Expert Systems With Applications 212 (2023) 118817

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

A novel flexible feature extraction algorithm for Spanish tweet sentiment analysis based on the context of words

García-Díaz Pilar [a,*], Sánchez-Berriel Isabel [b], Pontiel-Martín Diego [c], González-Ávila José Luis [d]

[a] Universidad de Alcalá, Escuela Politécnica Superior, Departamento de Teoría de la Señal y Comunicaciones, 28805 Madrid, Spain
[b] Universidad de La Laguna, Facultad de Ciencias, Departamento de Ingeniería Informática y de Sistemas, 38200 San Cristobal de La Laguna, S/C de Tenerife, Spain
[c] Universidad de Alcalá, Escuela Politécnica Superior, Departamento de Teoría de la Señal y Comunicaciones, 28805 Madrid, Spain
[d] Universidad de La Laguna, Facultad de Ciencias, Departamento de Ingeniería Informática y de Sistemas, 38200 San Cristobal de La Laguna, S/C de Tenerife, Spain

ARTICLE INFO

Keywords:
Sentiment analysis
Context semantics
Grouping Genetic Algorithm
Flexible feature extraction
Twitter

ABSTRACT

A tweet polarity classifier is presented with four categories: positive, neutral, negative and no opinion. A grouping genetic algorithm performs feature extraction on the reviews. The feature definition is based on entropy and semantic context described as the relative positions between words. The feature selection is flexible because it is customized to each word studied in the reviews. The algorithm has been applied with two corpuses written in Spanish, of 3,413 tweets and more than 63,000 tweets, to classify an evaluation set of 1,899 reviews. The results were evaluated by the metrics M−F1 and accuracy. The algorithm has improved the results of both metrics and on both corpuses compared to the previous literature works, achieving a M−F1 of 0.640 and an accuracy of 0.689. The flexibility property in feature extraction has been the major qualitative improvement of the classifier.

1. Introduction

The last decade has seen the proliferation of numerous Social Media Sites and Apps. Nowadays, Twitter,[1] CiteULike,[2] Mendeley,[3] Facebook,[4] YouTube,[5] WhatsApp, Instagram, Tiktok, Telegram, Snapchat, Pinterest and more are universally known (Auxier and Anderson, 2021). A world population of over 7.9 billion was estimated at the end of 2021.[6] Facebook reported 2.895 billion monthly active users during that year.[7] Social networks are accessible to a significant portion of the world's population. Moreover, users upload an increasing amount of information to social networks on a daily basis. The huge number of users connected and interacting with each other, as well as the high speed of information broadcast that characterizes Internet, result in a very high influence of the opinions of social networks users on the decisions of other users and on the actions of many companies, regardless of their size (Kwayu et al., 2021). The opinions of many users encourage many others to also express theirs, generating a stimulation in masses of consumers (Naeem and Ozuem, 2021; Varghese and Agrawal, 2021), and a very strong influence on the economy. During the last decade, efforts have been made in research and development of systems for the automatic extraction of information from opinion texts. Data and information on the Internet have recently become a key target for companies and organizations of many different purposes (Ahuja et al., 2015; Rathika and Soranamageswari, 2022; Starosta, 2022; Xue et al., 2021; Madhu, 2018; Daas and Puts, 2014a). Data and user information are considered a decisive resource for advertising campaigns, service

Peer review under responsibility of Submissions with the production note 'Please add the Reproducibility Badge for this item' the Badge and the following footnote to be added:The code (and data) in this article has been certified as Reproducible by the CodeOcean: https://codeocean.com. More information on the Reproducibility Badge Initiative is available at https://www.elsevier.com/physicalsciencesandengineering/computerscience/journals..

* Corresponding author.
E-mail addresses: pilar.garcia@uah.es (G.-D. Pilar), isanchez@ull.edu.es (S.-B. Isabel), diego.pontiel@edu.uah.es (P.-M. Diego), jlgavila@ull.es (G.-Á. José Luis).

[1] https://twitter.com.
[2] https://www.citeulike.org.
[3] https://www.mendeley.com.
[4] https://www.facebook.com.
[5] https://www.youtube.com.
[6] https://www.worldometers.
[7] https://www.statista.com.

https://doi.org/10.1016/j.eswa.2022.118817
Received 23 May 2022; Received in revised form 6 September 2022; Accepted 8 September 2022
Available online 14 September 2022

improvement (Nilashi et al., 2021; Trivedi and Singh, 2021), new product orientation and implementation of new services (Mudinas et al., 2019; Vanaja and Belwal, 2018), public communications (Devi and Chingangbam, 2021), and others (Hossain et al., 2018; Khosravinik, 2018). Therefore, sentiment analysis in texts is a very useful method in many different disciplines to obtain quantitative indexes very quickly or even immediately, compared to other techniques that require more processing time, estimation or computation. For example, the consumer confidence index in economics (Van den Brakel et al., 2017; Daas and Puts, 2014b), altmetrics in biometrics (Barbounaki et al., 2021; Hassan et al., 2020; Wouters et al., 2019; Colón-Ruiz et al., 2019; Imran et al., 2018) and (Sarker et al., 2015).

Twitter has been more intensively studied by researchers than other social networks due to some characteristics: high population penetration, behavior as a repository of opinion of a large number of users, and broad spectrum of specific topics and interests on the platform (Passi and Motisariya, 2022; Villavicencio et al., 2021; Batista and Ribeiro, 2013). Those reasons have been a strong incentive to stimulate research in sentiment analysis methods (Ikram et al., 2022; Carvalho and Plastino, 2021; Nistor, et al., 2021; Hassan et al., 2020; Naseem et al., 2020; Rahman et al., 2019).

Sentiment analysis involves the techniques in the field of Natural Language Processing (NLP) to detect subjective information in a document (Pintas et al., 2021). An interesting overview on sentiment analysis can be read in (Liu, 2012). Traditionally, analyzing algorithms produced a numerical value representing the polarity of a single sentence regarding the sentiment, classified among three classes as positive, negative, or neutral polarity (Choi and Cardie, 2009; Tan et al., 2011). Afterwards, more specialized systems were developed with classifiers of three to five polarities for texts composed of several sentences (Srivastava et al., 2022). Nowadays, there has been a strong interest in sentiment analysis in short texts or micro-blogs written by users expressing their opinions of different products and services: online sales, hotels, travel, restaurants, etc. (Gokalp et al., 2020; Gu et al., 2018). Automatic text classification algorithms have been developed based on lexicons (Li, 2020) and based on machine learning, either supervised or unsupervised (Ahuja and Sharma, 2022; Rahman et al., 2019; Vashishtha and Susan, 2019). Hybrid solutions of those have been used as well (Li et al., 2022). The use of deep learning has emerged strongly in recent years (Gondhi et al., 2022; Tan et al., 2022; Trisna and Jie, 2022) and (Zulqarnain et al., 2022). The decision on text polarity has been addressed in the literature by multiple approaches such as Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM) or Gated Recurrent Units (GRU). Convolutional Neural Networks (CNN) are the most widespread in recent research. In (Ni and Cao, 2020), LSTM and GRU techniques were combined to process long-term information by taking advantage of the computational efficiency of GRU. The authors used GloVe representations on the input data, then employed models with an LSTM layer and a GRU layer. (Onan, 2022) applied a convolutional recurrent neural network architecture also combining two layers LSTM and GRU to extract features at high level and reduce the feature space dimensionality. (Zulqarnain et al., 2022) implemented GRU in two states, including pre-feature attention and post-feature attention. For sentiment analysis in Spanish, systems submitted to competitions on sentiment analysis, such as TASS (Taller de Análisis Semántico, Semantic Analysis Workshop) 2018 and TASS 2019 for Spanish texts, mostly used Recurrent Neural Networks and Transformers (Díaz Galiano et al., 2019a; Díaz-Galiano et al., 2019b).

Aspect-Based Sentiment Analysis (ABSA) has been explored in the literature to understanding long complicated sentences and obtaining interaction between the sentiment polarity of aspects and contexts (Liang et al., 2021). Deep learning is the common factor in many such research (Cheng et al., 2022; Yang et al., 2019). Complex linguistic structures analysis provides more accurate solutions and can solve tasks such as the effect of negation on sentiment. Graph Convolutional Networks (GCN) allow detecting these structures in contextual information. (Zhang et al., 2021) combined GCN with BERT (Bidirectional Encoder Representations from Transformers) to complement syntactic information and contextual information with long-range word dependencies. (Zhao et al., 2022b) combined CGN, BERT and a dynamic multiple weight mechanism to extract sentence-level dependency relations in aspect-based sentiment analysis. (Zhao et al., 2022a) used ABSA with aggregator functions (max and mean aggregators) to obtain local node neighborhood information, they also established node subdependencies to capture long distance dependency information.

