Expert Systems With Applications 223 (2023) 119862

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Review

A review on sentiment analysis from social media platforms

Margarita Rodríguez-Ibánez [a,*], Antonio Casánez-Ventura [b], Félix Castejón-Mateos [b], Pedro-Manuel Cuenca-Jiménez [b]

[a] Department of Business Economics, Universidad Rey Juan Carlos, Madrid, Spain
[b] Department of Signal Theory and Communications, Telematics, and Computing Systems, Universidad Rey Juan Carlos, Madrid, Spain

ARTICLE INFO

Keywords:
Sentiment analysis
Social media
Twitter
Causality
Temporal sentiment analysis
Professional and academic methodologies
Reproducibility studies
Causal rule predictions

ABSTRACT

Sentiment analysis has proven to be a valuable tool to gauge public opinion in different disciplines. It has been successfully employed in financial market prediction, health issues, customer analytics, commercial valuation assessment, brand marketing, politics, crime prediction, and emergency management. Many of the published studies have focused on sentiment analysis of Twitter messages, mainly because a large and diverse population expresses opinions about almost any topic daily on this platform. This paper proposes a comprehensive review of the multifaceted reality of sentiment analysis in social networks. We not only review the existing methods for sentiment analysis in social networks from an academic perspective, but also explore new aspects such as temporal dynamics, causal relationships, and applications in industry. We also study domains where these techniques have been applied, and discuss the practical applicability of emerging Artificial Intelligence methods. This paper emphasizes the importance of temporal characterization and causal effects in sentiment analysis in social networks, and explores their applications in different contexts such as stock market value, politics, and cyberbullying in educational centers. A strong interest from industry in this discipline can be inferred by the intense activity we observe in the field of intellectual protection, with more than 8,000 patents issued on the topic in only five years. This interest compares positively with the effort from academia, with more than 2,300 articles published in 15 years. But these papers are unevenly split across domains: there is a strong presence in marketing, politics, economics, and health, but less activity in other domains such as emergencies. Regarding the techniques employed, traditional techniques such as dictionaries, neural networks, or Support Vector Machines are widely represented. In contrast, we could still not find a comparable representation of advanced state-of-the-art techniques such as Transformers-based systems like BERT, T5, T0++, or GPT-2/3. This reality is consistent with the results found by the authors of this work, where computationally expensive tools such as GPT-3 are challenging to apply to achieve competitive results compared to those from simpler, lighter and more conventional techniques. These results, together with the interest shown by industry and academia, suggest that there is still ample room for research opportunities on domains, techniques and practical applications, and we expect to keep observing a sustained cadence in the number of published papers, patents and commercial tools made available.

1. Introduction

Since 2008, sentiment analysis has become an active research area according to an increasing number of published papers, as it can be observed using research databases such as Elsevier's ScienceDirect, IEEE Xplore Digital Library, °Springer Link, ACM Digital Library, or Wiley Online Library. As an example, from 2008 to 2022 the number of published papers that include the concept "sentiment analysis in social networks", grew at a geometrical rate of 34 % year on year. An in-depth study of a number of reviews of sentiment analysis with a special focus on social networks, showed us that until 2020, published revisions usually dealt with the following two main issues: on the one hand the techniques used, namely machine learning or lexicon-based methods; and on the other hand, on specific domains of applications such as emergencies, business intelligence, marketing, prediction of election results, etc. A scheme and classification of this structure of articles is presented in Fig. 1 for illustrative purposes.

An illustrative example is the publication by Ravi and Ravi Ravi &

* Corresponding author.
E-mail address: margarita.rodriguez@urjc.es (M. Rodríguez-Ibánez).

https://doi.org/10.1016/j.eswa.2023.119862
Received 5 August 2022; Received in revised form 8 March 2023; Accepted 10 March 2023
Available online 14 March 2023

Ravi (2015), who conducted a study summarizing over one hundred papers published between 2002 and 2015, focusing on the applications of sentiment analysis, the different approaches and open issues in the field. In 2016, Balazs and Velazquez Balaz & Velazquez (2016) highlighted the relevant value of the opinion mining and the fusion of information that can be found in sentiment analysis.

Rajalakshimi et al. in 2017 Rajalakshmi et al. (2017) focused on sentiment analysis methods, application domains and challenges. Also, subjectivity detection was analyzed by Chaturvedi et al. in 2018 Chaturvedi et al. (2018). And finally, Hemmatian and Sohrabi Hemmatian & Sohrabi (2019), in 2019, reviewed more specifically the classification techniques for opinion mining, sentiment analysis and characteristics extraction. Different from these surveys, Ashime et al. Ashima & Kumar (2020) in 2019 aimed to cover the significant and widespread approaches which are introduced in the field of sentiment analysis using deep learning – an extensive work including over 130 research papers, that provides a detailed survey of the most popular deep learning techniques at the time.

In 2020, Morone Birjali Birjali et al. (2021) and colleagues published a paper on the existing tools and a full inventory of the most common sentiment analysis techniques (machine learning, lexicon-based, hybrid and others), describing their advantages and disadvantages in detail. More specific was the approach by Garg et al. S. Garg et al. (2020), whose extensive guide of natural language processing focused on sentiment analysis on a Twitter dataset. We can also mention other papers, like the one focused on polarity in Twitter Singh (2020), the forecast on the opinion of the UK parliament and EU over Brexit Chandio & Sah (2020), or the predictions about the price return of nine cryptocurrencies Kraaijeveld & Smedt (2020).

As mentioned above, there have been many studies that have analyzed sentiment in written texts, particularly on social media. These studies provide a broad overview of the sentiment and perspective of users. However, this approach does not allow us to evaluate how sentiment changes over time or any causality relationships. Additionally, it may not accurately reflect the unique context and realities of the specific domains where the analyses are being applied. Many of these studies do not provide a thorough, comparative analyses of different techniques used by different authors, often because it is difficult to reproduce the methods. Lastly, few articles discuss how the industry approaches sentiment analysis, suggesting a disconnect between academic research and commercial practices. For these reasons, this paper proposes a different strategy that integrates and combines both an academic review of the literature and methods, as well as a vision of the applicability of these techniques in the industry.

Starting with the academic approach, temporal analysis, as described in the literature, allows for the formulation of relationships between observed sentiment and other underlying events or realities. This allows for a more effective use of the information on extracted sentiment. In other words, this dynamic analysis of sentiment enables us to understand how events produce changes on sentiment over time, and can provide insight on behavior patterns Preethi et al. (2015). This type of analysis has been successfully studied in various disciplines such as politics Gupta & Sandhane (2022), Park et al. (2021), suicide prevention in educational systems Yu et al. (2020), cyberbullying Chatterjee & Das (2020), and customer service Sharuee et al. (2021), among others. In addition to examining the relationship between events and sentiment at different points in time, temporal analysis can also be combined with spatial analysis, allowing for a more comprehensive understanding of the relationships between sentiment and other factors in a given area over time.

Causality is another type of analysis that examines the relationship between events that occur over time. Granger's causality is the most common approach found in the literature. It involves testing whether the results of one variable can be used to predict another variable, and whether the relationship between the variables is unidirectional or bidirectional. To use Granger's causality, it is necessary to compare and evaluate the current and past behavior of a time series (A) to determine whether it can predict the behavior of another time series (B). Many

