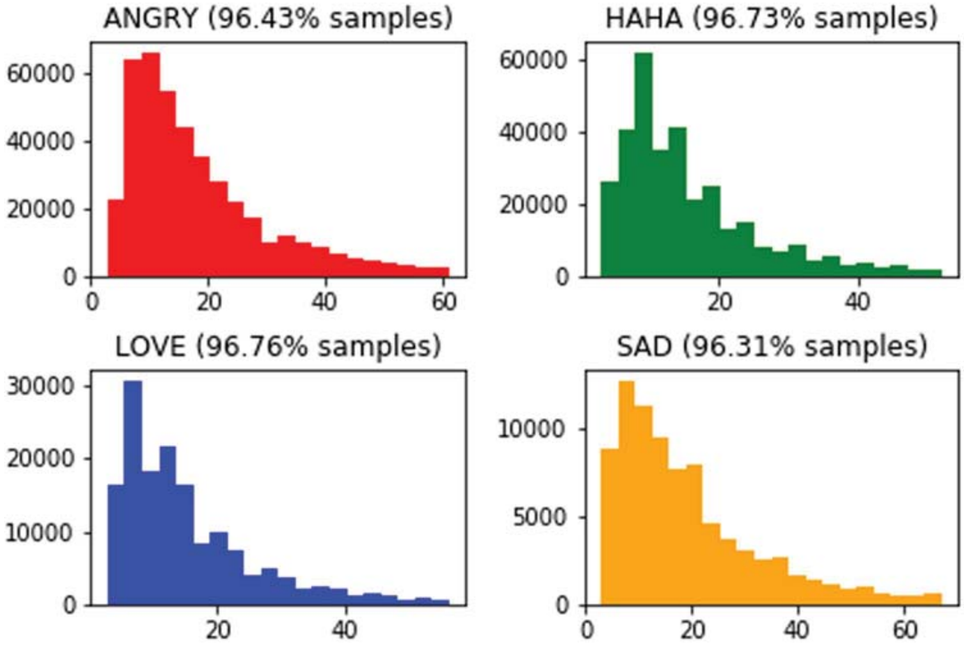


Fig. 4 Token level, post comments histogram

comments linked to HAHA reaction are contained in the unique tokens of comments related to the SAD reaction and that, inversely, 62% of the unique tokens of comments that correspond to the SAD reaction are present in the unique tokens of comments related to the HAHA reaction.

Most Common Unique Comment Tokens in Each Class

For the last metric extracted from the dataset, the frequency of each unique term in each reaction was identified, i.e., those terms that are linked only to a specific

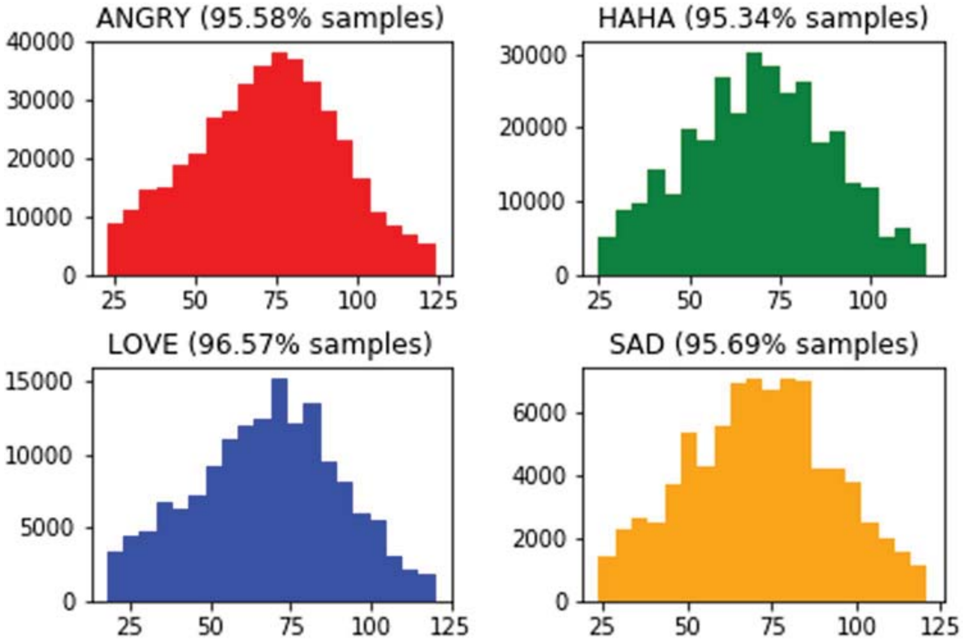
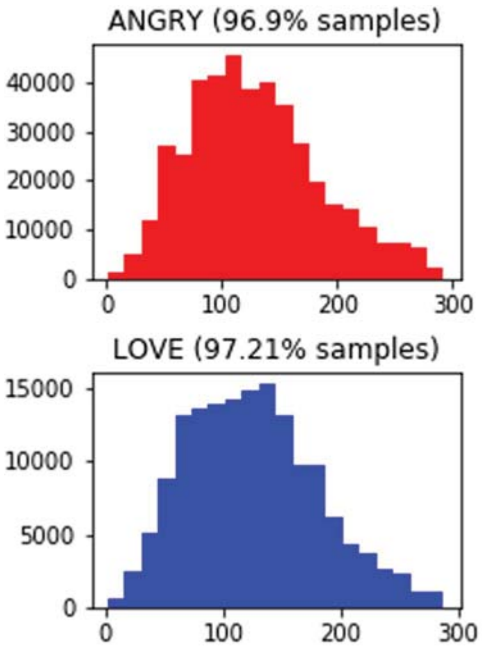


Fig. 5 Character level, post titles histogram

Fig. 6 Character level, post subtitles histogram



reaction. The following figures show the word clouds for these terms, considering each reaction. Figure 8 corresponds to the word cloud for comment terms associated with the HAHA reaction, Fig. 9 corresponds to the ANGRY reaction, Fig. 10 represents the SAD reaction and, finally, Fig. 11 shows the LOVE reaction.