Detection of neutrality (neutral opinions) is still an open issue in sentiment analysis (Chan et al., 2022). (Valdivia et al., 2018) applied a consensus voting method to improve classification accuracy. Another issue to consider in sentiment analysis is opinion ambivalence, referring to cases where the same opinion contains both positive and negative sentiments. (Zhang et al., 2021) approached this matter using GCN. (Wang et al., 2020) addressed the ambivalence problem through introducing ambivalent classes in classification.

The black box behavior of algorithms is a work in progress by Explainable Artificial Intelligence in sentiment analysis area. In this research line, (Zucco et al., 2018) extracted the contributing words in polarity prediction. (Cambria et al., 2022) used unsupervised and reproducible subsymbolic techniques such as autoregressive language models to overcome explainable limitation in this field.

Text classification is characterized by a high level of complexity. The very large solution space, the wide variability of human language and high noise in texts cause the complexity of sentiment analysis in texts. Texts written in social networks are unstructured data. The noise comes from incomplete word content, abbreviations, grammatical errors, misspellings, etc. Input data cleaning and preprocessing must necessarily be carried out before sentiment analysis (Birjali et al., 2021; de Oliveira and Merschmann, 2021; Duong and Nguyen-Thi, 2021; Mhamed et al., 2021; Ahuja et al., 2019).

The large size of the solution space follows the well-known Curse of Dimensionality or Hughes Effect (Hughes, 1968), which states that the higher the dimensionality, the lower the reliability of the estimation of the statistical parameters. Dimensionality reduction has been addressed for years through feature selection techniques that exclude the least relevant characteristics in the classification operation (Onan and Korukoğlu, 2017; Rui et al., 2016; Agarwal and Mittal, 2013). Feature extraction is a widely applied technique in many areas of science with very high performance (Fan et al., 2022; Sachadev and Bhatnagar, 2022). Feature selection has been successfully applied in sentiment analysis of documents as well (Jain and Jain, 2022; Osmani et al., 2022; Wang and Hong, 2019; Madasu and Elango, 2020; Wang and Lin, 2020). (Gokalp et al., 2020) used a wrapper method of feature selection oriented to sentiment analysis to reduce dimensionality and improve results. The features in the classification state correspond to bags of words, tf-idf and unigrams. Today it remains a crucial focus for improving performance and an open work line in this area. Wrapper methods for feature extraction are more computationally expensive than filter selection methods (feature rankings). (Setya Rintyarna et al., 2019) applied ranking-type feature selection methods to analyze the sentiment of polysemic words using supervised learning. However, wrapper methods provide higher accuracy (Gokalp, 2002). To reduce their computational requirements, combined feature space exploration methods have been used (Rathika and Soranamageswari, 2022; Salam and Ali, 2020; Coban et al., 2018). Feature selection is implemented on different criteria: entropy study (Ahuja et al., 2019), semantic study (Singh et al., 2020) or semantic study through contexts (Corallo et al., 2020). The combination of criteria may offer greater accuracy in the assessment of the text sentiment but to the best of our knoweledge, no research has been carried out that combines the entropy study with semantics in the context of words.

Sentiment analysis has also been approached with evolutionary algorithms. (García-Mendoza et al., 2020) presents a multiclass classification alternative to deep learning algorithms when large volumes of

Table 1
Distribution of tweets from the training, development and testing datasets in InterTASS corpus according to their polarity. Data source: (Diaz Galiano, 2018).

InterTASS corpus	Training		Development		Training + Development		Testing	
Label	# Tweets	%	# Tweets	%	# Tweets	%	# Tweets	%
P	321	31.84	156	30.83	477	31.51	642	33.81
NEU	133	13.20	69	13.64	202	13.34	216	11.37
N	416	41.27	219	43.28	635	41.94	767	40.39
NONE	138	13.69	62	12.25	200	13.21	274	14.43
Total	1,008	100	506	100	1,514	100	1,899	100

data are not available. The authors use a differential evolutionary algorithm to estimate the optimal weights of multiple classifiers. A voting system unifies their decision. In Onan and Korukoğlu (2017) the authors design a multi-objective differential evolutionary algorithm combining several techniques and a voting system applied to sentiment analysis problems as well as to other text classification problems. On the other hand, the feature extraction techniques using genetic algorithms show a good compromise between accuracy, reliability and convergence (Jain and Jain, 2022; Iqbal et al., 2019; Das et al., 2018).

Research in sentiment analysis has reached a maturity point. However, plenty of work is still needed to obtain more robust, effective and efficient solutions, especially in the accuracy of polarity classification in short texts. Human communication, both oral and written and in all languages, is extremely elaborate and complex to automate, characterized by a wide diversity of accents, dialects and an infinite number of variations that change in a soft or hard way the message sentiment. This intrinsic property of human language causes a lower precision in the application of equivalent artificial intelligence algorithms in areas in which very high precision solutions were obtained (García-Díaz et al., 2020). Another outstanding issue is that sentiment analysis methods have been mostly developed for the English language, and most of the literature concerns the analysis of texts written in English. Nevertheless, the information in other languages published on social networks is significant. Twitter has half of the messages written in languages different from English. Spanish is the second language used on Twitter and Facebook (Fernández Vítores, 2020). Research in sentiment analysis must also progress for languages other than English. Most of the research has been implemented on English data, with interest in the application of developed methods to other languages growing progressively (Poria et al., 2020). NLP research in Spanish is still far from the advance in English. SEPLN[8] created in 2012 the TASS to promote the development of specific NLP techniques for the processing of opinion in texts written in Spanish.

In this paper, a metaheuristic algorithm for sentiment analysis applied to Twitter texts, called Neighbor-sentiment algorithm, is presented. This research work does: (1) classify Spanish tweets according to message sentiment, (2) no opinion in tweets is an added category in the classification, (3) feature extraction is based on entropy and semantic context and (4) feature selection is flexible, since each word has its independent feature extraction, variable in size and identity. The words in the texts are defined with a group of features associated with the entropy and semantic context. The algorithm is trained with a corpus and on which it performs a flexible extraction of the most relevant features for classification. The customized group of features for each word constitutes an individual classifier. A text is analyzed by as many classifiers as words it contains. The final classification decision of the text or review is weighted by the vote of all the classifiers applied on it. Feature extraction is carried out by a hybrid Genetic Grouping Algorithm (GGA) with an Extreme Machine Learning (ELM). A new text to be classified could not contain words considered relevant by the algorithm because it did not train with enough texts containing the words of the new one. In that case, the algorithm does not have enough information to provide a classification decision and the text is labeled as unclassified. The results obtained by the developed algorithm are compared with previous literature works (Díaz Galiano et al., 2018).

The paper is structured as follows. Section 2 details the corpus used in the implemented algorithm. The metrics to be used for the comparison of results with other studies are also described in this section. Section 3 explains the criteria applied by the algorithm in the classification of text polarity. Section 4 details the classification algorithm developed, especially the genetic grouping algorithm that integrates it and the special coding used for the reviews. Section 5 discusses the results obtained in the evaluation of the classification algorithm. These results are compared with the results in the literature. Finally, section 6 summarizes the conclusions of this work and future research lines.

2. Material and methods

This paper describes a text classifier algorithm according to the analysis of the sentiment expressed by the text. The classification criterion is oriented to the study of the word context by analyzing the most relevant words (it also considers entropy as a criterion). The algorithm includes an evolutionary algorithm that requires a training corpus to identify the most relevant words in the documents. Once the system is trained, the algorithm is evaluated by classifying an unknown set of documents.

The authors have selected a challenge published as " Task 1" ofTASS-2017[9] to evaluate the performance of the developed algorithm. The challenge is detailed in (Díaz Galiano et al., 2018). Both corpuses used in this work was published by TASS: InterTASS and General corpus. They are composed of Twitter reviews written in Spanish language. The mixed dialects of the same language could negatively impact the sentiment analysis. Each corpus consists of tweets manually evaluated and classified at document level by a team of experts of the SEPLN. The classification of tweets according to opinion establishes four categories: positive (P), negative (N), regular or neutral (NEU) and no sentiment or no opinion (NONE).

InterTASS corpus is composed of a total of 3,413 tweets published in the years 2016 and 2017. All tweets are characterized by containing at least one adjective and longer than 3 words. InterTASS corpus has been specifically configured for the TASS-2017 challenge. TASS group them into three disjoint sets, named training set C_{Train}, development set C_{Dev} and testing set C_{Test} (see Table 1).

The General corpus consists of 57,832 tweets, which were published in 2011 and 2012 (García-Cumbreras et al., 2016a). The tweets collect opinions on different topics: politics, economy, communication, and culture among others. The general corpus is appropriate for evaluating sentiment analysis algorithms because of its large item number. Previous editions of TASS have used this corpus. (García-Cumbreras et al., 2016b; Villena-Román et al., 2015). A team of experts labeled tweets in six categories according to the opinion of the text: strong positive (P +), positive (P), strong negative (N +), negative (N), neutral (NEU) and no opinion (NONE). The General corpus was adjusted to the TASS-2017

[8] SEPLN (Sociedad Española de Procesamiento del lenguaje Natural) Spanish Society of Natural Language Processing: https://www.sepln.org/ (Last access in April 2022).