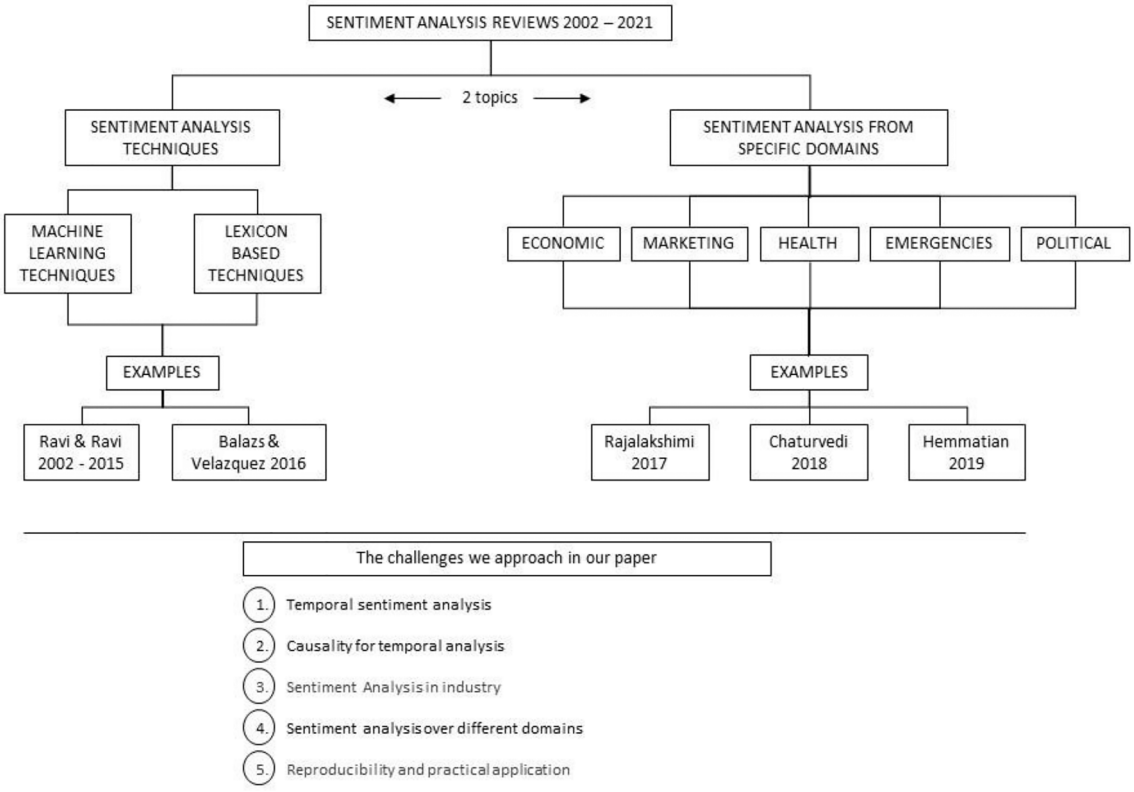

Fig. 1. Sentiment Analysis reviews and new challenges.

researchers use predictive causality in sentiment analysis, and very frequently in stock market analysis Granger (1969). One of the advantages of using Granger's causality is that it allows for the inclusion of exogenous variables, in analyses that otherwise could only be considered to include intrinsic information. However, it is important to carefully consider the type of exogenous variables being included to improve performance without creating noise in the final conclusion.

Regarding the commercial application of sentiment analysis techniques, to our knowledge, there are no scientific publications that incorporate this aspect in their studies or summaries. However, these techniques have proven successful and there are a good number of applications available, especially with the following purposes: a) social media management, b) monitoring and active listening and c) analysis of what is published. Organizations like SAS, SAP, Microsoft and Google are examples of business organizations that have built these kinds of tools Nazir et al. (2022). In this work, we elaborate on the different tools and information available on this topic.

Another question of interest, beyond the techniques and functionalities used, could be, what are the disciplines or domains where this type of analysis is of growing interest and being applied with greater frequency? To answer this question, we collected 2,306 papers related to sentiment analysis from 2008 to 2021, and we scrutinized the top six domains we found in the academic literature. This evaluation included not only this domain clustering, but also the trends, the types of publications and the techniques used in each domain.

The main motivation of our survey is to complement the rest of the scientific publications on sentiment analysis by focusing on issues that we believe may have been treated superficially in other works. We have conducted a systematic, detailed and thorough study on the temporal analysis of sentiment analysis which shows an important relationship with other non-temporal variables, such as location. This research survey also provides a new perspective on the causality of temporal sentiment analysis in social media, which we believe should stimulate further discussion and consideration on how to improve future studies in this area. A novel part of this work that we are planning to continue in future research is a section that aims to offer a practical vision regarding the applicability or reproducibility of some of the existing and upcoming techniques. We pay special attention to promising research directions such as zero-shot inference using large language models, which despite showcasing impressive result can be challenging to apply in practice. Finally, an in-depth overview of the domains where sentiment analyses has been used also contributes to a better understanding of the applicability of these techniques.

This paper is structured as follows: First, we describe the temporal dynamics of sentiment analysis. Second, we devote a section to review contributions on causality to sentiment analysis in social media. Third, we summarize the different commercial sentiment analysis software tools used in industry and then we evaluate the different domains where sentiment analysis is currently being used, and the methodologies applied for each one of them. We follow with practical considerations on reproducibility and applicability of state-of-the-art methods, and conclude with a brief discussion.

2. On the temporal dynamics of sentiment analysis

Textual data from sources and tools such as social media, review websites, blogs, forums, and interview transcripts is requiblue for the initial stage of sentiment analysis Birjali et al. (2021). Furthermore, as previous research has shown, combining sentiment analysis with temporal dynamics can yield important outcomes. Temporal sentiment analysis is useful for summarizing people's feelings about events (or any other issue) in relation to time (i.e., when they happened) Preethi et al. (2015).

In the context of health, a recent study has shown the temporal patterns in emotional fluctuations during the COVID-19 pandemic, measuring changes in citizens' sentiments during different periods in every single day for over 4 consecutive months Yu et al. (2021). They achieved this by dividing the day into six groups: early morning, morning, afternoon, evening, night, and late night. They categorized all of the general sentiment data by day and hour, and then counted the frequencies of each category. They used heat maps to show the distribution of each emotion across time in order to better understand the temporal distribution of emotions. Furthermore, the temporal distribution data for intradays was displayed in a matrix diagram.

In the field of politics, authors in Park et al. (2021) studied the temporal dynamics of emotional appeals in Russian campaign communications during the 2016 election. They questioned, "How does emotion in Facebook ads and Twitter affect the sentiments of users and their online comments over time?". They followed a methodology that combined word-level sentiment analysis with natural language processing (NLP) techniques.

In the education sector, Yu et al. Yu et al. (2020) analyzed the communication patterns of learners, focusing on the temporal analysis of their sentiments on public (Twitter) versus private (MS Teams) platforms. They used a pre-trained Bidirectional Encoder Representations from Transformers (BERT) model Devlin et al. (2019) to study the aggregation of sentiment in a time window. Cyberbullicide is another of the topics that have been studied using temporal sentiment analysis. In Chatterjee & Das (2020), they conducted temporal sentiment analysis of data from social media to achieve early detection of cyberbullying and suicide ideation of a victim by using graph-based data mining approaches.

A perhaps trite area, but no less important, is product evaluation. Sharuee et al. Sharuee et al. (2021) sought to formalize a chronological sentiment analysis of product reviews. They used the automatic contextual analysis and ensemble clustering (ACAEC) algorithm. ACAEC is a clustering algorithm which utilizes contextual analysis and clustering ensemble learning. They performed chronological sentiment analysis using window sequential clustering (WSC) and segregated window clustering (SWC). WSC is a dynamic analysis, whereas SWC is solely based on the temporal characteristic of reviews. The results show that WSC and SWC are competitively accurate compared to supervised methods. Temporality in real time sentiment analysis is achieved by formulating rules based on metadata as well as the linguistic context of words. An example of work in this area is Kaur & Mohana (2019), who carried out temporal sentiment analysis based on *meta*-data in conjunction with the linguistic context of words. They developed an algorithm for evaluating Tempo-Sentiscore.

A new research domain is, "spatio-temporal data mining", which aims to analyze spatial and temporal data. Observing the temporal sentiment changes in different locations helps both to examine the emotional changes in the locations and to understand the thoughts of the social media (or any other source) users in these locations. In this respect, Ecemis et al. Ecemis et al. (2021) used a method called *Temporal Sentiment Analysis of Socially Important Locations* (TS-SIL) to discover socially important locations from an aggregated Twitter dataset, and performed sentiment analysis using a dictionary-based approach and machine learning algorithms. Hu et al. Hu et al. (2020) explored local users' sentiments extracted from Geo-tweets data during one year and analyzed them from a spatial–temporal perspective.

Finally, in the area of event prediction, Preethi et al. Preethi et al. (2015) presented a prediction model based on the temporal sentiment analysis of tweets in order to identify the causal relation between events and the intervals between them.