Experimental Setup

Over the 1,020,557 remaining elements, a random sample of quadruples (title, subtitle, comment, reaction) was selected; in order to estimate the value of the desired parameter, this is the Fleiss kappa agreement measure [28]

Fig. 7 Character level, post comments histogram

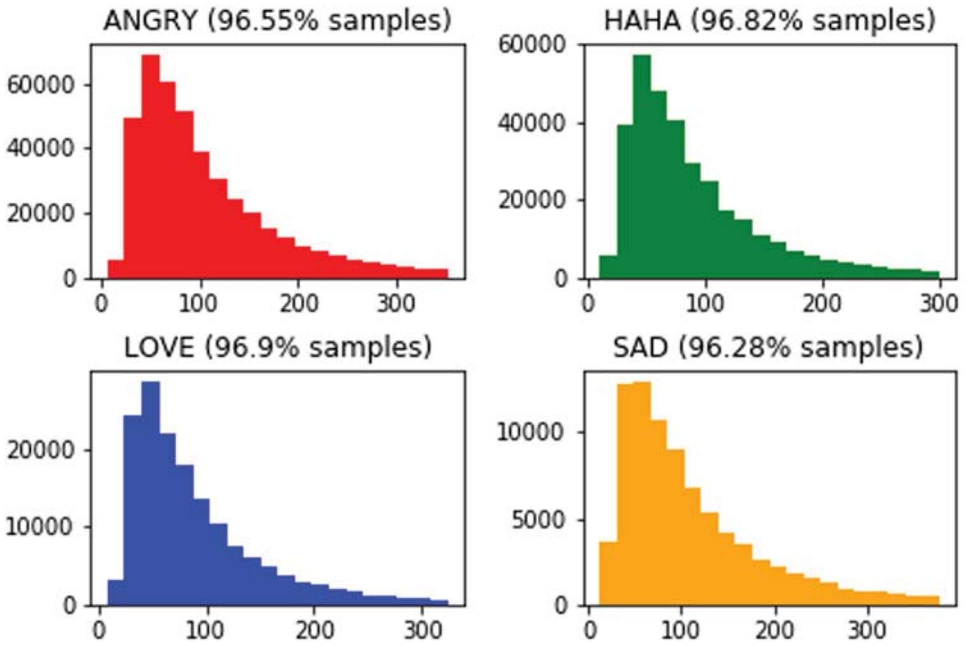


Table 3 Vocabulary overlapping level among comments

	HAHA	SAD	LOVE	ANGRY
HAHA	1.00	0.27	0.35	0.48
SAD	0.62	1.00	0.54	0.66
LOVE	0.57	0.39	1.00	0.59
ANGRY	0.42	0.26	0.31	1.00

for this particular dataset. To determine sample size, the finite population determination formula was used. This is shown in equation (2), where n is the size of the resulting sample, N is population size, Z is the statistical parameter depending on the confidence level, e is the margin of error, p is the probability for success, and $q = (1 - p)$.

$$n = \frac{N \times Z_{\alpha}^{2} \times p \times q}{e^{2} \times (N-1) \times Z_{\alpha}^{2} \times p \times q} \tag{2}$$

As there were no known proportions from p and q for this dataset, their values were set as 0.5. The confidence level was set at 95% and the maximum allowable error, at 5.23%. These last two parameters were the best possible with the resources available, but still considered acceptable for this work. The resulting sample size was 247.

The sample was split into 10 sets of 24 or 25 quadruples. Each set was then used to build a Google Form [62] with one classification task per quadruple. Set size was determined experimentally, since it was observed that larger sets yielded poorer agreement results; this could be due to human rater loss of concentration.

Hsueh et al. [63] stated that, to carry out the manual tagging phase, it is appropriate to involve experts in the task. Following this suggestion, psychologists were asked to carry out the manual classification task. Participants were shown a news title, a subtitle and comment, and then were asked to select what reaction, among four possible options (ANGRY, SAD, HAHA, and LOVE) the comment conveyed. Participants were allowed to select a second choice, but this was optional. Around 25 psychologists took part of this classification task. Each comment was reviewed at least by three individuals, as annotation quality can be improved through cross-validation and verification by several annotators [30].

acuchillo
foquistas
hidratante
enjendros
quevivanlosdos
mantera
cortaria
reformatorios
multaria
sueldono
hdremilp
dañadas
atorrantesss
disciplinaria
despreciable
soportaremos
arrechos
pueble
desgenerado
heredarán
atarla

Fig. 9 Word cloud of unique comment tokens for ANGRY

After the review, the Fleiss kappa agreement measure was calculated globally, and then, each reaction was considered versus the others, i.e., considering one reaction as a category and the remaining three as another category. Another measure to calculate the agreement, as done in [54], was considering the most voted class for each comment as the valid label. Then, the Fleiss kappa agreement measure was calculated for the original reaction and the reaction that received the most rater votes for each comment. In the case of a draw for most voted reaction, the optional secondary response and the original tag were used, if necessary, to break the tie.

Finally, to gain some insight about challenging cases, comments were analyzed using BabelSenticNet [32] to extract the main concepts and the overall polarity of each class. A global

humosvaldo
ajajajajajajajajasjajaj
bocabaret
sparring
jujuyyy
jaajajajajajajajajajaja
jajajjajajajaja
jajajajajajajjaajajjajajaja
vocabaret
pisala
suegras
auzqui
bengue
cabarulo
friomessi
jajaajajajaja

Fig. 8 Word cloud of unique comment tokens for HAHA

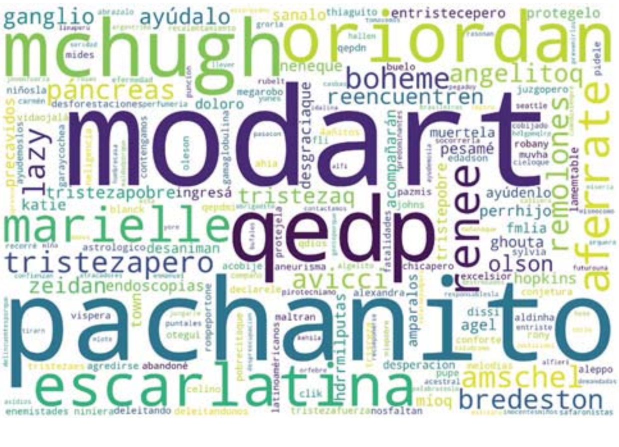


Fig. 10 Word cloud of unique comment tokens for SAD

tesón
pumita
lindaaa
grité
balá
felicitacionesque
hermosísimo
phil
bellooo
leoncitas
medicucho
atardecer
aripuca
collins
nairo
bendigaaa
coiro
verlaaa
petropresidente
nogales
felicitciones
mirkito
resi
chido
teleo
iturrieta
jacobs
esports
perfectono
encantó
expectacular
exelenteee
felicitacionesy
felicitacionesun
maravillo
bravisimo
muñeeecooo
braaavooo
proelección
fuerzalula
telenoche
hermosuraaa
acompañandolo
riverplatenses
peregrino

Fig. 11 Word cloud of unique comment tokens for LOVE

analysis was performed using concept clouds; also, several representative samples were analyzed for each class.