[9] https://www.sepln.org/workshops/tass/2017/.

Table 2
Distribution of tweets of the General corpus data for the four categories of opinion and sentiment. Data source: (Diaz Galiano, 2018).

Label	# Tweets	%
P	21,262	36.76
NEU	1,300	2.25
N	15,124	26.15
NONE	20,146	34.83
Total	57,832	99.99

challenge by moving from 6 classes to 4, unifying P with P + on the one hand, and unifying N with N + on the other hand. Table 2 shows the percentages of tweets for each category.

The proposed algorithm processes the two corpuses in two different operations to classify InterTASS testing set C_{Test} in both cases (last two columns in Table 1). In the case of InterTASS corpus, the algorithm trains with the training and development sets (columns #6 and #7 in Table 1). In the case of General corpus, the algorithm can train with the complete corpus as well as with InterTASS training and development sets, while keeping the tweets from the C_{Test} unknown (last four columns in Table 3). All tweets run a cleaning and preprocessing operation before the classification. The applied text preprocessing techniques are indicated in (Birjali et al., 2021; Ahuja et al., 2019): tokenization, normalization, lemmatization, 'stop words' removal and noise removal.

Note that the distribution of the four polarity categories is not balanced in Tables 1-3. Both corpuses contain a majority of positive (P) and negative (N) category items. The General corpus has less than 3 % of the tweets with neutral category (NEU), this is a strong bias that will increase the difficulty of classifying this class for any algorithm trained with this corpus. The unbalanced frequency of polarities in the corpus may impact the results obtained by previous researchers (Díaz Galiano et al., 2018). Firstly, the algorithms were ordered in the literature according to M−F1 metrics first, and then according to accuracy. None of the algorithms exceed a M−F1 of 60 % nor an accuracy of 65 %. These limits indicate the complexity of the sentiment analysis problem. Both metrics are defined in (Ahuja et al., 2019). Accuracy is a weighted calculation between the total number of successful predictions and the total number of failures, according to equation (1):

$$accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)} \tag{1}$$

where TP (True Positive) indicates the number of items correctly classified, FP represents the number of false positives, FN is the number of false negatives and TN is the number of true negatives. M−F1 metric is the value F1 or F1-score, which is a harmonic mean between precision and recall. Precision and recall are computed according to equations (2) and (3):

$$precision = \frac{TP}{(TP + FP)} \tag{2}$$

$$recall = \frac{TP}{(TP + FN)} \tag{3}$$

with harmonic mean defined to M−F1 as:

$$M\text{-}F1 = \frac{(2{*}precision{*}recall)}{(precision + recall)} \tag{4}$$

3. Theory

The classification of a short text (from 4 to 20 words) according to sentiment can be approached in different ways. The decision to classify the message or review consists of choosing one among several categories or sentiments. The number of categories is usually between three (positive, neutral, negative) and five (very positive, positive, neutral, negative, negative, very negative). The classification decision can be resolved by a single classifier or as a result of a combination of several classifiers through a weighted voting system. When a set of decisors is considered, each one could be associated with a word appearing in the text. Words are selected according to their influence on the polarity classification of the message. Some words with weak influence on sentiment are excluded, such as articles, pronouns and other words depending on the topic of the review (in some cases they are words with numerical meaning, in other cases they are certain extremely frequent verbs such as "to be" or "to have").

The study and analysis of human language is a highly complex task. The message meaning and the expressed feelings are easily altered by multiple factors such as polysemy, the presence of particular words and even their relative position. Varying the position of a given word in the message can convey the opposite polarity to the initial one. As an example, compare the sentiment of the following sentences: "the seller was very little tolerant" and "The little seller was very tolerant".

Natural language is so sophisticated that the message sentiment does not always match the sentiment of the words that it contains. A simple example is the sentence: "This is not a nice view". The word "*not*" is decisive in classifying the message sentiment as negative, overriding the positive sentiment of the word "*nice*". Other times the sentiment associated with a particular word is a catalyst for the opposite sentiment that the message would have in the absence of that word. Consider the ironic statement: "The street was covered with waste offering a very *nice* view".

The word meaning is sometimes associated with a sentiment that is different from the one conveyed by the message. For example, the word "difficult" is generally associated to a negative opinion. In the sentence "The exam was very difficult", the sentiment of the word "difficult" is consistent with the negative sentiment of the whole sentence. However, the text "Difficult roads often lead to beautiful destinations" denotes typically positive opinion.

The word position in the sentence can also influence the evaluation of the text sentiment. For example, the negation "not/no" has a different effect on the overall sentiment in the following sentences:

- "The swimming pool was not clean" expresses a negative sentiment.
- "The swimming pool was clean, no insects" conveys a positive sentiment.

The complexity of natural language allows situations in which one or more neighbor words are able to flip the generalized text sentiment. An example is the sentence: "The seller was very little tolerant". Note that the word "tolerant" is associated with a positive sentiment. The word "very" before "tolerant" increases the positive sentiment. However, the

Table 3
Distribution of tweets from General corpus and InterTASS corpus according to the four polarity categories (P, NEU, N, NONE).

	General corpus		Training + Development InterTASS corpus		General + Training + Develp. InterTASS		Testing InterTASS	
Label	# Tweets	%	# Tweets	%	# Tweets	%	# Tweets	%
P	21,262	36.76	502	33.16	21,764	36.67	642	33.81
NEU	1,300	2.25	193	12.75	1,493	2.52	216	11.37
N	15,124	26.15	609	40.22	15,733	26.51	767	40.39
NONE	20,146	34.83	210	13.87	20,356	34.30	274	14.43
Total	57,832	99.99	1,514	100.00	59,346	100.00	1,899	100.00

word "little" between them reverses this sentiment and becomes predominant in the evaluation of the message polarity. Now, the word "very" emphasizes the sentiment of "little". Consequently, the text is evaluated with a negative sentiment even though it contains the word "tolerant" with opposite sentiment. J. R. Firth's distributional hypothesis holds that the meaning of a word can be derived from the contexts in which the word is used. For this reason, the words preceding and/or following each relevant word are often studied. The number of previous and/or later words is defined by a maximum distance to the relevant word. The sequence of neighboring words is defined as *n*-grams.

A word is defined as "influential" if its presence in the review or text is relevant to the text polarity. The impact of an influential word on the message sentiment is associated both to its semantics and to the context semantics.

The presence of a word in the review is usually quantified through the entropy or word frequency in the review (lower occurrence frequency implies higher entropy and vice versa), also through the frequency in the corpus (El-Halees, 2015; Nigam et al., 1999). Absolute frequency concerns to the number of word occurrences in the document. Relative frequency refers to the number of word instances in the corpus. Numerous literature research use entropy as the key criterion in sentiment analysis (Xue et al., 2021; Xie et al., 2019). In some works, the entropy is added to the study of other parameters to improve the classifier quality (Jagdale et al., 2022; Devi and Chingangbam, 2021). An interesting theoretical background of the maximum entropy calculation in assigning classes to words as a function of the document in the corpus to which they belong can be read in (El-Halees, 2015).

Summarizing the above comments, there are two fundamental factors that classify the document polarity in one way or another: entropy and semantics through contexts. The combined study of both elements offers a higher accuracy guarantee in the sentiment evaluation. The document representation for the polarity study should consider the word position and the proximity between them. The algorithm developed by the authors is based on the semantic study of both the relevant words and their neighbors (word context), also considering the word frequency. In addition, the approach used in this work is novel because a customized feature extraction is carried out for each word in the document. For each word, the features that provide higher accuracy in the classification are extracted. This selection is variable in size and independent of the feature extraction for the rest of the words. In the literature there is research based on the study of entropy or based on the context of words (Corallo, 2020; Singh, 2020; Ahuja et al., 2019), but we are not aware of works in which text representation is carried out from contexts, in the sense of word embeddings algorithms such as word2vec (Mikolov et al., 2013), Glove (Pennington et al., 2014) or BERT (Devlin et al., 2018), nor is flexible feature extraction performed.