3. On the causality of temporal sentiment analysis

We have seen in the preceding sections that current methods allow us to create signals with time by aggregating sentiment in tweets, according to some index evolving with time. We can denote this sentiment signal $s(t)$, and it is assumed to evolve with time. When this sentiment time signal on some topic is scrutinized in the context of some other

measured variable of interest, which we will denote here by $y(t)$, one of the most recurrent research hypotheses is to determine to what extent the sentiment signal is actually affecting the signal of interest. A typical example would be to determine whether the sentiment signal ($s(t)$) condensed from the short messages related to some given company could affect the stock market value of said company ($y(t)$). Note that other exogenous variables can be present, and for notational simplicity we will consider just one exogenous variable, and denote it by $x(t)$. We obviate the sampling nature of all the data for notational simplicity here.

Obviously, the correlation between the time samples of the sentiment signal and the signal of interest is not statistically solid enough from an stochastic-process point of view, and even from a practical point of view. The most useful way would be to establish the causality of the sentiment on the signal of interest. It may not be surprising that a really vast amount of works about sentiment analysis in the last years have leveraged on the concept of causality in the sense of Granger from econometrics Granger (1969). A time series is said to be Granger-cause of a time series of interest if it can be shown that the former provides statistically significant information about future values of the second.

Granger causality is established in terms of a statistical hypothesis test with the null hypothesis that $s(t)$ does not Granger-cause $y(t)$, where $s(t)$ and $y(t)$ are stationary time series. By considering a set of time lags M, the mathematical method described by M. Chvostekova (2019) first adjusts a univariaté autoregression model for the signal of interest, which is,

$$y(t) = a_o + \sum_{m=1}^{M} a_m y(t - mt_o) + e_o(t) \tag{1}$$

and then the model is completed with N lagged samples of the sentiment signal $s(t)$,

$$y(t) = a_o + \sum_{m=1}^{M} a_m y(t - mt_o) + \sum_{n=1}^{N} b_n s(t - nt_1) + e_1(t) \tag{2}$$

where t_0 and t_1 are basic lags for each signal and $e_0(t)$, $e_1(t)$ are the noise residuals of the regression. Note that exogenous variables can be readily included in the preceding data model, by just using a vector autoregressive model. The null hypothesis that $s(t)$ does not Granger-cause $y(t)$ is accepted if and only if no lagged values of $s(t)$ are retained in the regression S. Zhao et al. (2016). Note that the involved time series need to be stationary, and that they have to be previously transformed otherwise. We should keep in mind that Granger causality in econometrics is understood with the sense of *predictive causality*, rather than Physics causality. Granger causality has been the most widely used technique to test causal relationships among time series Hu et al. (2020), and it has been paid increasing attention in the last 10 years, playing an important role in understanding the behavior of time series processes in many different fields, such as stock market prediction in economics Gujarati & Porter (2009), genetics Zhu et al. (2010), climate studies Elsner (2007), and neuroscience Gao et al. (2011), to name just a few of them.

There is a large number of research studies that have used this predictive causality in sentiment analysis. Some representative research efforts are compiled for the prediction of stock market trends using social media data through Granger causality tests, as this is one of the most active application areas. Early works can be retrieved, as the first proposals scrutinizing the correlation and the prediction capabilities of investor sentiment and the Shangai Composite Index using vector multivariate causality data models Zheng & He (2010).

The basic statement of Granger causality tests has evolved with time, as well as the scope and specificity of the studies. As an example, the relationship between investor sentiment and stock prices in the Growth Enterprise Market Finger Bar of Eastmoney Zhang et al. (2020) considered scenarios where noise traders and rational arbitrageurs impact on markets are subject to theoretical debate, such as semi-efficient markets in China. By using Bayesian data models relating stock prices with their investor bullish index, using labels in 6000 posts, they determined that the forums affect the price over time, whereas the price affects the forums in the short run. We would like to issue a word of caution at this point, with respect to this and in other works where they are using only sentiment signals for prediction.

Other works have aimed to improve the flexibility of the statistical learning data model used as the basis for the causality tests by considering nonlinear predictors, rather than linear regression Marinazzo et al. (2008); J. Park et al. (2017, 2019). A nonlinear approach to Granger causality using Support Vector Machines for prediction Preethi et al. (2015) used these learning models in comparison with linear model to predict the stock market using sentiment information about four companies (Apple, Google, Amazon, and Microsoft). They showed that larger lags were more significant for nonlinear than for linear data models, and that incorporating sentiment signals in these conditions to nonlinear models increased the prediction accuracy for stock prices. The same work determined that a minimum number of 2500 of messages per day and company is desirable to reduce the statistical fluctuations due to sentiment under sampling.

The huge amount of works on sentiment and stock market causality has brought theoretical controversy and strong diversity of methodological analysis. A detailed study in S. Bouktif et al. (2020) aimed to determine which is more consistent with the data observations between the two poles inherited from old theories, either (a) the hypothesis that stock markets react instantaneously to any given news (prediction advocates), or (b) the random walk hypothesis that prices are determined randomly and stock prediction is impossible. In 10 influential companies of NASDAQ, several advanced analysis tools were considered, such as the use of Latent Dirichlet Analysis for tweets corpus authentication, use of N-grams, stationary transformations, algorithmic feature selection, and a variety of nonlinear models. Statistically significant causality was obtained for different lags (which were different for different companies). Authors conclude halfway point from the data, as stock prices of some companies are more susceptible to public sentiment than others, and they emphasis the use of appropriate processing techniques.

The vast literature on sentiment causality on stock markets is currently becoming very advanced, as pointed out S. Bouktif et al. (2020), starting with the consideration of which techniques are better for prediction of stock markets. Whereas evidence exists that sentiment increases predictability in multivariate data models, it also seems to heavily depend on the company context (domain, volumes, or origins of posts). The adequate representation of sentiment signals in stock market prediction has been shown to be a key factor, though not the only one, and specific aspects clearly matter. Other application fields of causality based on sentiment signals can surely benefit from the lessons learnt so far in the stock market prediction arena.

4. Sentiment analysis in industry

Sentiment analysis in social networks is a rapidly growing field that extends beyond academic interest, reaching companies and organizations that have found in social networks new ways to better manage their marketing and communication policies. The applications created to analyze social networks have evolved to the point where today they enable companies to carry out two key functions: to obtain information of great interest about their market and customers and to communicate commercial and marketing information, effectively. From this perspective, there are many commercial tools available today, either for purchase or free, that allow companies to approach this new form of bidirectional communication by making intensive use of social networks. Depending on the target market, very large companies have traditionally opted for customized developments and powerful tools that are often provided in conjunction with consulting services. According to the Gartner Group, the main players in this field are IBM, SAS, Microsoft, and SAP Gartner Group Inc. (2021); Blanco (2021). To get a real

perspective of the importance and market potential of these types of solutions, we only have to look at how much IBM invested in developing its Watson tool: over 3 billion dollars! McKinsey & Co (2018a), McKinsey & Co (2018b). So, it is not surprising that the great expectations for, and the potential value of this market, have meant that today, not only large companies can make use of these tools, but also a large number of entrepreneurs have made their initiatives available to the public. So that now there are more than 200 solutions on the market, including applications, APIs, and tools, that offer free, freemium and paid alternatives to entities of all shapes and sizes G2.com (2022). In an increasingly developed market McKinsey & Co, 2018a; McKinsey & Co, 2018b; G2.com (2022), the different social networking applications for companies offer a wide range of services that end up being consolidated in Social Media Suites (SMS). These SMS incorporate tools that provide extensive coverage for the different needs of companies, such as: (i) Managing, monitoring and analyzing information related to one or more social media accounts and enabling marketing and communication teams to manage marketing campaigns including post automation, community engagement and account integration across media, (ii) Monitoring listening, tracking and information gathering on social media channels. (iii) Aggregating and analyzing to measure the effectiveness of social campaigns, sales, marketing and customer service. But not all tools on the market manage to provide all of these services, which is why we can classify them according to which of the 3 following functions they offer: (a) the management of social networks, (b) monitoring and active listening, and (c) the analysis of what is published.