Results and Discussion

In this section, the main characteristics of the results, obtained after the manual tagging process, are described. These results are included as Electronic supplementary material 2.

- Filename: “manual_tagging_results.csv”
- Title: “Comments manually tagged by psychologists”
- Number of instances: 247 (“ANGRY”: 101, “HAHA”: 86, “LOVE”: 41, “SAD”: 19)
- Number of attributes per instance: 7
- Attribute information:

1. Sample code number (type: numeric) (reference to the sample obtained from “facebook_automatically_tagged_dataset.csv”)
2. Primary reaction selected by human rater #1 (HAHA, LOVE, ANGRY, SAD)
3. Secondary reaction selected by human rater #1 (HAHA, LOVE, ANGRY, SAD, None) (if selected, otherwise empty)
4. Primary reaction selected by human rater #2 (HAHA, LOVE, ANGRY, SAD)
5. Secondary reaction selected by human rater #2 (HAHA, LOVE, ANGRY, SAD, None) (if selected, otherwise empty)
6. Primary reaction selected by human rater #3 (HAHA, LOVE, ANGRY, SAD)
7. Secondary reaction selected by human rater #3 (HAHA, LOVE, ANGRY, SAD, None) (if selected, otherwise empty)

Table 4 Agreement among human raters

Metric	Scored result
Fleiss kappa global result	0.4911
Fleiss kappa ANGRY vs all	0.4933
Fleiss kappa HAHA vs all	0.4989
Fleiss kappa LOVE vs all	0.5332
Fleiss kappa SAD vs all	0.4240

Table 5 Agreement between human raters and original tag

Metric	Scored result
Fleiss kappa global result	0.4426
Fleiss kappa ANGRY vs all	0.4071
Fleiss kappa HAHA vs all	0.4415
Fleiss kappa LOVE vs all	0.5452
Fleiss kappa SAD vs all	0.4081

Overall, results are presented in Table 4, showing that agreement is moderate. The global Fleiss kappa score is 0.49 and, if individual reactions are considered, LOVE is the highest and SAD is the lowest.

As it can be seen in Table 5, if the original reaction in the dataset is considered as another reviewer, the global Fleiss kappa score drops to 0.4426, but still within the moderate agreement zone, the individual reaction with the highest value is still LOVE, but the lowest value is now shared by ANGRY and SAD.

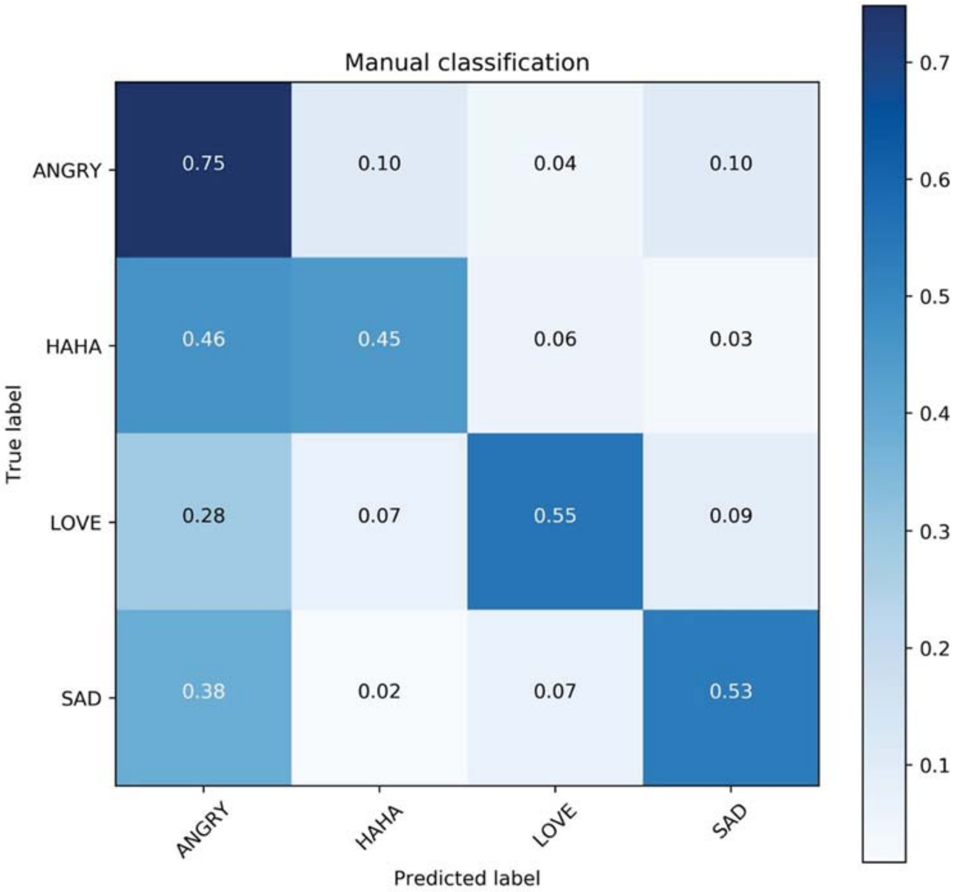
In the results presented in Table 6, to give more weight to the original reaction and to filter possible manual classification outliers, the manually classified reaction for every comment was decided by a vote. Fleiss kappa was then calculated among the most voted reaction for every sample and the original dataset reaction.

The second measure presented in Table 6 also considers the secondary reaction (if selected) as a vote. In the last two measures of the same table, if the voting is tied and the original reaction is among the most voted one, then the voting result is set to the original reaction. As it can be seen, all measures are also within the moderate agreement zone.

Table 6 Agreement between the most voted reaction and original tag

Metric	Scored result
Fleiss kappa vote first response	0.4409
Fleiss kappa vote first and second responses	0.4036
Fleiss kappa vote first response, ties as correct	0.4701
Fleiss kappa vote first and second responses, ties as correct	0.4922

Fig. 12 Manual classification vs. original reaction



To visualize where the disagreements between the original tag and the human raters were, a confusion matrix was built; the result is presented in Fig. 12. As it can be seen, ANGRY was the most accurately predicted but also the one with more false positives. Every reaction was confused with ANGRY, HAHA being the worst case. The remaining three reactions did not present classification problems between each other.