3.1. Defining features based on the proximity of words

Consider a set of categories to approach the classification of texts or reviews, written in the same language, according to sentiment analysis.

$$categories = \{P, NEU, N, NONE\} \tag{5}$$

where P denotes positive sentiment, NEU indicates that the sentiment is neutral, N refers to negative sentiment and NONE indicates that the text conveys no opinion or sentiment. Let us consider a corpus C consisting of a collection of reviews R_i:

$$C = \{R_i\}, i = 1, 2, \cdots, n \tag{6}$$

where n is the total number of reviews in the corpus C. A review is expressed as an array of lemmatized words w_j of a dictionary D. The dictionary D records all the different lemmatized words existing in the corpus C. The dictionary may also contain words from another corpus does not present in C.

$$R_i = \{w_j\}, w_j \in D \tag{7}$$

Note that a review may contain duplicate words. The word w_j is identified in the dictionary by a unique numeric identifier id_x, where x represents the position of the word in the dictionary D. Thus, the review can also be described as:

$$R_i = \{id_x\}, id_x \in D$$

$$D = \{id_x\}, x = 1, 2, \cdots, \#D \tag{8}$$

The corpus C is divided into three groups of documents to be processed by the classification algorithm: *Training* set or C_{Train}, *Development* set or C_{Dev}, and *Testing* set or C_{Test}. C_{Train} is composed by the set of the reviews used to train the algorithm, C_{Dev} contains the reviews which will be used to test the algorithm to adjust its parameters during the training and C_{Test} is composed by a set of unknown reviews for the algorithm and to evaluate its performance as a classifier. The three sets are expected to be disjoint, so that each review R_i belongs to only one of the three, especially for reviews of C_{Test}.

$$C = C_{Train} \bigcup C_{Dev} \bigcup C_{Test} \tag{9}$$

The influence or relevance of a word w_j on the polarity of a review R_i is related to the proximity of other words in R_i. Equation (10) defines an *n*-gram as the collection of words before and after w_j with maximum distance or threshold Th in the review. Note that for a value $Th = 0$ you have the word w_j. Given a value of Th, the optimization algorithm will process the *n*-grams of the review to be classified, one *n*-gram per w_j.

$$\{w_{j+s}\}, s \in Z, |s| \leq Th \tag{10}$$

Note that each element of the *n*-gram will be represented by its corresponding id_x identifier according to the dictionary. Consider as an example the review: "The water in the hotel pool is clean and crystal blue every morning". The process of lemmatization will use the infinitive form "to be" for the word "is". Equation (11) describes the content of the *n*-gram associated to the word w_6 = "pool" with a $Th = 2$. No numeric identifiers have been assigned in the *n*-gram for simplicity.

$$\{w_{6-2}, w_{6-1}, w_6, w_{6+1}, w_{6+2}\}, j = 6, Th = 2$$

$$\{the, hotel, pool, be, clean\}$$

$$\{id_a, id_b, id_c, id_d, id_e\} \tag{11}$$

Intuitively, the occurrence of the word "clean", at position + 2 to j = 6 in the review, will have a stronger influence on the general sentiment of R_i than the occurrence of the article "the" at position −2. Therefore, the algorithm records for the word "pool" that the feature position + 2 is more effective than position −2 for the classification of future reviews containing "pool". By increasing the threshold value up $Th = 5$, the *n*-gram has a larger size:

$$\{w_{6-5}, w_{6-4}, w_{6-3}, w_{6-2}, w_{6-1}, w_6, w_{6+1}, w_{6+2}, w_{6+3}, w_{6+4}, w_{6+5}\},$$

$$j = 6, Th = 5$$

$$\{the, water, in, the, hotel, pool, be, clean, and, crystal, blue\} \tag{12}$$

The influence of the word w_j on the classification of a review is related to the existence of w_j itself, as well as to the presence of other words in some proximity to it. The use of a higher threshold value increases the number of features for the words in the review, which may improve the classification accuracy. The maximum threshold is reached when words far away from each other do not affect the semantics of the context.

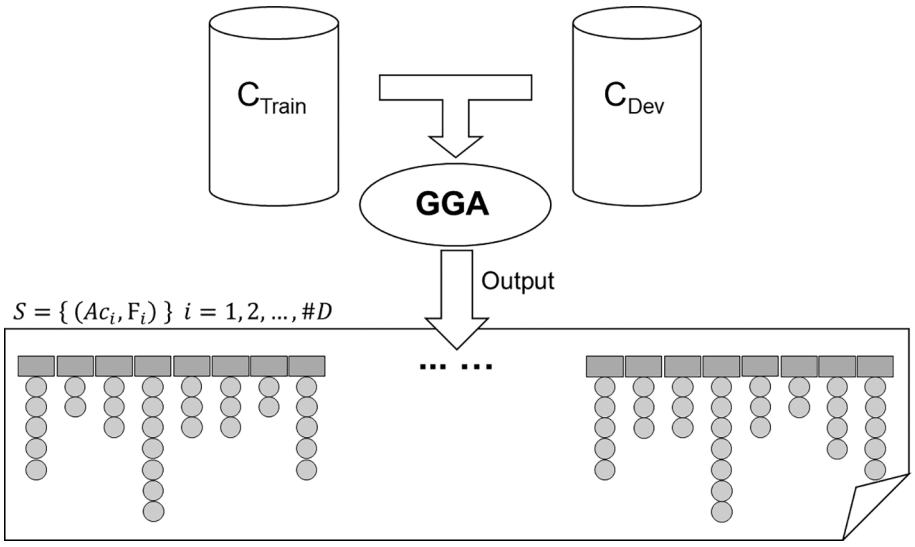

Fig. 1. Feature extraction for the most relevant words set for sentiment analysis for C_{Train} and C_{Dev} by GGA.

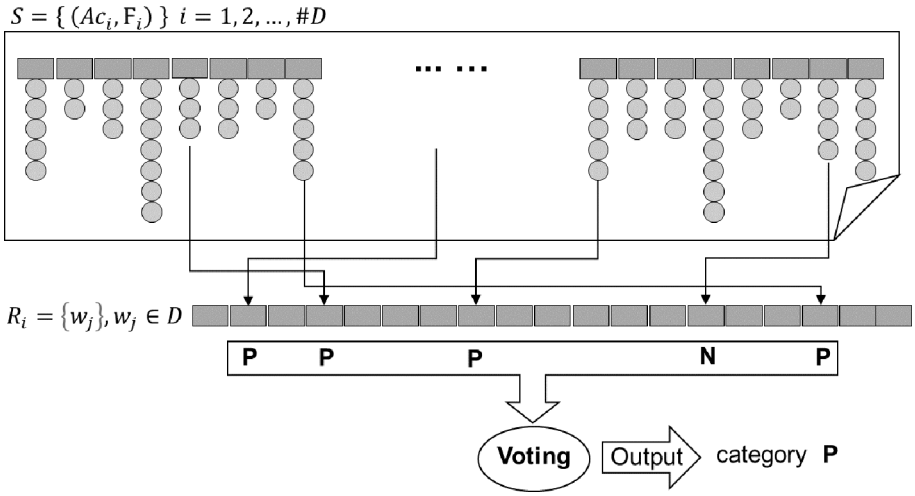

Fig. 2. Classification process of a review R_i using the classifier set S provided by GGA.

4. Algorithm for polarity or sentiment classification

The classification algorithm developed by the authors, named Neighbor-sentiment algorithm, consists of a classifier that performs a flexible and customized feature extraction for the words of the reviews contained in a corpus. Fig. 1 represents graphically the feature extraction from a set of reviews and Fig. 2 shows the classification process of new reviews from the previous feature selection.

Fig. 1 illustrates the algorithm application on the review sets C_{Train} and C_{Dev}. The algorithm returns as solution $S = \{S_i\}$ the most relevant words set for sentiment analysis for C_{Train} and C_{Dev}. Each word is described by the measured classification accuracy during training (filled rectangles in Fig. 1) and customized feature selection (columns of filled circles in Fig. 1). The Neighbor-sentiment algorithm defines the feature set of a word as the aggregate of the neighbor words not farther than Th positions, together with the absolute frequency and the relative frequency of the word.

A word and the flexibly selected features together constitute an individual classifier S_i. The algorithm provides a set of classifiers varying in number and composition. For each solution S_i, the algorithm has recorded its average accuracy on the analyzed reviews (C_{Train} and C_{Dev} sets). Classifiers with low average accuracy are discarded. Classifiers with acceptable classification average accuracy are considered useful, being associated to relevant words for the polarity study.

Varying the classification accuracy threshold that is considered acceptable changes the number of individual classifiers. A higher threshold guarantees higher quality classifiers, but the number of classifiers is smaller, i.e., fewer words are examined in the reviews. In this case, there may be reviews that do not contain words registered in the classifiers and the algorithm does not have enough information to make a classifying decision. The quality of the individual classifiers during the feature extraction process depends on the size of the corpus the algorithm is trained on. A larger corpus size provides more information about the words and a larger number of words studied.

A genetic grouping algorithm performs feature extraction for the words under study. The operation of this algorithm is described in the next section. For a word w_j with identifier id_x, the algorithm provides as a solution S_i the classifier (Ac_x, F_x), where Ac_x is the average accuracy of the classifier associated with the word and F_x is the set of selected features or descriptors of that word, as expressed in equation (13).

$$S_i = (Ac_x, F_x), F_x = \{d_i\}, i = 0, 1, 2, \cdots, Max_F \tag{13}$$

where Max_F is the maximum number of features or descriptors defined for the word with identifier id_x. After processing the genetic grouping

algorithm for all the words in the training and development sets, the solution S is composed, which consists of a set of classifiers associated to words in the dictionary D as shown in equation (14).

$$S = \{(Ac_i, F_i)\}i = 1, 2, \cdots, \#D \tag{14}$$

In Fig. 1, the features associated with each classifier S_i are represented by a circle column in the set S. The variability of the number of descriptors of S_i is plotted with variable length circle columns. The number of elements in S is fixed by defining a minimum accuracy threshold (Ac_{min}) for Ac_i.