Management-focused tools, or so-called *social media management software*, are the most abundant and the IT reference consolidation website G2.com lists more than 270 of them. They are often used by social media, marketing and communications departments to increase brand awareness and develop new business. Such tools provide customers, partners and suppliers with essential communication services through activity automation. The monitoring software tools include functionalities for listening, tracking and inventorying content. These products are again used by marketing and communications departments to identify trends, track competitors and understand customer sentiment. Artificial intelligence and other techniques have made it possible to use these types of tools to extract information (not just content), from social networks that can be used to effectively improve company policies and bring them real value. The number of tools that offer these types of functions is somewhat smaller than the number of management focused tools, but still high: more than 200 according to G2.com. This is because sometimes they include and count not only fully functional tools, but also APIs developed for this purpose that can be used in conjunction with other solutions. Among other functions, these tools provide: active listening in networks; identification of trends, detection of customer sentiment; customer characterization and classification and identification of opinion leaders. Finally, the most sophisticated set of tools are those that are enabled by analytic software. These, with a higher level of technical sophistication, provide detailed tracking and cross-analysis with other sources, audience evaluation, tracking / engagement, sentiment analysis, statistical consolidation of information and comparative analysis of campaigns, publications, and content. In this case, the number of programs claiming to offer these types of services, according to G2.com, is over 200.

From the perspective of the techniques used by the various platforms or social network analysis tools, proprietary companies do not provide detailed information on the techniques, algorithms or dictionaries they use when performing their analysis. On the other hand, there are some aggregators and comparators that analyze the results achieved by a limited number of these tools. These studies are rarely published in academic environments, though they are made public with some frequency in congresses. From the published studies to which the authors of this paper have had access, the information seems to confirm a commonly held opinion (consensus) that there are significant differences in the results of sentiment analysis produced by these tools which are related to the different techniques and vocabulary used Abbasi et al. (2014).

In addition to the large number of applications and investments of large firms in this discipline, the dynamism of the sentiment analysis sector and the activity in social networks is also reflected in the number of patents related to sentiment analysis. To evaluate it, authors executed different queries on Google Patents, containing the terms"sentiment analysis" and"social networks". The query carried out was: "Sentiment Analysis" & "Social Network". The inclusion criteria for the patents to be assessed was the year of granting, between 2016 and 2021 and corresponding to the technologies included in the study, while the exclusion criteria were that the patents were pending granting. The queries turned up in 8349 published patents registered by more than 3,000 different entities between the years 2016 to 2021 (See Fig. 2). It is additionally noteworthy that only 14 of these 3,000 + entities owned 25% of the patents, offering a clear picture of the heavy investment of these 15 entities in the field. The first of these was Google itself, with 447 patents, followed by Facebook, with 283 patents, Microsoft with 260, Apple with 217, IBM with 149, Samsung with 136, Oracle with 105, Commvault with 101, Amazon with 86, Visa with 68, Tencent with 61, One Trust with 61, Salesforce with 60, and Intel with 54.

5. Sentiment analysis over different domains

As stated before, sentiment analysis in social networks is a growing field of study that continues to generate interest beyond the academic community due to its wide range of applications and the ease of obtaining data from social networks like Twitter, Facebook, or Instagram. This type of analysis can be used for many purposes, including social.

studies, epidemiological analysis, and market research. To assess the interest generated by the different disciplines in sentiment analysis, an exhaustive search was carried out in the Scopus database on 10 topics between the years 2007 and 2021. The query was carried out on February 15, 2022.

The inclusion criteria for this query was: have the descriptors of Sentiment Analysis, Social Network and technologies associated with these concepts, be written in English, be Articles, Book Chapter, Conference Paper or Reviews and published between 2007 and 2021, both inclusive.

More concretely, we searched the Scopus website using a query with the terms"Sentiment Analysis" and"Social Network" restricted to the keywords"machine learning","artificial intelligence","neural networks","deep learning","NLP", "Support Vector Machines", "Autoencoders", "Causality", "Bayesian", "Logistic Regression", "Optimization", "AI", "Statistical Learning", "Self-organized Networks", "Decission trees", "PCA", "ICA" and "Feature selection". The specific query carried out in the scopus database was as follows: TITLE-ABS-KEY (sentiment AND analysis, AND social AND networks, AND (machine AND learning) OR (artificial AND intelligence) OR (neural AND networks) OR (deep AND learning) OR (nlp) OR (support AND vector AND machines) OR (autoencoders) OR (causality) OR (bayesian) OR (logistic AND regression) OR (optimization) OR (ai) OR (statistical AND learning) OR (selforganizing AND networks) OR (decision AND trees) OR pca OR ica OR (features AND selection). This search was applied to Title, Abstract and keywords, identifying 2306 articles, book chapters, contributions to conferences or reviews. No paper was rejected as the scope was very restricted. The elements were downloaded in csv format from the Scopus search engine and processed in Excel, classifying them by domains and techniques used filtering by associated concepts in title, abstract and keywords. Fig. 3 shows the growth of the different topics over the years, while in Fig. 4 we can see the distribution per year.

Among the different domains, marketing was found the most reproduced discipline with 613 publications. Following this one, we found significant increase in the politics domain (459 publications in the period), economics (200 publications), Health and medicine (196) and

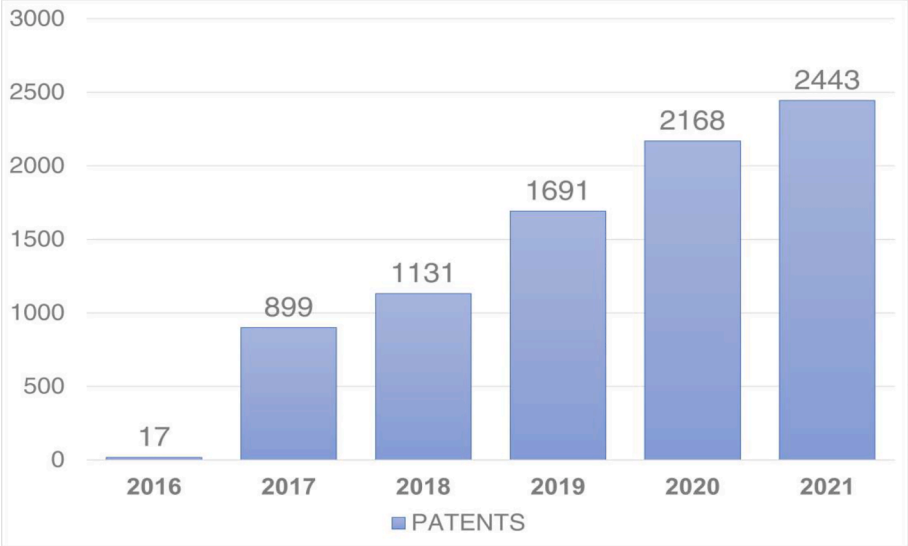

Fig. 2. Issued patents from 2016 to 2021 related to Sentiment Analysis in Social Networks.

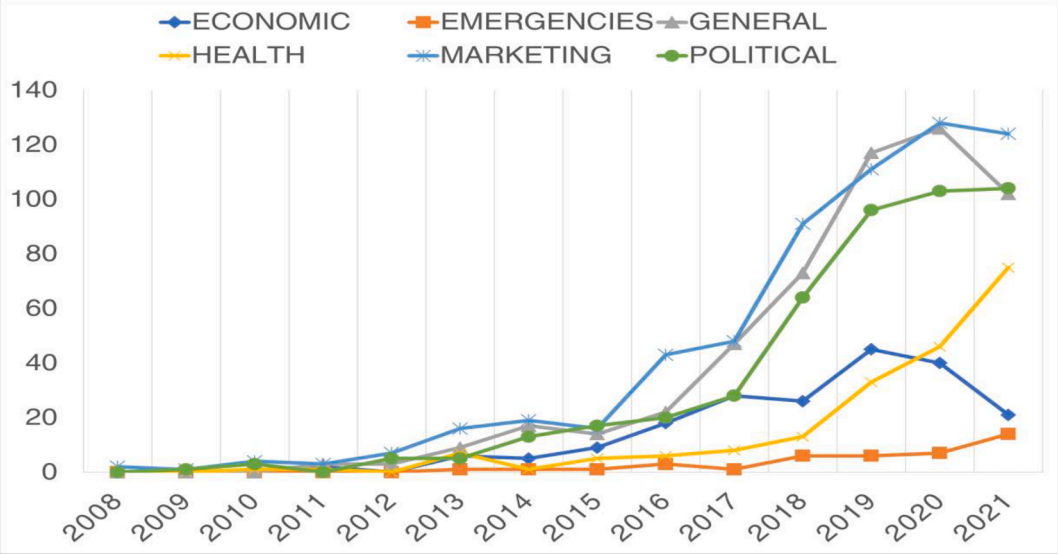

Fig. 3. Sentiment Analysis in Social Networks publications published between 2008 and 2021, according to Scopus.

finally emergencies and disasters, with 40 publications. The rest of the publications, counting up to 533 publications were classified as General, because they do not fall into a specific domain, but rather analyze various technologies.