As it can be seen in the confusion matrix, most errors were between ANGRY and the rest of the reactions. To analyze the potential cause for these misclassifications, the Spanish version of BabelSenticNet [32] was used to extract and evaluate the polarity of the concepts mentioned in the comments.

The results are presented in Figs. 13, 14, 15, and 16, which show a concept cloud for every reaction. The color of the concepts in the cloud indicates their polarity. The more intense the red, the more negative the concept is; the more intense the green, the more positive concept is; gray concepts are close to neutral.

In addition to this, for every reaction, the average polarity of all concepts was also calculated. The reaction with the most negative average polarity is HAHA (- 0.00725), followed by ANGRY (0.02047), SAD (0.03534), and LOVE (0.0989). This could explain why psychologists misclassified many of the comments tagged with HAHA as ANGRY. On the other hand, the reaction with the most positive average polarity, LOVE, was the least confused with ANGRY.

Table 7 presents some of the most common misclassifications detected, which are HAHA, SAD, and LOVE classified as ANGRY. The other regions of the confusion matrix did not present a relevant number of misclassifications.

In some of the comments tagged as ANGRY but misclassified as HAHA, the author of the comment probably considers that the person mentioned in the topic has little or none credibility, and everything that person does or says makes the commenter laugh. Still, these comments contain some words with a very negative connotation that may have misled the human reviewers. This is the case of comments 4 and 8 in Table 7. Other comments misclassified in this way were sarcastic (such as comment #1).

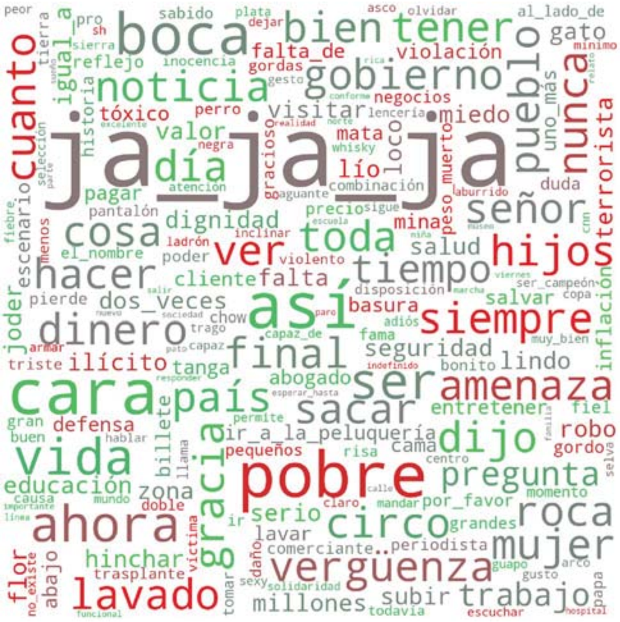


Fig. 13 Polarity analysis for HAHA reaction

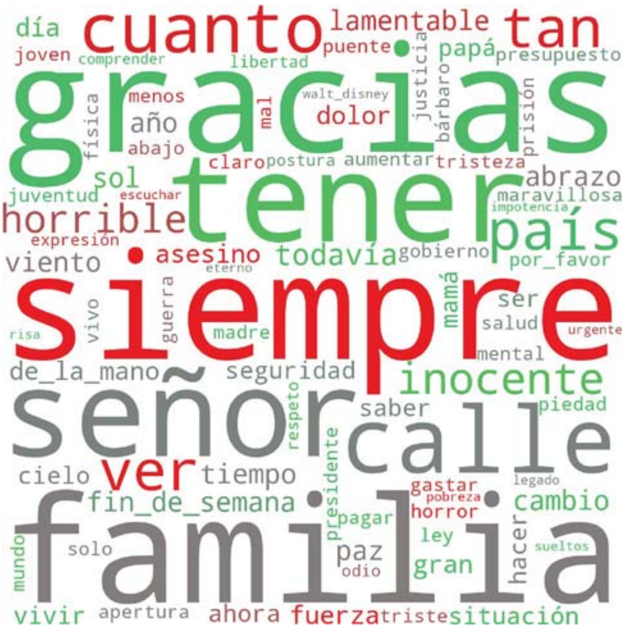


Fig. 15 Polarity analysis for SAD reaction

As regards the comments tagged as LOVE but classified as ANGRY, the person who posted the comment agrees with the event reported in the news, but in doing so, they also criticize something or someone. Examples of this behavior can be seen in comments 2, 7, and 9 in Table 7.

SAD and ANGRY comments are difficult to distinguish. An example of this is comment 3 in Table 7. A hint to distinguish them may be that SAD comments are written in a more respectful way than ANGRY comments.

Finally, comments 4, 5, and 6 in Table 7 present obvious hints (highlighted in bold) about what the actual reaction is. The misclassification in those cases could be due to the lack of concentration of the human reviewers. This may mean that the questionnaires should probably be shorter.