The classifier set S, obtained from the GGA training, is evaluated by classifying the C_{Test} reviews. Fig. 2 shows the classification operation of a review R_i, belonging to C_{Test}, by applying S. R_i is evaluated by the individual classifiers of S_i associated to the words of R_i. The final classification decision of R_i is the majority vote of the applied classifiers. In the example of Fig. 2, five classifiers have been applied to R_i: four classifiers decide category P and another classifier decides category N. The final vote classifies R_i with category P. If the review contains no words associated with S, the review is not classified because the algorithm has no information to decide. This situation is handled a false negative in the classification process.

4.1. Genetic grouping algorithm for feature extraction

In clustering problems, elements are classified into several categories. Each element is defined by a set of features of size Max_F. When Max_F is very large, classification using all the features does not have high accuracy due to the Huges effect (Hughes, 1968). Then a selection of features common to all elements is carried out to improve accuracy by discarding features that are not relevant or influential in the classification. The solutions have variable length, since neither the number of features nor their identification are defined.

In this research work, feature extraction has not been performed because of the high dimensionality, rather because the decisive features for polarity classification depend on the context semantics. It is efficient to select the best features to improve the classification accuracy. On the other hand, in the application of sentiment analysis, it is known that the most appropriate features to recognize the word sentiment in the context of the review R_i may not always coincide with the most competent features for other words of the same review. This is the reason the authors implement flexible feature selection. Flexibility means that the features selected for each word are different, customized based specially on the context of the words. GGA performs this flexible feature extraction.

The generic GGA algorithm was first developed by (Falkenauer, 1993). The GGA has been broadly used in optimization problems in many different fields of knowledge, obtaining excellent results. It consists of a genetic algorithm with a coding specially adapted to solve clustering problems.

The genetic algorithm is an evolutionary algorithm in which a population of solutions progresses through a series of consecutive generations. Each individual represents a coded solution to the optimization problem. Each solution is characterized by a fitness value that quantifies its adaptation to the environment (Forrest, 1996). The goal of the individuals is to survive through each generation, adapting to the environment and overcoming competitions against other individuals. After a maximum number of generations, the algorithm returns one or more of the solutions with the best fitness value discovered. Several actions are taken during a generation: matching of individuals for crossover, crossover generating offspring, mutation with an established probability, calculating the fitness function (cost function) of the new individuals, and finally, selection of individuals from the total population to participate in the coming generation. Individuals with better fitness are more likely to overcome the selection processes. However, randomness is present and can help individuals with poor fitness value to pass to the next generation. A repair function may be required to ensure that the individuals correspond to feasible solutions to the optimization problem. Many variations of these operations have been implemented in the years of research (Sohail, 2021). In high complexity problems, simple alternatives of these operations, such as random parent matching, single point crossover or two random points crossover, are often applied.

The authors have adapted the GGA developed in (García-Díaz et al., 2020). The encoding of the solution is a key issue to ensure the efficiency of the genetic algorithm as well as the cost function. This function evaluates the average accuracy of the classification of the elements of the training set with the features selected by the algorithm. The fitness value calculation is carried out a very large number of times during the algorithm, so it must be fast and efficient. This function has been implemented with an Extreme Learning Machine that ensures both requirements. A description of the GGA, especially the encoding of the solutions, the crossover and mutation operators and the fitness function can be read in (García-Díaz et al., 2020).

4.2. Specific coding for the classification of reviews

A review $R = \left\{id_a, id_b, \cdots, id_y\right\}$ defined by id_x identifiers of its words by Equation (8) is encoded in the Neighbor-sentiment algorithm as a matrix of size $N \times Max_F$. Equation (15) represents the encoding matrix, where N is the number of words in R and Max_F is the total number of features for each word.

$$\#\left\{id_a, id_b, \cdots, id_y\right\} = N$$

$$\begin{bmatrix} id_{a,-Th} id_{a,-Th+1} \cdots id_{a,0} id_{a,1} \cdots id_{a,Th} f_a^{abs} f_a^{rel} \\ id_{b,-Th} id_{b,-Th+1} \cdots id_{b,0} id_{b,1} \cdots id_{b,Th} f_b^{abs} f_b^{rel} \\ \cdots \\ id_{y,-Th} id_{y,-Th+1} \cdots id_{y,0} id_{y,1} \cdots id_{y,Th} f_y^{abs} f_y^{rel} \end{bmatrix} \tag{15}$$

where Th is the threshold or maximum proximity distance between words, $id_{i,j}$ is the identifier of the word located j positions after the corresponding word with identifier id_i. The component f_i^{abs} is the absolute frequency of the word with identifier id_i. The component f_i^{rel} is the relative frequency of the word with identifier id_i. Note that the number of components is $Max_F = 2*Th + 3$, since the neighbor Th to the left of the word, the neighbor Th to its right, the word itself, its absolute frequency and its relative frequency are considered. When a word with identifier id_i lacks a neighbor at position j, the value of the component $id_{i,j}$ is null.

As an example, the review R = "The water in the hotel pool is clean" consists of a total of N = 8 words. Given a threshold $Th = 3$, the coding matrix for R has a size $N \times Max_F = 8 \times 9$. Three properties are observed in the use of this encoding: a) A word can be repeated two or more times in the same review; b) Numerous components of the matrix have the same non-null value; c) Several components of the matrix have a null value.

a) Note that the word "the" has two instances w_1 and w_4 in the review. Each instance is processed individually even with the same identifier in the dictionary. Since they have different neighbor words, each one contributes its own components in the encoding matrix.
b) The identifiers of the words in the review will be repeated in the matrix multiple times as the identifiers of the neighbors to each word are recorded. This information is stored in the first $2*Th+1$ columns in the matrix.
c) As mentioned before, when a word with identifier id_i has no neighbor at position j, the value of the component $id_{i,j}$ in the matrix is null.

Table 4
Results of the Neighbor-sentiment algorithm run on the InterTASS corpus for the classification of tweets from the InterTASS Testing set C_{Test}.

Label	# Tweets	% Tweets	Prediction	No answer	TP	FP	FN	% Precision	% Recall	% M−F1
P	642	33.81	429	98	317	112	325	73.89	49.38	59.20
NEU	216	11.37	82	58	16	66	200	19.51	7.41	10.74
N	767	40.39	1,036	65	624	412	143	60.23	81.36	69.22
NONE	274	14.43	44	87	19	25	255	43.18	6.93	11.95
Total	1,899	100.00	1,591	308	976	615	923			

Table 5
Neighbor-sentiment algorithm metrics comparison with the two best results in the literature on the classification of tweets in the C_{Test} set using InterTASS corpus.

Algorithm	Precision	Recall	M−F1	Accuracy
ELiRF-UPV-run1	–	–	49.3	60.7
RETUYT-svmcnn	–	–	47.1	59.6
Neighbor-sentiment	57.76	51.39	54.39	61.35

elements are considered false negatives for metrics in the last three columns of Table 4: accuracy, recall and M−F1. Data from columns #6 to #8 (true positive as TP, false positive as FP and false negative as FN) are used to compute the accuracy and recall values.

The results show a notable difference in accuracy and M−F1 for the categories positive (P) and negative (N) versus neutral (NEU) and no opinion (NONE). This is in accordance with the unbalanced distribution of categories in the InterTASS corpus described above. Some specialization of the algorithm in the detection of negative category tweets is also noticeable: 624 of the 767 tweets with negative polarity were successfully classified, and only 143 were not detected. This represents a recall value of 81.36 %.

Table 6
Results of the Neighbor-sentiment algorithm in classifying tweets in C_{Test} set trained with the General corpus.

Label	# Tweets	% Tweets	Prediction	No answer	TP	FP	FN	% Precision	% Recall	% M−F1
P	642	33.81	493	64	403	90	239	81.74	62.77	71.01
NEU	216	11.37	100	42	49	51	167	49.00	22.68	31.01
N	767	40.39	1,025	37	669	356	98	65.27	87.22	74.66
NONE	274	14.43	54	84	31	23	243	57.41	11.31	18.90
Total	1,899	100.00	1,672	277	1,152	520	747			

5. Results and discussion

This section presents the results obtained by the Neighbor-sentiment algorithm. The results are compared with previous literature (Díaz Galiano et al., 2018), where several algorithms were applied for the classification of the same tweet set (InterTASS testing set C_{Test}) by using two corpuses of different size: the InterTASS (Table 1) and the General corpus (Table 3). Four categories were used in the sentiment analysis, as shown in equation (5).