Under a temporal perspective, the maximum number of publications occurred in 2020, with 484, followed by 2021 (462) and 2019 (443). Entering into a deeper analysis, political domain studies are carried out mainly in relations with elections, political parties, hate and specific processes in certain countries. Health domain studies included disease perception, public perception of vaccinations, mental illness, vaccination acceptance, suicide and in 2021 the pandemic caused by COVID 19. Marketing domain studies were very much related to customer perception, sales,

movies, music and specific markets, while economics domain contributions deal with social networks, stock market, social responsibility, brand awareness, commodities and cryptocurrency. Finally, the domain of emergencies and disasters deal with emergency management and intervention in natural and technological disasters.

As we see in Fig. 5, in terms of Feature Extraction, the use of lexicons and other classic NLP techniques stands out throughout the period as the most relevant methods, followed by embeddings and transformers-based methods such as BERT, RoBERTa or BERTweet (especially in recent years), and other technologies such as Bag of Words, Word2Vec, N-Grams and Tokens. New state-of-the-art auto-regressive and large language models (T5, T0++, GPT-3, GPT-J) are rarely used in the evaluated publications. As far as processing technologies are concerned (see Fig. 6), Machine Learning (ML), Deep Learning, SVM and Bayesian analysis, are frequently present during the period under observation. At a much lesser extent we also find logistic regression and decision trees. More specific methods such as self-organizing maps, causal analysis or autoencoders, are scarce. If we focus on the breakdown by document type, we can see that the mix broadly generalizes over the period (see Fig. 7).

On the other hand, no significant relationship was observed between the domains and the techniques used. Instead, we observed a homogeneous deployment of all technologies over the different domains (see Fig. 8), if we except the domains where the number of contributions was not significant enough to represent all techniques. For this reason, in

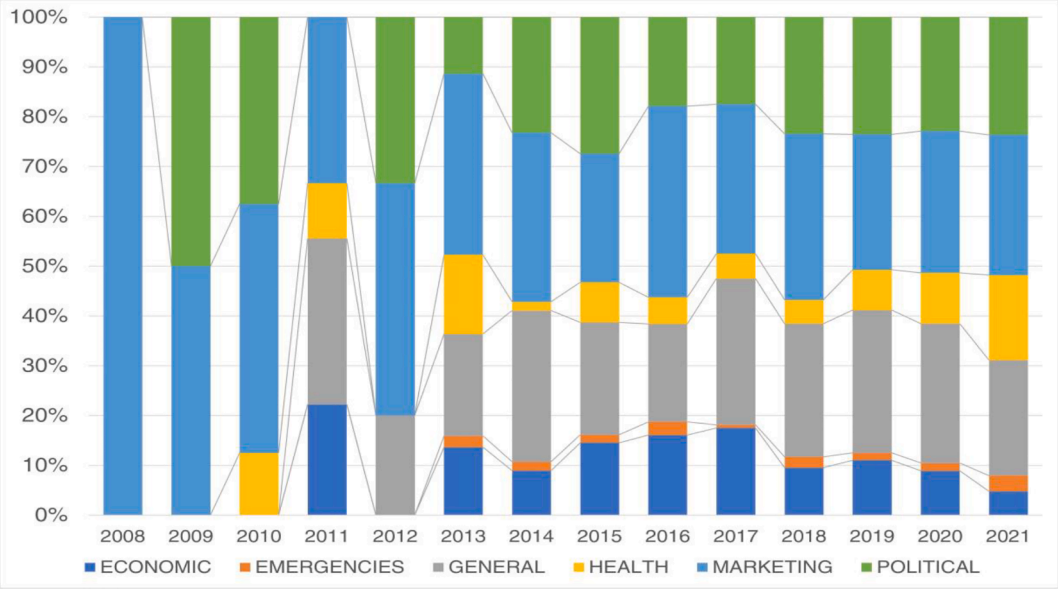

Fig. 4. Distribution of papers from 2008 to 2021 related to Sentiment Analysis in Social Networks obtained from Scopus DataBase.

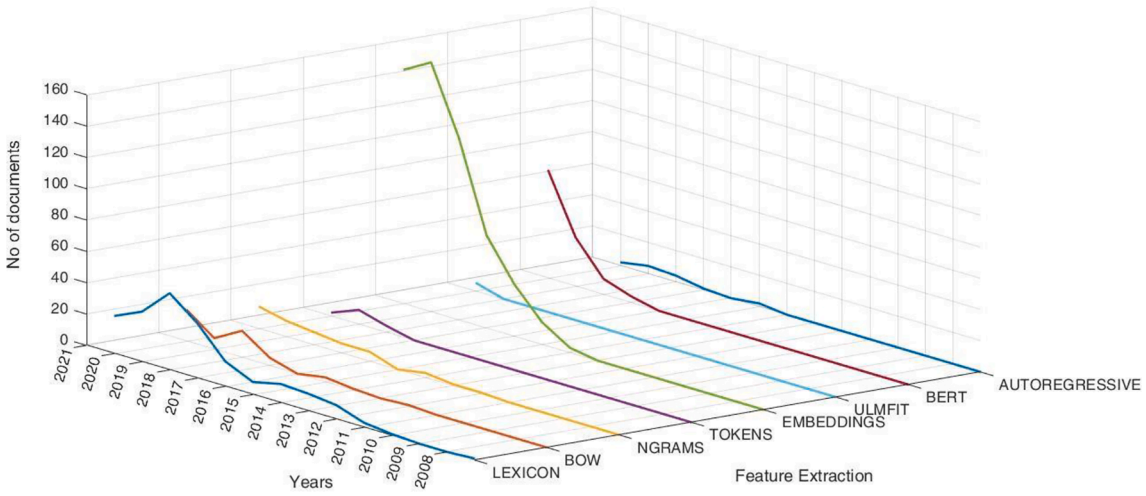

Fig. 5. Feature Extraction Technologies used in documents from 2008 to 2021 related to Sentiment Analysis in Social Networks obtained from Scopus DataBase.

marketing and health almost all the technologies were detected; but just a few are depicted in the emergency domain.

6. On the reproducibility and practical application

In previous sections it became apparent that sentiment analysis, and in particular sentiment analysis on social media such as Twitter, is a rich field with many applications that fulfill a variety of goals across industries. However, from the point of view of the practitioner or researcher, there are a myriad of methods with different trade-offs, and it is hard to determine what is the best method to apply to a particular situation. Research literature tends to focus on hard performance figures that improve the previous state-of-the-art, but this quest increasingly comes at the cost of ever bigger and more expensive models that are usually outside the reach of most organizations.

Even if we just focus on dry performance metrics and ignore the trade-offs to achieve them, it is usually difficult to compare methods, as they are measured on different tasks and datasets.

In this section we explore some of the most representative method families that are available in the practitioner's toolbox, and examine the current existence (or lack thereof) of performance and usage comparisons against competing methods. We provide some basic intuition on complexity and cost, but we conclude that more work needs to be done in this area to fully understand the advantages and limitations of all the different techniques.