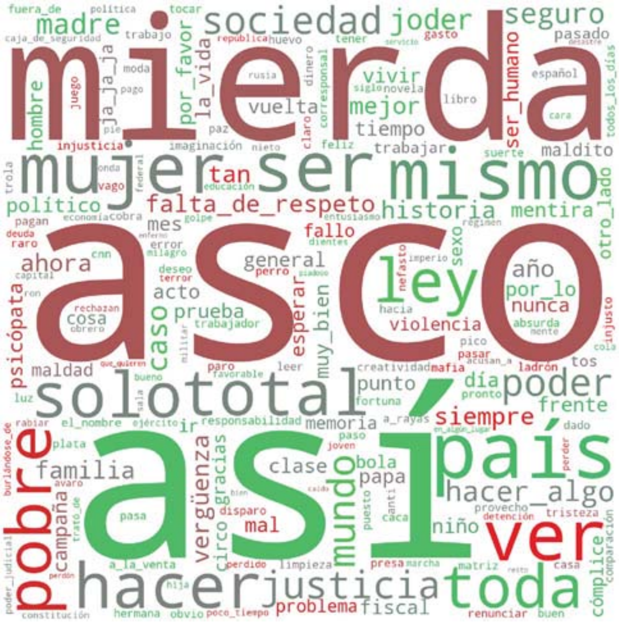


Fig. 14 Polarity analysis for ANGRY reaction

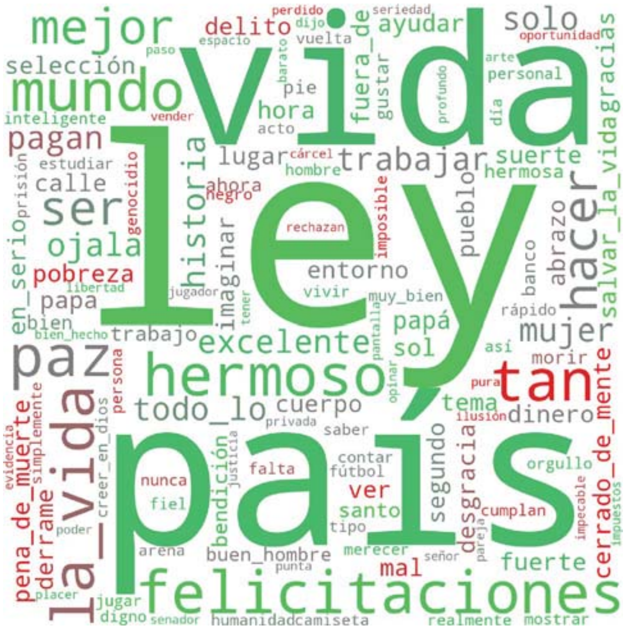


Fig. 16 Polarity analysis for LOVE reaction

Table 7 Representative cases of misclassifications, with English translation below

#	Reaction	Content		Errors
1	HAHA	Title	*El incómodo momento que vivió Pedro Pablo Kuczynski al intentar besar el anillo del papa Francisco* The uncomfortable moment when Pedro Pablo Kuczynski tried to kiss Pope Francis' ring	2/3 as ANGRY
		Comment	*Rarísimo que un papa se guíe por cuestiones políticas. Rarísimo* Very uncommon for a pope to act based on political issues. Very uncommon	
2	LOVE	Title	*Violador de nena atropellado por camión tenía herida de bala* Rapist of girl hit by truck had gunshot wound	2/3 as ANGRY
		Comment	*q se joda ojala q haya sufrido mucho antes de morir* Fuck him I hope he suffered a lot before he died	
3	SAD	Title	*A Maru Botana le echan en cara la muerte de su hijo* Maru Botana is blamed for the death of her son	2/3 as ANGRY
		Comment	*Qué barbaro cuanto odio, estamos en democracia, por si no lo recuerdan, hay liberdad de expresión y Maru defendió su postura, más allá si se comparte o no, se merece todo el respeto. RESPETO que se esta perdiendo!* Unbelievable, so much hatred, we are in a democracy, in case you don't remember that, there is freedom of expression and Maru defended her position, and regardless of whether we share it or not, she deserves all the respect. RESPECT, which is being lost!	
4	HAHA	Title	*Un hombre condenado por matar al bebé de su amante marchó contra el aborto en Río Grande* A man convicted for killing his lover's baby marched against abortion in Rio Grande	2/3 as ANGRY
		Comment	***Jajajajajajajajaja** la doble moral de los pro vida no acaba nunca. No se olviden que también marchan al lado de pedófilos y pederastas que en algún momento les van a meter las manos adentro del pantalón de sus propios hijos. Tienen el cerebro lavado. #YoNoMeMetoEnUteroAjeno* **Hahahahahahahahaha** the double standard of pro-life people never ends. Do not forget that they also march alongside pedophiles who at some point are going to put their hands inside their own children's pants. They are brainwashed. #IDoNotGetIntoSomeoneElsesUterus	
5	SAD	Title	*Se entregó el hombre acusado de violar a un nene en Chaco* Man accused of raping a boy in Chaco surrendered	2/3 as ANGRY
		Comment	*6 y 15 años de prisión? Nada más?!!!! Degenerado horrible y el nene arruinada su salud física y mental por siempre, que horror, **estoy triste!!!!*** 6 and 15 years in prison? Only that?!!!! Horrible degenerate and the boy has his physical and mental health ruined forever, what a horror, **I'm sad !!!!**	
6	HAHA	Title	*Maju Lozano explotó contra Baby Etchecopar* Maju Lozano exploded against Baby Etchecopar	3/3 as ANGRY
		Comment	*Esta gorda haciéndose la víctima por ser mujer **causa mucha mucha gracia**. Marmota, las bardeo por boludas, no por mujeres* This fat woman pretending to be a victim **is very very funny**. You are dumb, he questioned them for being stupid, not because they are women	
7	LOVE	Title	*Operativo contra los manteros en Liniers: desalojaron 475 puestos ilegales* Operation against street sellers in Liniers: 475 illegal stands removed	2/3 as ANGRY
		Comment	*Si quieren trabajar qué paguen impuestos como todo comerciantes x eso ellos venden mas barato x que usan espacio publico y no pagan impuesto y de paso se traen la droga para vender en Argentina como nadie controla nada en este pais. Estamos fritos con esta gente* If they want to work, they must pay taxes like all shop owners, they offer cheaper prices because they use public space and do not pay taxes, and they also bring drugs to sell in Argentina, since there are no controls for anything in this country. We are hopeless with these people	
8	HAHA	Title	*Marcha contra ajuste de planes sociales* March against the adjustment of social welfare plans	2/3 as ANGRY
		Comment	*El día que hagan un reclamo legítimo capaz que el pueblo los acompañe* The day they make a legitimate claim maybe the people will march with them	
9	LOVE	Title	*Comienza el juicio a Lázaro Báez por la ruta del dinero K* Lazaro Báez's trial begins on K money route	2/3 as ANGRY
		Comment	*Tiene que ser rápido las pruebas son contundentes hay pruebas de sobra no sé qué tanto tienen que estudiar* It has to be fast, the evidence is conclusive, there is plenty of evidence, I don't know what it is they have to analyze so much	

Conclusions and Future Work

In this paper, a new dataset of news, comments, and emotional reactions was presented. The dataset consists of 1,020,557 comments, each one tied to a news article (title and subtitle) and a specific reaction (the true value class). The number of entries is significantly larger than other manually tagged sentiment datasets that have been built for the Spanish language [44], which can be easily achieved by using noisy labels as content tags. However, no studies, at least for the Spanish language, compare the reliability of those tags versus manual tags, nor has any study linked tags directly to the comment instead of linking them to the originating news article.