A single dictionary with more than 53,000 words was generated for both corpuses. The GGA algorithm was run for all words in the dictionary with the following configuration: 60 generations with a population of 50 individuals, two random points crossover and mutation probability of 10 %. Solutions were discarded when the calculated accuracy did not exceed a minimum threshold of $Ac_{min} = 67\%$. The maximum number of features in the review coding was $Max_F = 13$, corresponding to a threshold value $Th = 5$. The Neighbor-sentiment algorithm has been implemented in Python (version 3.8.8) on the Spyder 4.2.1 framework in Anaconda Navigator.[10] It has been executed with distributed programming with an Intel Core i5 computer connected to a virtual server of the University of Alcalá.

Table 4 shows the classification results of the tweets in C_{Test} from the InterTASS corpus discussed in Section 2. The categories in the InterTASS corpus are unbalanced (see Table 1): most of the tweets are labeled with negative sentiment (N), about a third of the reviews have positive sentiment (P) and approximately a quarter of the total belong to the minority categories, neutral (NEU) and no opinion (NONE). These percentages are also read in column #3 (% Tweets) of Table 4. Column #4 (Prediction) gives the total number of tweets that the algorithm classifies in each category. The column labeled "No answer" lists the number of reviews of each category that the algorithm is not able to classify due to lack of information in the solution set S (the review does not contain words that overcome the minimum accuracy threshold Ac_{min}). These

Table 5 shows the final metrics of the Neighbor-sentiment algorithm, calculated from the last three columns of Table 4 and the percentage indicated in column #3 of the same table. Table 5 also shows the metrics of the two algorithms from the literature (Díaz Galiano et al., 2018) that obtained the highest M−F1 and accuracy values. (Díaz Galiano et al., 2018) M−F1 ranks ahead of accuracy. However, the Neighbor-sentiment algorithm offers the best values for both metrics.

Table 6 presents the results of the classification of the same tweet set C_{Test} when the Neighbor-sentiment algorithm was trained with the General corpus of TASS. General corpus holds more balance in the number of tweets of positive, negative and no opinion categories compared to the InterTASS corpus. Table 6 has the same structure as Table 4. Note that the algorithm obtains better accuracy, recall and M−F1 metrics in all categories in comparison to the InterTASS corpus in Table 4. This is consistent with a larger corpus size, which the algorithm is expected to generate a larger and more efficient set of solutions. On the other hand, a higher improvement is observed in the classification of the neutral and positive categories than in the negative and non-opinion classes. Specifically, for the neutral category: accuracy has been multiplied by a factor of 2.5 (from 19.51 % in InterTASS to 49 % in General corpus), whereas recall and M−F1 have tripled.

A reduction in the unclassified tweet number in all categories due to lack of information in the set of algorithm solutions is observed, since a larger corpus was used. The reduction in negative polarity is

Table 7
Neighbor-sentiment algorithm metrics comparison with the two best results in the literature on the classification of tweets in the C_{Test} set using the General Corpus.

Algorithm	Precision	Recall	M−F1	Accuracy
INGEOTEC-evodag 003	–	–	57.7	64.5
jacerong-run-1	–	–	56.9	70.6
Neighbor-sentiment	67.85	60.66	64.06	68.90

[10] https://www.anaconda.com.

noteworthy: 65 unclassified tweets with InterTASS compared to 37 unclassified tweets with the General corpus. However, the improvement in the no opinion category is weak. This point will guide one of the future work lines.

Table 7 shows the metrics of the Neighbor-sentiment algorithm in C_{Test} classification from General corpus knowledge, together with the metrics of the two best algorithms from (Díaz Galiano et al., 2018). Neighbor-sentiment is characterized by better metrics in both parameters: M−F1 and accuracy. The algorithms in the literature included in Table 5 do not match those in Table 7, however the Neighbor-sentiment algorithm achieves the best metrics in both corpuses.

6. Conclusions

A novel algorithm for polarity classification of 1,899 tweets into four categories: positive, neutral, negative and no opinion has been described. Two corpuses published by TASS were used: the InterTASS corpus with 3,413 tweets and the General corpus with 57,832 tweets, both written in Spanish. The algorithm has obtained better metrics than previous research in both corpus: using the InterTASS corpus it obtained a M−F1 of 54.39 % and an accuracy of 61.35 % and working with the General corpus M−F1 and accuracy increase to 64.06 % and 68.90 %, respectively.

The text words to be classified are analyzed based on the entropy and semantics in the word context. The innovative aspect of the algorithm consists of a flexible feature extraction, where the feature selection for each word is customized and based on the frequency and the semantics in the word context. Feature extraction is therefore variable in number and independent of the rest of the words in the corpus. The authors consider that the flexibility in the feature selection process has been fundamental for the improvement of metrics compared to the algorithms published in the literature.

However, future versions of the algorithm must improve the metrics for the category “no opinion”, which indicates the missing sentiment or that the text does not express any opinion. In the current version the number of true positives for this category with the General corpus does not increase in the same proportion as for the other categories when compared to a corpus of smaller size. The responsibility for this seems to belong to the algorithm.

Future work is also focused on the application of this classification technique to corpus written in languages different from Spanish. Another future research line will be the specialization of independent binary classifiers working in a coordinated system on the same text. It would develop a binary classifier for each polarity to answer true or false to the corresponding category membership. The outputs of these classifiers would be combined appropriately as a multiclass classifier. The authors have already initiated this work, and good results have been obtained for positive and negative categories. Binary classifiers for neutral and no opinion classes involve more effort.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data are attached as supplementary data.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.eswa.2022.118817.

References

Agarwal, B., & Mittal, N. (2013). Optimal feature selection for sentiment analysis. In *International conference on intelligent text processing and computational linguistics* (pp. 13–24). Berlin: Springer.

Ahuja, R., Chug, A., Kohli, S., Gupta, S., & Ahuja, P. (2019). The impact of features extraction on the sentiment analysis. *Procedia Computer Science, 152*, 341–348.

Ahuja, R., Rastogi, H., Choudhuri, A., & Garg, B. (2015, March). Stock market forecast using sentiment analysis. In 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom) (pp 1008-1010). IEEE.

Ahuja, R., & Sharma, S. C. (2022). *Sentiment Analysis on Different Domains Using Machine Learning Algorithms. In Advances in Data and Information Sciences* (pp. 143–153). Singapore: Springer.

Auxier, B., & Anderson, M. (2021). Social media use in 2021. *Pew Research Center, 1*, 1–4.

Barbounaki, S. G., Gourounti, K., & Sarantaki, A. (2021). Advances of Sentiment Analysis Applications in Obstetrics/Gynecology and Midwifery. *Materia Socio-Medica, 33*(3), 225.

Batista, F., & Ribeiro, R. (2013). Sentiment analysis and topic classification based on binary maximum entropy classifiers. *Procesamiento del lenguaje natural, 50*, 77–84.

Birjali, M., Kasri, M., & Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems, 226*, Article 107134.

Cambria, E., Liu, Q., Decherchi, S., Xing, F., & Kwok, K. (2022). SenticNet 7: a commonsense-based neurosymbolic AI framework for explainable sentiment analysis. Proceedings of LREC 2022.

Carvalho, J., & Plastino, A. (2021). On the evaluation and combination of state-of-the-art features in twitter sentiment analysis. *Artificial Intelligence Review, 54*(3), 1887–1936.

Chan, J. Y. L., Bea, K. T., Leow, S. M. H., Phoong, S. W., & Cheng, W. K. (2022). State of the art: A review of sentiment analysis based on sequential transfer learning. *Artificial Intelligence Review*, 1–32.

Cheng, L. C., Chen, Y. L., & Liao, Y. Y. (2022). Aspect-based sentiment analysis with component focusing multi-head co-attention networks. *Neurocomputing, 489*, 9–17.

Choi, Y., & Cardie, C. (2009). Adapting a polarity lexicon using integer linear programming for domain specific sentiment classification. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2, pages 590–598. Association for Computational Linguistics.

Coban, O., Ozyildirim, B. M., & Ozel, S. A. (2018). An empirical study of the extreme learning machine for Twitter sentiment analysis. *International Journal of Intelligent Systems and Applications in Engineering, 6*(3), 178–184.

Colón-Ruiz, C., Segura-Bedmar, I., & Martínez, P. (2019). Análisis de Sentimiento en el dominio salud: Analizando comentarios sobre fármacos. *Procesamiento del Lenguaje Natural, 63*, 15–22.

Corallo, A., Fortunato, L., Massafra, A., Pasca, P., Angelelli, M., Hobbs, M., … Ciavolino, E. (2020). Sentiment analysis of expectation and perception of MILANO EXPO2015 in twitter data: A generalized cross entropy approach. *Soft Computing, 24*(18), 13597–13607.

Daas, P., & Puts, M. (2014a). Big data as a source of statistical information. *The Survey Statistician, 69*, 22–31.

Daas, P., & Puts, M. (2014b). Social media sentiment and consumer confidence. European Central Bank Statistics paper series No. 5, Frankfurt Germany.