6.1. Sentiment analysis methods: An evolution towards large transformer models

The history of sentiment analysis is the history of language classification, which in turn is tied to the progress achieved in the NLP field. We find it useful to group method families according to the major breakthroughs in NLP, as those breakthrough techniques illustrate important paradigm changes in the way text corpora are being processed and evaluated. In addition, all the major methods discussed in this summary can still be used for modern problems, and the timeline at which they appeared is a rough proxy for the complexity of the method. Users looking for adequate performance with fast inference times and reasonable costs may still apply proven methods such as SVM Cortes & Vapnik (1995) classification of word embeddings; those users whose

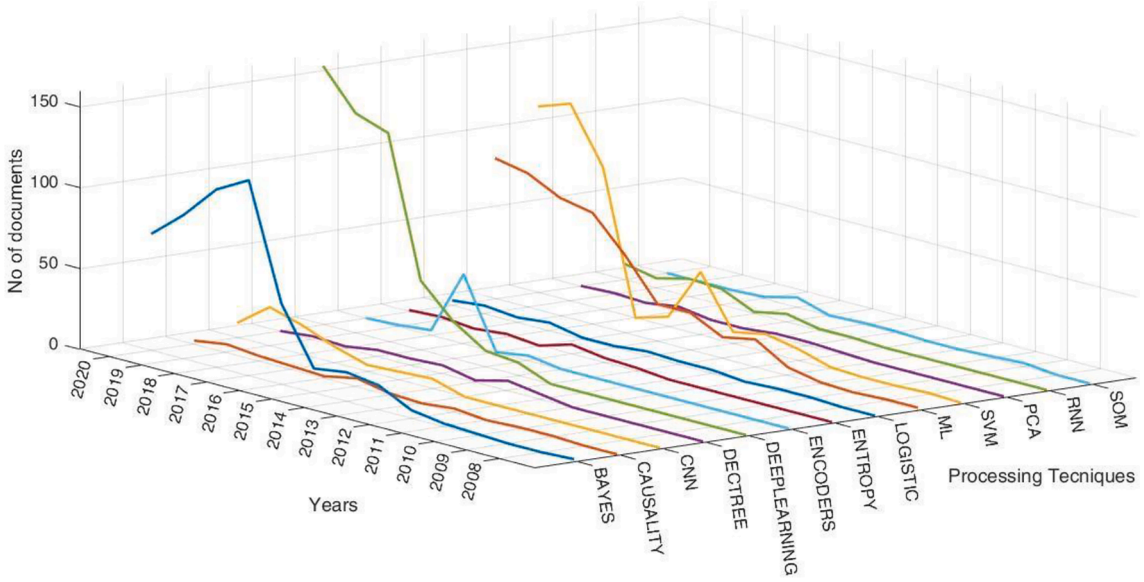

Fig. 6. Feature Processing Technologies used in documents from 2008 to 2021 related to Sentiment Analysis in Social Networks obtained from Scopus DataBase.

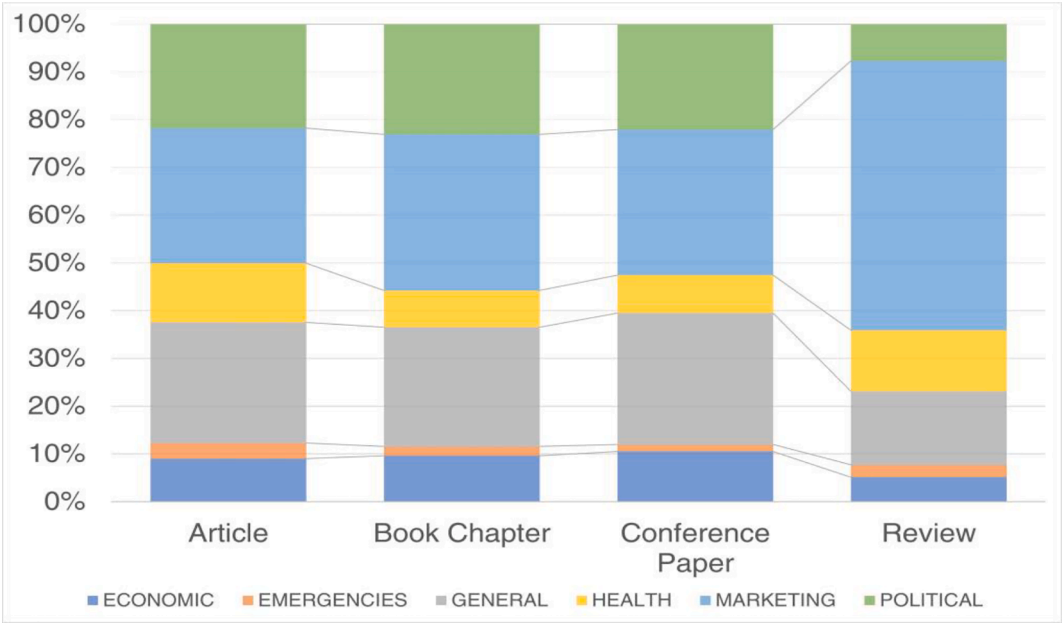

Fig. 7. Document type by domain from 2008 to 2021 related to Sentiment Analysis in Social Networks obtained from Scopus DataBase.

goals demand the best performance possible would have to look into the more modern, complex and expensive techniques.

Table 1 shows a summary of some of the most important milestones in the progress of NLP, and how they translate to the problem of sentiment classification.

The first group of methods (*(a) in* Table 1) is what we consider to be the classic era of NLP. Text processing was heavily based on language knowledge, coupled with extensive use of rules and heuristics to make the problem tractable. For example, the use of *stop words* (that are essentially ignored) is frequent to reduce complexity, as is the use of stemming and lemmatization techniques to simplify the vast and rich landscape of human utterances. These rules have to be carefully crafted and adapted to the problem at hand, striking a balance between the expressiveness.

of the representation and the computing complexity. These techniques drew a major influence from the statistical and probabilistic approach coming from Information Theory, where text was separated into n-grams and the most probable successor to a given sequence of n-grams could be computed, extracting language knowledge from those distributions *(b)*. Techniques such as tf-idf Sammut & Webb (2010) made it possible to successfully apply computation techniques for complex tasks such as document and information retrieval, or search engines.

A major step in the evolution of NLP was the use of embedding vectors *(c)*, which in our opinion marks the *first coming* of neural networks for NLP. Neural networks are not directly applied to make predictions on text; instead, they are trained to map language terms to vectors in a multidimensional vector space by leveraging co-occurrence matrices, and are therefore used as a tool to simplify the construction of high-dimensional vector spaces. These vector spaces usually consist of a few hundred dimensions (300 being a typical value), as larger dimensionality yields diminishing returns. One of the major advantages of

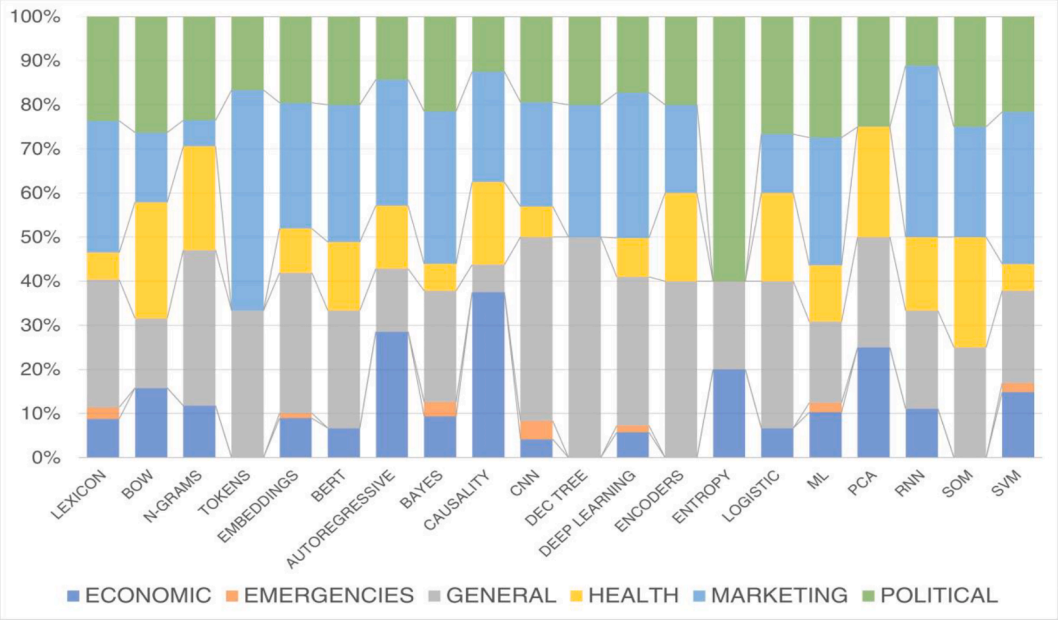

Fig. 8. Technologies used by domain from 2008 to 2021 related to Sentiment Analysis in Social Networks obtained from Scopus Database.

Table 1
Schematic overview of historic NLP methods and techniques.