As seen in the previous sections, emotionally tagged distantly supervised datasets can be automatically collected from social media articles. The agreement measured between the human raters is very similar to the one between human raters and the original tag; every measure presented is within the moderate agreement zone, which other authors [53, 54] considered suitable for sentiment classification training.

Although the agreement measure is a little lower compared with fully manually classified datasets, larger datasets can be built by using the guidelines presented in this article, as less or none manual tagging is required.

Filtering out duplicate comments and trolls improves the agreement measures presented. Many of the social media users that perform such practices are trolls whose input cannot be trusted.

The ANGRY reaction presented a significant number of false positives; the authors assume that this may be the consequence of unfiltered troll activity, so refining the troll filtering process may help improve this issue. This confusion could also be caused by misinterpreted sarcasm in the comments. Therefore, the presence of sarcastic comments in the dataset should be explored further, and sarcasm detection could be performed following the recommendations of Majumder et al. [5]. In addition, a more detailed polarity analysis by class could be performed by applying Sentic patterns [64] along with BabelSenticNet [32].

The next step of this research is to train machine-learning algorithms that can predict the emotion using a comment as input and can explain it as well [65]. As seen in SemEval 2019 [54], contextual information can be used to improve classification accuracy in textual dialogs; this could also be the case for interactions in social media, as responses to news articles are a form of communication or dialog. The use of semantic information should also be explored, as it may help improve classification accuracy [66].

Supplementary Information The online version contains supplementary material available at (https://doi.org/10.1007/s12559-020-09800-x).

Acknowledgments This publication was supported by public funds, and the authors would like to thank the following institutions and/or agencies:

• Instituto de Investigación y Transferencia en Tecnología (Centro CICPBA), Universidad Nacional del Noroeste de Buenos Aires, Argentina.

• Secretaría de Investigación, Desarrollo y Transferencia de la Universidad Nacional del Noroeste de Buenos Aires, Argentina.

• Gobierno de Aragón, a través del proyecto AffectiveLab-T60-20R, España. Instituto de Investigación en Informática LIDI (Centro CICPBA), Facultad de Informática, Universidad Nacional de La Plata, Argentina.

• Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina.

• Comisión de Investigaciones Científicas de la Provincia de Buenos Aires (CIC-PBA), Argentina.

• Ministerio de Ciencia, Innovación y Universidades (MCIU), España (contrato RTI2018-096986-B-C31).

The authors would also like to especially thank the psychologists who collaborated with this research by participating in the dataset validation process and the *Colegio de Psicólogos de la Provincia de Buenos Aires-Distrito III*, for liaising with the professionals.

Funding This study was funded by:

• Consejo Nacional de Investigaciones Científicas y Técnicas, Argentina, (RESOL-2020-131-APN-DIR#CONICET)

• Comisión de Investigaciones Científicas de la Provincia de Buenos Aires, Argentina (Acta No 1491).

• Convocatoria para la Acreditación de Proyectos y Solicitud de Subsidios de Investigación Bianuales 2019, Secretaría de Investigación Desarrollo y Transferencia, Universidad Nacional del Noroeste de la Provincia de Buenos Aires, Argentina, (EXP 536-2019).

• Gobierno de Aragón, España (AffectiveLab-T60-20R).

• Ministerio de Ciencia, Innovación y Universidades, España, (contrato RTI2018-096986-B-C31).

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

Ethical Approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Cambria E. Affective computing and sentiment analysis. IEEE Intell Syst. 2016;31(2):102–7.
2. Picard R. Affective Computing. MIT Press; 1997.
3. Cambria E, Poria S, Gelbukh A, Thelwall M. Sentiment analysis is a big suitcase. IEEE Intell Syst. 2017;32(6):74–80.
4. Chaturvedi I, Cambria E, Vilares D. Lyapunov filtering of objectivity for Spanish Sentiment Model. In: 2016 International Joint Conference on Neural Networks (IJCNN). Vancouver, British Columbia, Canada: IEEE; 2016. p. 4474–4481.
5. Majumder N, Poria S, Peng H, Chhaya N, Cambria E, Gelbukh A, et al. Sentiment and Sarcasm Classification With Multitask Learning. IEEE Intell Syst. 2019 May-June 1;34(3):38–43.