Das, A. K., Sengupta, S., & Bhattacharyya, S. (2018). A group incremental feature selection for classification using rough set theory based genetic algorithm. *Applied Soft Computing, 65*, 400–411.

de Oliveira, D. N., & Merschmann, L. H. D. C. (2021). Joint evaluation of preprocessing tasks with classifiers for sentiment analysis in Brazilian Portuguese language. *Multimedia Tools and Applications, 80*(10), 15391–15412.

Devi, W. R., & Chingangbam, C. (2021). Sentiment Analysis for Electoral Prediction Using Twitter Data. In Emerging Technologies in Data Mining and Information Security (pp. 25-33). Springer, Singapore.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv: 1810.04805.

Díaz Galiano, M. C., García Cumbreras, M.Á., García Vega, M., Gutiérrez, Y., Cámara, E. M., Piad Morffis, A., & Villena Román, J. (2019). TASS 2018: The strength of deep learning in language understanding tasks. *Procesamiento del Lenguaje Natural, 62*, 77–84.

Díaz-Galiano, M. C., Vega, M. G., Casasola, E., Chiruzzo, L., Cumbreras, M. Á. G., Cámara, E. M., ... & Miranda-Jiménez, S. (2019b). Overview of TASS 2019: One More Further for the Global Spanish Sentiment Analysis Corpus. In IberLEF@ SEPLN (pp. 550-560).

Díaz Galiano, M. C., Martínez Cámara, E., García Cumbreras, M. Á., García Vega, M., & Villena Román, J. (2018). The democratization of deep learning in TASS 2017.

Duong, H. T., & Nguyen-Thi, T. A. (2021). A review: Preprocessing techniques and data augmentation for sentiment analysis. *Computational Social Networks, 8*(1), 1–16.

El Rahman, Sahar A., Alotaibi, F. A., & Alshehri, W. A. (2019). Sentiment analysis of twitter data. In 2019 international conference on computer and information sciences (ICCIS). IEEE, pp 1-4.

El-Halees, A. M. (2015). Arabic text classification using maximum entropy. *IUG Journal of Natural Studies, 15*(1).

Falkenauer, E. (1993). The grouping genetic algorithms: Widening the scope of the GA's. *JORBEL-Belgian Journal of Operations Research, Statistics, and Computer Science, 33* (1–2), 79–102.
Fan, Q., Bi, Y., Xue, B., & Zhang, M. (2022). Genetic programming for feature extraction and construction in image classification. *Applied Soft Computing, 118*, Article 108509.
Fernández Vítores, D. (2020). El español: una lengua viva. Informe 2019. Instituto Cervantes. https://cvc.cervantes.es/lengua/espanol_lengua_viva/pdf/espanol_lengua_viva_2019.pdf.
Forrest, S. (1996). Genetic algorithms. ACM Computing Surveys (CSUR), 28(1), 77-80.
García-Cumbreras, M., Martínez-Cámara, E., Villena-Román, J., & Morera, J. G. (2016). TASS 2015 - The evolution of the Spanish opinion mining systems. *Procesamiento de Lenguaje Natural, 56*, 33–40.
García-Cumbreras, M. A., Villena-Román, J., Martínez-Cámara, E., Díaz-Galiano, M. C., Martín-Valdivia, M. T. & Ureña López, L. A. (2016b). Overview of tass 2016. In TASS 2016: Workshop on Sentiment Analysis at SEPLN, pp 13-21.
García-Díaz, P., Sánchez-Berriel, I., Martínez-Rojas, J. A., & Diez-Pascual, A. M. (2020). Unsupervised feature selection algorithm for multiclass cancer classification of gene expression RNA-Seq data. *Genomics, 112*(2), 1916–1925.
García-Mendoza, C. V., Gambino, O. J., Villarreal-Cervantes, M. G., & Calvo, H. (2020). Evolutionary optimization of ensemble learning to determine sentiment polarity in an unbalanced multiclass corpus. *Entropy, 22*(9), 1020.
Gokalp, O., Tasci, E., & Ugur, A. (2020). A novel wrapper feature selection algorithm based on iterated greedy metaheuristic for sentiment classification. *Expert Systems with Applications, 146*, Article 113176.
Gondhi, N. K., Sharma, E., Alharbi, A. H., Verma, R., & Shah, M. A. (2022). Efficient Long Short-Term Memory-Based Sentiment Analysis of E-Commerce Reviews. Computational Intelligence and Neuroscience, 2022.
Gu, Y. H., Yoo, S. J., Jiang, Z., Lee, Y. J., Piao, Z., Yin, H., & Jeon, S. (2018, January). Sentiment analysis and visualization of Chinese tourism blogs and reviews. In 2018 International Conference on Electronics, Information, and Communication (ICEIC), pp 1-4. IEEE.
Hassan, S. U., Aljohani, N. R., Idrees, N., Sarwar, R., Nawaz, R., Martínez-Cámara, E., … Herrera, F. (2020). Predicting literature's early impact with sentiment analysis in Twitter. *Knowledge-Based Systems, 192*, Article 105383.
Hossain, M. A., Dwivedi, Y. K., Chan, C., Standing, C., & Olanrewaju, A. S. (2018). Sharing political content in online social media: A planned and unplanned behaviour approach. *Information Systems Frontiers, 20*(3), 485–501.
Hughes, G. F. (1968). On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory, 14*(1), 55–63. https://doi.org/10.1109/TIT.1968.1054102
Ikram, A., Kumar, M., & Munjal, G. (2022). Twitter Sentiment Analysis using Machine Learning. In *In 2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence). IEEE* (pp. 629–634).
Imran, M., Akhtar, A., Said, A., Safder, I., Hassan, S. U., & Aljohani, N. R. (2018, September). Exploiting social networks of Twitter in altmetrics big data. In STI 2018 Conference Proceedings (pp. 1339-1344). Centre for Science and Technology Studies (CWTS).
Iqbal, F., Hashmi, J. M., Fung, B. C., Batool, R., Khattak, A. M., Aleem, S., & Hung, P. C. (2019). A hybrid framework for sentiment analysis using genetic algorithm based feature reduction. *IEEE Access, 7*, 14637–14652.
Jagdale, J., Reha, A. Y., & Emmanuel, M. (2022). Sentimental Evaluation of Sensitive Tweets Using Hybrid Sentiment Analysis Model. In Proceedings of Third International Conference on Intelligent Computing, Information and Control Systems (pp 889-897). Springer, Singapore.
Jain, A., & Jain, V. (2022). Sentiment classification using hybrid feature selection and ensemble classifier. *Journal of Intelligent & Fuzzy Systems, 42*(2), 659–668.
Khosravinik, M. (2018). Social media techno-discursive design, affective communication and contemporary politics. *Fudan Journal of the Humanities and Social Sciences, 11*(4), 427–442.
Kwayu, S., Abubakre, M., & Lal, B. (2021). The influence of informal social media practices on knowledge sharing and work processes within organizations. *International Journal of Information Management, 58*, Article 102280.
Li, H., Chen, Q., Zhong, Z., Gong, R., & Han, G. (2022). E-word of mouth sentiment analysis for user behavior studies. *Information Processing & Management, 59*(1), Article 102784.
Liang, B., Yin, R., Du, J., Gui, L., He, Y., Yang, M., & Xu, R. (2021). Embedding Refinement Framework for Targeted Aspect-based Sentiment Analysis. *IEEE Transactions on Affective Computing*.
Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies, 5*(1), 1–167.
Madasu, A., & Elango, S. (2020). Efficient feature selection techniques for sentiment analysis. *Multimedia Tools and Applications, 79*(9), 6313–6335.
Madhu, S. (2018). An approach to analyze suicidal tendency in blogs and tweets using Sentiment Analysis. *International Journal of Scientific Research in Computer Science, 6* (4), 34–36.
Mhamed, M., Sutcliffe, R., Sun, X., Feng, J., Almekhlafi, E., & Retta, E. A. (2021). Improving Arabic Sentiment Analysis Using CNN-Based Architectures and Text Preprocessing. Computational Intelligence and Neuroscience, 2021.
Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, 26.
A. Mudinas D. Zhang M. Levene Market trend prediction using sentiment analysis: Lessons learned and paths forward 2019 arXiv preprint arXiv:1903.05440.
Naeem, M., & Ozuem, W. (2021). Customers' social interactions and panic buying behavior: Insights from social media practices. *Journal of Consumer Behaviour, 20*(5), 1191–1203.
Naseem, U., Razzak, I., Musial, K., & Imran, M. (2020). Transformer based deep intelligent contextual embedding for twitter sentiment analysis. *Future Generation Computer Systems, 113*, 58–69.
Ni, R., & Cao, H. (2020). Sentiment Analysis based on GloVe and LSTM-GRU. In 2020 39th Chinese Control Conference (CCC) pp 7492-7497. IEEE.
Nigam, K., Lafferty, J., & McCallum, A. (1999, August). Using maximum entropy for text classification. In IJCAI-99 workshop on machine learning for information filtering Vol. 1(1), pp 61-67.
Nilashi, M., Ahmadi, H., Arji, G., Alsalem, K. O., Samad, S., Ghabban, F., … Alarood, A. A. (2021). Big social data and customer decision making in vegetarian restaurants: A combined machine learning method. *Journal of Retailing and Consumer Services, 62*, Article 102630.
Nistor, S. C., Moca, M., Moldovan, D., Oprean, D. B., & Nistor, R. L. (2021). Building a twitter sentiment analysis system with recurrent neural networks. *Sensors, 21*(7), 2266.
Onan, A. (2022). Bidirectional convolutional recurrent neural network architecture with group-wise enhancement mechanism for text sentiment classification. *Journal of King Saud University-Computer and Information Sciences, 34*(5), 2098–2117.
Onan, A., & Korukoğlu, S. (2017). A feature selection model based on genetic rank aggregation for text sentiment classification. *Journal of Information Science, 43*(1), 25–38.
Osmani, A., Mohasefi, J. B., & Gharehchopogh, F. S. (2022). Sentiment classification using two effective optimization methods derived from the artificial bee colony optimization and imperialist competitive algorithm. *The Computer Journal, 65*(1), 18–66.
Passi, K., & Motisariya, J. (2022). Twitter Sentiment Analysis of the 2019 Indian Election. In IOT with Smart Systems. *Springer, Singapore, 2022*, 805–814.
Pennington, J., Socher, R., & Manning, C. D. (2014). October). Glove: Global vectors for word representation. In *In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543).
Pintas, J. T., Fernades, L. A. F.; Garcia, A. C. B. (2021). Feature selection methods for text classification: a systematic literature review. Artificial Intelligence Review, 2021, vol. 54(8), pp 6149-6200.
Poria, S., Hazarika, D., Majumder, N., & Mihalcea, R. (2020). *Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research*. IEEE Transactions on Affective Computing.
Rathika, J., & Soranamageswari, M. (2022). Intensified Gray Wolf Optimization-based Extreme Learning Machine for Sentiment Analysis in Big Data. In Evolution in Signal Processing and Telecommunication Networks (pp 103-114). Springer, Singapore.
Rui, W., Liu, J., & Jia, Y. (2016). Unsupervised feature selection for text classification via word embedding. In *In 2016 IEEE International Conference on Big Data Analysis (ICBDA)* (pp. 1–5).
Sachadev, J. S., & Bhatnagar, R. (2022). A Comprehensive Review on Brain Disease Mapping—The Underlying Technologies and AI Based Techniques for Feature Extraction and Classification Using EEG Signals. *Medical Informatics and Bioimaging Using Artificial Intelligence*, 73–91.
Salam, M. A., & Ali, M. (2020). Optimizing Extreme Learning Machine using GWO Algorithm for Sentiment Analysis. *International Journal of Computer Applications, 975*, 8887.
Sarker, A., Ginn, R., Nikfarjam, A., O'Connor, K., Smith, K., Jayaraman, S., … Gonzalez, G. (2015). Utilizing social media data for pharmacovigilance: A review. *Journal of biomedical informatics, 54*, 202–212.
Setya Rintyarna, B., Sarno, R., & Fatichah, C. (2019). Semantic features for optimizing supervised approach of sentiment analysis on product reviews. *Computers, 8*(3), 55.
Singh, N. K., Tomar, D. S., & Sangaiah, A. K. (2020). Sentiment analysis: A review and comparative analysis over social media. *Journal of Ambient Intelligence and Humanized Computing, 11*(1), 97–117.
Sohail, A. (2021). Genetic algorithms in the fields of artificial intelligence and data sciences. *Annals of Data Science*, 1–12.
Srivastava, R., Bharti, P. K., & Verma, P. (2022). A Review on Multipolarity in Sentiment Analysis. In Information and Communication Technology for Competitive Strategies (ICTCS 2020). *Springer, Singapore, 2022*, 163–172.
Starosta, K. (2022). Sentiment Analysis as a New Source of Information. In *Measuring the Impact of Online Media on Consumers, Businesses and Society. Sustainable Management, Wertschöpfung und Effizienz*. Wiesbaden: Springer Gabler. https://doi.org/10.1007/978-3-658-36729-9_4.
Tan, K. L., Lee, C. P., Anbananthen, K. S. M., & Lim, K. M. (2022). RoBERTa-LSTM: A Hybrid Model for Sentiment Analysis With Transformer and Recurrent Neural Network. *IEEE Access, 10*, 21517–21525.
Tan, L. K. W., Na, J. C., Theng, Y. L., & Chang, K. (2011). In *October). Sentence-level sentiment polarity classification using a linguistic approach* (pp. 77–87). Berlin, Heidelberg: Springer.
TASS-2017: Workshop on Semantic Analysis at SEPLN. Available: http://www.sepln.org/workshops/tass/2017/ (Last access in April 2022).
Trisna, K. W., & Jie, H. J. (2022). Deep Learning Approach for Aspect-Based Sentiment Classification: A Comparative Review. *Applied Artificial Intelligence, 2022*, 1–37.
Trivedi, S. K., & Singh, A. (2021). Twitter sentiment analysis of app based online food delivery companies. Global Knowledge, Memory and Communication.
Van den Brakel, J., Söhler, E., Daas, P., & Buelens, B. (2017). Social media as a data source for official statistics; the Dutch Consumer Confidence Index. *Survey Methodology, 43*(2), 183–210.
Valdivia, A., Luzón, M. V., Cambria, E., & Herrera, F. (2018). Consensus vote models for detecting and filtering neutrality in sentiment analysis. *Information Fusion, 44*, 126–135.