Ref.	Feature Extraction / Representation	Usual Classification Strategy
(a)	Bag of words, dictionary	Sum of per-word polarities. Can be assigned using domain knowledge.
(b)	n-grams, subword tokens	Same.
(c)	Embeddings: GloVe, word2vec, fastText	SVM (typical), and others.
(d)	LSTM networks	Built-in (dense layer).
(e)	ULMFiT (language model fine-tuning)	Built-in. Can be used as a feature extractor only.
(f)	Transformer-based encoders (BERT, RoBERTa)	Built-in during fine-tuning, or feature extraction.
(g)	Auto-regressive models: GPT-3, GPT-J	Zero-shot / few-shot prompts.
(h)	Encoder-decoder transformers (T5, T0++)	Zero-shot / few-shot prompts.
(i)	Dense embeddings from auto-regressive models	Classification methods on extracted vectors.

embeddings was that similar or related concepts were placed close together in the embedding space, and it was possible to find related words or even apply language transformations by performing vector operations in the embedded space. Examples of these systems are word2vec Mikolov et al. (2013) (Google, 2013), GloVe Pennington et al. (2014) (Stanford, 2014), or fastText Bojanowski et al. (2016) (Facebook, 2016). From the point of view of text classification, solutions are achieved by applying classification techniques, such as SVM and others, to the embedded representations Joulin et al. (2016).

The use of neural networks and deep learning never left the field of NLP, and it has been a major source of innovation ever since. The next step was the use of Recurrent Neural Networks (RNNs) as a natural way to process text sequences, introducing the capability for the network to acquire or "remember" some context about previous fragments. RNNs come in a variety of forms, and specialized layers such as LSTM *(d)* Hochreiter & Schmidhuber (1997) or GRU Cho et al. (2014) are frequent. These techniques were quickly adopted by industry in assistants like Alexa. The text classification task can be addressed by incorporating a dense layer at the end of the text-processing network, and train that layer, together with the rest of the network, to classify source sentences or paragraphs into an arbitrary set of classes.

The next major and very significant milestone appeared in around 2018 *(e)*, when it was shown that large language models could be pre-trained and fine-tuned Howard & Ruder (2018). Just like what had happened with their computer vision cousins a few years before, language models were now starting to achieve state-of-the-art performance in a variety of downstream tasks. The key contribution was that the lengthy, complex and expensive task of pre-training a large language model, which requires vast amounts of text, can be performed once. The pre-trained network can then be adapted or fine-tuned to concrete tasks (such as classification), on specific corpora (such as Twitter messages) when necessary. Even though pre-training is very expensive, fine-tuning is much easier, requires less data and is therefore feasible for small organizations. Similarly, to what was described in the case of RNNs, classification can be performed by introducing a dense layer at the end of the network. In this case, however, most of the network has already been pre-trained, so it already knows how to extract good representations from the input text. As a consequence, the classification layers can be trained very quickly.

Soon after language-model fine-tuning was described, the so-called Transformer architecture *(f)* was introduced. It is based on a component called *Attention* Vaswani et al. (2017), which can broadly be considered an improved version of recurrent layers with memory. The attention mechanism is used to measure the contribution of each term to the meaning of the sentence, making it possible to disambiguate senses in settings that were previously very difficult to solve. In addition to the effectiveness of attention, another major contribution of the Transformer architecture is its efficient scalability: transformer-based networks are easy and stable to train, and can be configured with multiple transformer units to achieve better performance. These factors made it possible for transformer-based models to achieve spectacular results in complex tasks such as machine translation.

Transformer-based encoder models such as BERT Devlin et al. (2019) are capable of producing representations of input texts suitable for classification, either by fine-tuning or by applying traditional classification methods to the representations. Transformers can also be used in auto-regressive, decoder-only configurations, trained to generate text based on an input context called a *prompt (g)*. The text is generated from the input context according to the statistical properties of language that were learned through exposure to vast corpora during the training process. These models essentially predict the most probable

continuation to the input sequence, and this can be leveraged to address a diverse set of tasks, including classification. Auto-regressive models have been one of the main subjects for large scale experiments, where the main components of the architecture are based on transformer blocks, but model size –and, correspondingly, time and cost of pre-training–, are being continually increased. Fig. 9 shows the evolution of the number of parameters in large language models in the past few years.

Models as large as GPT-3 Brown et al. (2020) can only be trained by a handful of large companies and research labs, using vast amounts of data and after significant expense and engineering effort on distributed computing resources. Relatively lower-cost alternatives such as fine-tuning are also impractical. As a matter of fact, even simple inference using pre-trained models is very challenging, as it requires powerful hardware dedicated to the task. The use of these huge models is predicated on two principles:

- The generalization capabilities of models so large, and trained on extremely extensive corpora, are presumed to be much higher than those of smaller models. The promise of these models is being able to use them in zeroshot or few-shot settings, without having to go through any type of fine-tuning. Zero-shot refers to the model being able to provide satisfactory answers for a new task it was not trained on, without ever seeing examples of the task. Few-shot learning is similar, but the user can supply a few examples of inputs and expected outputs of the task. When the model examines this context, it is able to provide reasonable responses to new inputs.
- The models are used as a service provided by a company, and available through an API on top of which new applications can be built. Companies offering these services host the pre-trained models in their premises, and they typically offer access for a price per thousand queries. Users (including researchers, practitioners and companies) have to weigh the benefits of using third-party infrastructure against cost and other factors such as privacy considerations, but they will rarely be able to replicate this infrastructure themselves.

In-between the encoder-only models (such as BERT Devlin et al. (2019) or RoBERTa Liu et al. (2019)) and the auto-regressive models such as GPT-3 Brown et al. (2020), there is a family of encoder-decoder models *(h)* that combine the characteristics of both. The input texts are encoded into intermediate representations, and the representations are decoded as text outputs that serve as responses to the task the user is trying to solve. All potential tasks, including classification and regression, are modelled as text-to-text operations. This approach is taken by models such as T5 Raffel et al. (2020) (Google) and T0++ Sanh et al. (2022) (BigScience). T0++ essentially follows the same architecture as T5, but training is performed on more data and, crucially, more diverse tasks and prompts, with the goal to make the model robust to unseen tasks the user might want to perform.

A recent alternative to text-output APIs comes in the form of access to the internal dense embeddings of large language models *(i)*, including GPT-3. Those embeddings or representations can be retrieved, and new systems can be built on top of them for tasks such as classification or search.

All of these options are available to any sentiment analysis or text classification project, but it's not straightforward to determine what the best solution is for a particular problem. Are APIs to large language models suitable for twitter sentiment analysis, is a hosted deep learning model better, or is it enough to build a custom solution using SVMs and word embeddings? Next subsection briefly looks at the current state of model evaluation.

6.2. Evaluation of methods for sentiment analysis

In order for practitioners and researchers to make informed decisions about the best suited methods for the problem at hand, it would be necessary to have comprehensive tests and benchmarks available for examination. These benchmarks should be performed on comparable datasets, and would not only provide information about classification metrics, but also about the limitations and trade-offs associated to each method. These include an estimation of compute resources to fine-tune a model on a particular domain or dataset (when fine-tuning is necessary), but also the amount of computing required for training from scratch, in those cases where the model architecture is described but the pre-trained weights are not made available for other researchers to use. Performance metrics should not only cover the accuracy of the model, but also the time, cost and complexity required for performing predictions at scale. In addition, transparency about training datasets is important to identify biases the model may have incorporated during training – as it has been demonstrated, social and cultural biases are not just merely reproduced by models exposed to them, but they become