6. Majumder N, Poria S, Gelbukh A, Cambria E. Deep learning-based document modeling for personality detection from text. IEEE Intell Syst. 2017;32(2):74–9.
7. Medhat W, Hassan A, Korashy H. Sentiment analysis algorithms and applications: a survey. Ain Shams Eng J. 2014;5(4):1093–113.
8. Cambria E, Hussain A, Havasi C, Eckl C. Sentic Computing: Exploitation of Common Sense for the Development of Emotion-Sensitive Systems. In: Esposito A, Campbell N, Vogel C, Hussain A, Nijholt A, editors. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Berlin, Heidelberg: Springer Berlin Heidelberg; 2010. p. 148–156. (Lecture Notes in Computer Science; vol. 5967).
9. Bi J-W, Liu Y, Fan Z-P, Cambria E. Modelling customer satisfaction from online reviews using ensemble neural network and effect-based Kano model. Int J Prod Res. 2019;57(22):7068–88.
10. Chen L, Qi L. Social opinion mining for supporting buyers' complex decision making: exploratory user study and algorithm comparison. Soc Netw Anal Min. 2011;1(4):301–20.
11. Bae Y, Lee H. Sentiment analysis of twitter audiences: measuring the positive or negative influence of popular twitterers. J Am Soc Inf Sci Technol. 2012;63(12):2521–35.
12. Mahata D, Friedrichs J, Hitkul, Shah RR. Phramacovigilance - exploring deep learning techniques for identifying mentions of medication intake from twitter. 2018. arXiv preprint arXiv 1805.06375
13. Wang Z, Chong CS, Lan L, Yang Y, Beng S, Ho JC. Tong Fine-grained sentiment analysis of social media with emotion sensing. In, 2016 Future Technologies Conference (FTC) [Internet] San Francisco, California, USA: IEEE 2016;1361-1364
14. Munezero M, Montero CS, Sutinen E, Pajunen J. Are they different? affect, feeling, emotion, sentiment, and opinion detection in text. IEEE Trans Affect Comput. 2014 Apr-June 1;5(2):101–111.
15. Wang Z, Ho S-B, Cambria E. A review of emotion sensing: categorization models and algorithms. Multimed Tools Appl. 2020;3:1–30.
16. Ekman P, Friesen WV. Constants across cultures in the face and emotion. J Pers Soc Psychol. 1971;17(2):124–9.
17. Susanto Y, Livingstone AG, Ng BC, Cambria E, Cambria E. The hourglass model revisited. IEEE Intell Syst. 2020 Sept-Oct 1;35(5):96–102.
18. Mintz M, Bills S, Snow R, Jurafsky D. Distant supervision for relation extraction without labeled data. In: Su K-Y, Su J, Wiebe J, Haizhou L, editors. Proceedings of the 47th Annual Meeting ofthe ACL and the 4th IJCNLP of the AFNLP. Suntec, Singapore: Association for Computational Linguistics and Asian Federation of Natural Language Processing Associations; 2009. p. 1003–1011.
19. Pool C, Nissim M. Distant supervision for emotion detection using Facebook reactions. 2016. arXiv preprint arXiv 1611.02988
20. Kaur W, Balakrishnan V, Rana O, Sinniah A. Liking, sharing, commenting and reacting on Facebook: user behaviors' impact on sentiment intensity. Telemat Informatics. 2019;39(June):25–36.
21. Tian Y, Galery T, Dulcinati G, Molimpakis E, Sun C. Facebook sentiment: reactions and emojis. In: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media. Stroudsburg, PA, USA: Association for Computational Linguistics; 2017. p. 11–16.
22. Balakrishnan V, Govindan V, Arshad NI, Shuib L, Cachia E. Facebook user reactions and emotion: an analysis of their relationships among the online diabetes community. Malaysian J Comput Sci. 2019;Special Issue 3:87–97.
23. Bilal M, Malik N, Bashir N, Marjani M, Hashem IAT, Gani A. Profiling social media campaigns and political influence: the case of pakistani politics. In: 2019 13th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS). Karachi, Pakistan, Pakistan: IEEE; 2019. p. 1–7.
24. Hoque MT, Islam A, Ahmed E, Mamun KA, Huda MN. Analyzing performance of different machine learning approaches with doc2vec for classifying sentiment of bengali natural language. In: 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE). Cox's Bazar, Bangladesh: IEEE; 2019. p. 1–5.
25. Raad BT, Philipp B, Patrick H, Christoph M. ASEDS: Towards Automatic Social Emotion Detection System Using Facebook Reactions. In: 2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS). Exeter, United Kingdom: IEEE; 2018. p. 860–866.
26. Baj-Rogowska A. Sentiment analysis of Facebook posts: The Uber case. In: 2017 Eighth International Conference on Intelligent Computing and Information Systems (ICICIS). Cairo, Egypt: IEEE; 2017. p. 391–395.
27. Sandoval-Almazan R, Valle-Cruz D. Facebook impact and sentiment analysis on political campaigns. In: Proceedings of the 19th Annual International Conference on Digital Government Research Governance in the Data Age - dgo '18. New York, New York, USA: ACM Press; 2018. p. 1–7.
28. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull. 1971;76(5):378–82.
29. Mercado V, Villagra A, Errecalde M. Political alignment identification: a study with documents of Argentinian journalists. J Comput Sci Technol. 2020;20(1):43–52.
30. Lo SL, Cambria E, Chiong R, Cornforth D. Multilingual sentiment analysis: from formal to informal and scarce resource languages. Artif Intell Rev. 2017;48(4):499–527.
31. Cambria E, Li Y, Xing FZ, Poria S, Kwok K. SenticNet 6: Ensemble application of symbolic and subsymbolic AI for sentiment analysis. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management. New York, NY, USA: ACM; 2020. p. 105–114.
32. Vilares D, Peng H, Satapathy R, CambriaE. BabelSenticNet: A commonsense reasoning framework for multilingual sentiment analysis. In, 2018 IEEE Symposium Series on Computational Intelligence (SSCI) Bangalore, India: IEEE 2018 1292 1298
33. Justo R, Alcaide JM, Torres MI, Walker M. Detection of sarcasm and nastiness: new resources for Spanish language. Cognit Comput. 2018;10(6):1135–51.
34. Dashtipour K, Poria S, Hussain A, Cambria E, Hawalah AYA, Gelbukh A, et al. Multilingual sentiment analysis: state of the art and independent comparison of techniques. Cognit Comput. 2016;8(4):757–71.
35. Roth B, Barth T, Wiegand M, Klakow D. A survey of noise reduction methods for distant supervision. In: AKBC 2013 - Proceedings of the 2013 Workshop on Automated Knowledge Base Construction, Co-located with CIKM 2013. San Francisco, California: Association for Computing Machinery; 2013. p. 73–77.
36. Go A, Bhayani R, Huang L. Twitter sentiment classification using distant supervision. Technical Report Stanford University, 2010. Available from: https://www-cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf. Accessed 15 May 2020.
37. Bandhakavi A, Wiratunga N, Massie S, Padmanabhan D. Lexicon generation for emotion detection from text. IEEE Intell Syst. 2017;32(1):102–8.
38. Ahmad Refaee EA. Sentiment analysis for micro-blogging platforms in arabic [dissertation on the Internet]. Edinburgh, United Kingdom: Heriot-Watt University; 2016. [cited 2020 May 15]. Available from: https://www.ros-test.hw.ac.uk/bitstream/handle/10399/3166/RefaeeE_0816_macs.pdf?sequence=1&isAllowed=y
39. Suttles J, Ide N. Distant supervision for emotion classification with discrete binary values. In: International Conference on

Intelligent Text Processing and Computational Linguistics. Berlin, Heidelberg: Springer; 2013. p. 121–136.