Vanaja, S., & Belwal, M. (2018). Aspect-level sentiment analysis on e-commerce data. In *In 2018 International Conference on Inventive Research in Computing Applications (ICIRCA). IEEE* (pp. 1275–1279).

Varghese, S., & Agrawal, M. (2021). Impact of Social Media on Consumer Buying Behavior. *Saudi Journal of Business and Management Studies (SJBMS), 6*(3), 51–55.

Vashishtha, S., & Susan, S. (2019). Fuzzy rule based unsupervised sentiment analysis from social media posts. Expert Systems with Applications, 2019, vol. 138, pp 112834.

Villavicencio, C., Macrohon, J. J., Inbaraj, X. A., Jeng, J. H., & Hsieh, J. G. (2021). Twitter sentiment analysis towards covid-19 vaccines in the Philippines using naïve bayes. *Information, 12*(5), 204.

Villena-Román, J., García-Morera, J., García-Cumbreras, M. A., Martínez-Cámara, E., Martín-Valdivia, M. T., & Ureña López, L. A. (2015). Overview of TASS 2015. In TASS 2015: Workshop on Sentiment Analysis at SEPLN, pp 13-21.

Wang, H., & Hong, M. (2019). Supervised Hebb rule based feature selection for text classification. Information Processing & Management, 2019, vol. 56(1) pp 167-191.

Wang, Z., Ho, S. B., & Cambria, E. (2020). Multi-level fine-scaled sentiment sensing with ambivalence handling. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 28*(04), 683–697.

Wang, Z., & Lin, Z. (2020). Optimal feature selection for learning-based algorithms for sentiment classification. *Cognitive Computation, 12*(1), 238–248.

Wouters, P., Zahedi, Z., & Costas, R. (2019). Social media metrics for new research evaluation. In *Springer handbook of science and technology indicators* (pp. 687–713). Cham: Springer.

Xie, X., Ge, S., Hu, F., Xie, M., & Jiang, N. (2019). An improved algorithm for sentiment analysis based on maximum entropy. *Soft Computing, 23*(2), 599–611.

Xue, L., Wang, H., Wang, F., & Ma, H. (2021, February). Sentiment Analysis of Stock Market Investors and Its Correlation with Stock Price Using Maximum Entropy. In International Conference on Intelligence Science (pp 29-44). Springer, Cham.

Yang, T., Yin, Q., Yang, L., & Wu, O. (2019). Aspect-based sentiment analysis with new target representation and dependency attention. *IEEE Transactions on Affective Computing*.

Zhang, M., Zhang, J., & Liu, L. (2021, December). Modelling Context with Graph Convolutional Networks for Aspect-based Sentiment Analysis. In 2021 International Conference on Data Mining Workshops (ICDMW) (pp 194-200). IEEE.

Zhao, M., Yang, J., Zhang, J., & Wang, S. (2022). Aggregated graph convolutional networks for aspect-based sentiment classification. *Information Sciences, 600*, 73–93.

Zhao, Z., Tang, M., Tang, W., Wang, C., & Chen, X. (2022). Graph convolutional network with multiple weight mechanisms for aspect-based sentiment analysis. *Neurocomputing*.

Zucco, C., Liang, H., Di Fatta, G., & Cannataro, M. (2018). Explainable sentiment analysis with applications in medicine. In 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 1740-1747). IEEE.

Zulqarnain, M., Ghazali, R., Aamir, M., & Hassim, Y. M. M. (2022). An efficient two-state GRU based on feature attention mechanism for sentiment analysis. *Multimedia Tools and Applications*, 1–26.