40. Felbo B, Mislove A, Søgaard A, Rahwan I, Lehmann S. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In: Palmer M, Hwa R, Riedel S, editors. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics; 2017. p. 1615–1625.
41. Moctezuma D, Graff M, Miranda-Jiménez S, Tellez ES, Coronado A, Sánchez CN, et al. A Genetic programming approach to sentiment analysis for twitter. In: Villena Román J, García Cumbreras MA, Martínez Cámara E, Díaz Galiano MC, García Vega M, editors. Proceedings of TASS 2017: Workshop on Sentiment Analysis at SEPLN co-located with 33nd SEPLN Conference [Internet]; 2017 Sept 19; CEUR Workshop Proc. Volume 1896, 2017 [cited 2020 May 15]. p. 23–28. Available from: http://ceur-ws.org/Vol-1896/p1_ingeotec_tass2017.pdf
42. Martín C, Aguilar RM, Torres JM, Díaz S. Supervisión remota en el entrenamiento de un clasificador de sentimientos en comentarios turísticos. In: XXXIX Jornadas de Automática [Internet]; 2018 Sept 7–9; Badajoz, Spain. Comité Español de Automática (CEA); 2018 [cited 2020 May 15]. p. 644–650. Available from: http://dehesa.unex.es/bitstream/handle/10662/8530/978-84-09-04460-3_644.pdf?sequence=1&isAllowed=y
43. Sociedad Española del Procesamiento del Lenguaje Natural (SEPLN). Taller de Análisis de sentimientos en Español (TASS) [Internet]. 2020 [cited 15 May 2020] Available from: http://tass.sepln.org
44. Cumbreras MÁG, Gonzalo J, Cámara EM, Unanue RM, Rosso P, Carrillo-de-Albornoz J, et al., editors. Proc Iber Lang Eval Forum (IberLEF 2019) co-located with 35th Conf Spanish Soc Nat Lang Process (SEPLN 2019) [Internet]. CEUR Workshop Proc. Volume 2421, 2019 [cited 2020 May 15]. Available from: http://ceur-ws.org/Vol-2421/
45. Broß J. Aspect-oriented sentiment analysis of customer reviews using distant supervision techniques [dissertation on the Internet]. Berlin, Germany: Free Universitat Berlin; 2013. [cited 2020 May 15] Available from: https://refubium.fu-berlin.de/bitstream/handle/fub188/6693/Dissertation_Juergen_Bross.pdf;jsessionid=C2E12B8B1868AA5AC7167DAB14296BAE?sequence=1
46. Sahni T, Chandak C, Reddy N, Singh M. Efficient twitter sentiment classification using subjective distant supervision. In: 2017 9th International Conference on Communication Systems and Networks (COMSNETS). Bangalore, India: IEEE; 2017. p. 548–553.
47. Refaee E, Rieser V. Evaluating distant supervision for subjectivity and sentiment analysis on arabic twitter feeds. In: Proceedings of the EMNLP 2014 Workshop on Arabic Natural Langauge Processing (ANLP). Stroudsburg, PA, USA: Association for Computational Linguistics; 2014. p. 174–179.
48. Deriu J, Lucchi A, De Luca V, Severyn A, Müller S, Cieliebak M, et al. Leveraging large amounts of weakly supervised data for multi-language sentiment classification. In: WWW '17 Companion: Proceedings of the 26th International Conference on World Wide Web Companion. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee; 2017. p. 1045–1052.
49. Marchetti-Bowick M, Chambers N. Learning for microblogs with distant supervision: political forecasting with twitter. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. Avignon, France: Association for Computational Linguistics; 2012. p. 603–612.
50. Carletta J. Squibs and discussions: assessing agreement on classification tasks: the kappa statistic. Comput Linguist. 1996;22(2):248–54.
51. Hearst MA. TextTiling: segmenting text into multi-paragraph subtopic passages. Comput Linguist. 1997;23(1):33–64.
52. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20(1):37–46.
53. Gambino OJ, Calvo H. Predicting emotional reactions to news articles in social networks. Comput Speech Lang. 2019;58:280–303.
54. Chatterjee A, Narahari KN, Joshi M, Agrawal P. SemEval-2019 Task 3: EmoContext Contextual Emotion Detection in Text. In: May J, Shutova E, Herbelot A, Zhu X, Apidianaki M, Mohammad SM, editors. Proceedings of the 13th International Workshop on Semantic Evaluation. Stroudsburg, PA, USA: Association for Computational Linguistics; 2019. p. 39–48.
55. Kemp S. Digital 2020: 3.8 billion people use social media [Internet]. We Are Social Ltd; 2020 [updated 2020 Jan 30; cited 2020 May 15]. Available from: https://wearesocial.com/blog/2020/01/digital-2020-3-8-billion-people-use-social-media
56. Becerra M. Medios digitales en Argentina: la película y la foto [Internet]. Letra P; 2018 [updated 2018 Sept 20; cited 2020 May 15]. Available from: https://www.letrap.com.ar/nota/2018-9-20-16-3-0-medios-digitales-en-argentina-la-pelicula-y-la-foto
57. Facebook. Facebook API Graph [Internet]. 2020 [cited 15 May 2020] Available from: http://developers.facebook.com
58. Bird S, Klein E, Loper E. Natural language processing with python. O'Reilly Media Inc.; 2009.
59. Al-Rfou R. PYCLD2 - Python bindings to CLD2 [Internet]. 2020 [cited 15 May 2020]. Available from: https://pypi.org/project/pycld2/
60. Han S. googletrans [Internet]. 2015 [cited 15 May 2020]. Available from: https://pypi.org/project/googletrans/
61. Craker N, March E. The dark side of Facebook®: The Dark Tetrad, negative social potency, and trolling behaviours. Pers Individ Dif. 2016;102:79–84.
62. Google. Google Forms [Internet]. 2020 [cited 15 May 2020] Available from: https://www.google.com/intl/es-419_ar/forms/about/
63. Hsueh P, Melville P, Sindhwani V. Data quality from crowdsourcing: A Study of Annotation Selection Criteria. In: Ringger E, Haertel R, Tomanek K, editors. Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing - HLT '09. Morristown, NJ, USA: Association for Computational Linguistics; 2009. p. 27–35. Available from: https://www.aclweb.org/anthology/W09-1904.pdf
64. Poria S, Cambria E, Gelbukh A, Bisio F, Hussain A. Sentiment data flow analysis by means of dynamic linguistic patterns. IEEE Comput Intell Mag. 2015;10(4):26–36.
65. Burdisso SG, Errecalde M, Montes-y-Gómez M. PySS3: A Python package implementing a novel text classifier with visualization tools for Explainable AI. 2019. arXiv preprint arXiv 1912.09322
66. Ferretti E, Errecalde M, Rosso P. Does semantic information help in the text categorization task? J Intell Syst. 2008;17(1–3):91–106.

Dataset Disclaimer Given that the dataset consists of a compilation of public user comments on Facebook, the authors of this article expressly declare that the users are solely responsible for their posts and that, under no circumstances, the content reflects the opinions of the authors or express verified information.