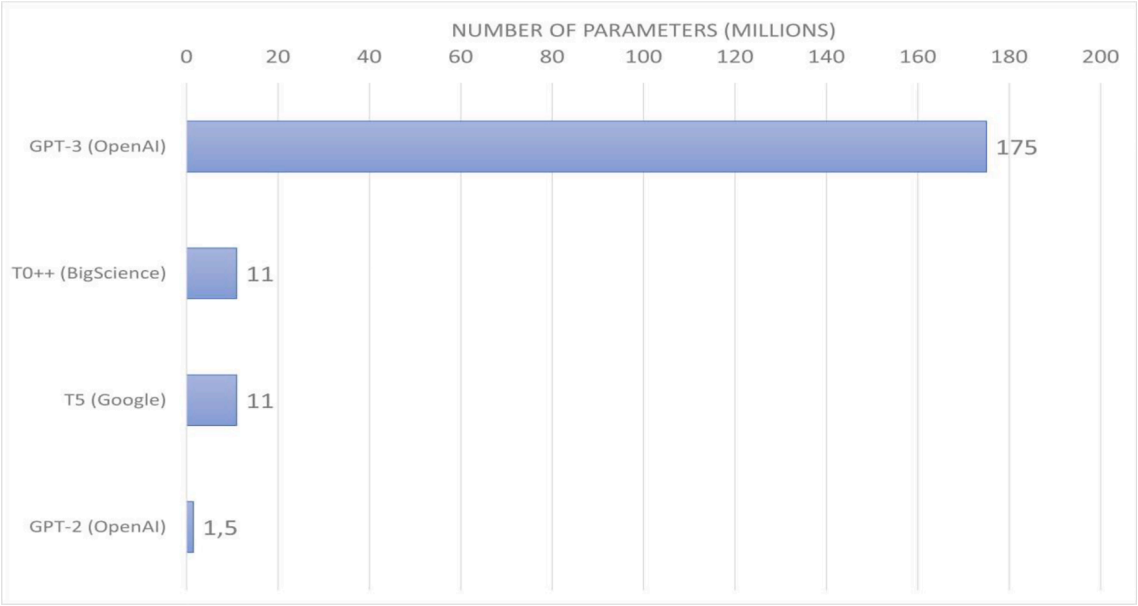

Fig. 9. Trends towards increasingly larger language models.

exacerbated Bender et al. (2021). Finally, it would be ideal to have access to sample code that reproduces the reported metrics and demonstrates how to use the model effectively, including pre-processing and normalization operations that need to be applied to input data.

In the case of sentiment analysis of twitter corpora, we found SemEval 2017 Rosenthal et al. (2017) to be a dataset frequently used as a reference and benchmark. It consists of about 62 thousand twitter messages collected on or before 2017, and it was intended to be used as a common ground for the SemEval competitions[1]. Tweets were classified into three classes (positive, negative, and neutral) using a crowdsourcing service. The age of the dataset might be a problem for newer applications; consider, for instance, that models using this dataset will have no exposure to data from recent events such as COVID-19 or the social and cultural changes of lockdowns.

Other datasets frequently used as benchmarks for sentiment analysis include IMDB and Rotten Tomatoes movie reviews, app reviews, and yelp reviews. Like SemEval 2017, these are mostly static datasets that, in addition, are constrained to very specific domains. Even though they can sometimes be used for fine-tuning, their use of language and the domains they cover differ from the messages that can usually be found in social media and Twitter. This is probably one of the reasons why SemEval 2017 remains the usual reference for Twitter Sentiment Analysis.

SemEval 2017 was the subject of the recent TweetEval Barbieri et al. (2020) benchmark, which explored the performance of various models and baselines on this dataset, and on a variety of tasks. The main results from this research have been reproduced in Table 2, including data for the BERTweet Nguyen et al. (2020) model that is not present in the published paper, but was made available later in the project's website.

TweetEvals focus is on performance metrics, but it does not cover some of the other considerations we mentioned, such as complexity and cost estimates. Fortunately, the authors released the pre-trained models, as well as code to test any other arbitrary models on the same dataset. However, the code does not include all training details for all the methods, so users that want to apply a particular method are compelled to reproduce the published results before they can apply the method on a different domain or dataset. We did that work ourselves for some of the methods in TweetEval, but we focused on the sentiment classification task. We include reference data for inference time and power consumption as a proxy for computing complexity, using the same hardware as a reference. We also include references to the pre-trained models we used, so they can be easily found by other researchers. Our results appear in Table 3.

In addition to BERTweet, there are other baselines that could be of interest to researchers and to the best of our knowledge have not yet been examined elsewhere. These include methods such as ULMFiT Howard & Ruder (2018), and an estimation of how large language models compare in a zero-shot/few-shot setting against specialized, fine-tuned models. To this extent, we performed initial tests on GPT-3 (running as a service) and GPT-J Wang & Komatsuzaki (2021) (running locally), a smaller auto-regressive model described in the next paragraph. Table 3 includes our results on these models.

GPT-J is an auto-regressive model in the spirit of GPT-2 Radford et al. (2019) and GPT-3. Compared to GPT-3, it has a size of 8 billion parameters instead of GPT-3's 175 billion, so it is feasible to host the model for inference in consumer-level hardware. Training is much harder in terms of resources, so we performed some tests based on its pre-trained behaviour. Using a naïve few-shot setting with a fixed prompt, we were able to achieve a recall of 57.0 on TweetEval's test set. We did not explore the sensitivity of this result to inference parameters (temperature, top p), the number of context samples in the prompt, or the polarity of those samples.

We did a similar test with GPT-3 using OpenAI's API access. The method used for classification in this model is much more sophisticated than the naïve approach we used for GPT-J. First, users need to upload a file with sample predictions. We used the complete train set from TweetEval. Then, for each subsequent query, the most relevant samples for that particular query are extracted from the reference set, and used as context for the prediction. Using this method, we achieved a recall of 58.4 over a random 6% sampling of TweetEval's test set. This performance is not a lot better than the naïve classification we performed on the much smaller GPT-J, and neither model appears to be close to the performance of specialized, fine-tuned models like BERTweet or RoBERTa-Retrained, or to small and convenient models such a SVM classifier on word embeddings obtained using fastText.

To allow other researchers to examine the details that led to these results, we will publish the source code for these experiments.

These results provide an idea of the order of magnitude that is achievable with the most representative sentiment classification methods. However, more work is required in this area to cover:

- Wider use, such as the fine-tuning of models like GPT-J or GPT-3, and the use of dense embeddings from large models.
- Further analysis on cost and complexity.
- Identification of social and cultural biases.

In addition, recent preliminary reports [2]suggest that the performance of large models, particularly in few-shot or zero-shot settings, depends on the text prompts used to query the system. We think that these topics

Table 2
TweetEval results for the sentiment classification task on the SemEval 2017 dataset. Sources: TweetEval paper Barbieri et al. (2020), https://github.com/cardiffnlp/tweeteval. The sentiment score is measured as the macro-averaged recall across the three classes.

Model	Sentiment Score	Notes
TimeLMs-2021	73.7	
BERTweet	73.4	
RoBERTa-Retrained	72.8	Fine-tuned on Twitter data.
RoBERTa-Base	71.3	Pre-trained model.
RoBERTa-Twitter	69.1]	Trained from scratch on Twitter data.
FastText	62.9	100 dimensions. Includes subword units.
LSTM	58.3	100-dimensional FastText embeddings.
SVM	62.9	Word and character n-gram features.

Table 3
Our results on the test set of SemEval 2017 (sentiment classification task). We could reproduce previous scores on BERTweet and RoBERTa-Retrained, and tried GPT-J and GPT-3 in few-shot settings. Additional experiments on GPT-J, GPT-3 and other large language models are necessary to assess their true potential. (1) Unbatched, sequential iteration time. Power consumption estimated at 1/8th of a TPU consuming about 200 W during inference. (2) Inference time extrapolated from 6.5% of test set.

Model	Sentiment Score	Inference Time	Power	Notes
BERTweet	73.4	5 s	1.75 kJ	
RoBERTa-Retrained	72.8	7 s	2.42 kJ	
GPT-J	57.0	2190 s [1]	55 kJ [1]	*temperature* = 0.3, top_p = 1.0
GPT-3	58.4	12912 s [2]	Unknown	https API

[1] https://alt.qcri.org/semeval2017/.

[2] https://wandb.ai/ivangoncharov/GPT-3/reports/Summary-Sentiment-Question-Answering-More-5-Creative-Tips-for-GPT-3-PromptEngineering–VmlldzoxODY0Nzky.

provide excellent opportunities for further research that we intend to address in future work.

7. Discussion and conclusions

As mentioned earlier in this paper, the interest in sentiment analysis in social networks has grown exponentially in recent years. This reality continues to be true even after hundreds of articles and extensive review papers Ravi & Ravi (2015); Ashima & Kumar (2020) have been published. A justification for this continuous increase could be found both in the progressive evolution and incorporation of new processing techniques and domains, and in the inherent difficulty to qualify the information in the different domains, as it was found in the unequal results obtained with various lexicons in a variety of research papers Rodriguez-Ibanez et al. (xxxx); Rodríguez-Ibánez et al. (2021).

According to the literature, the temporal perspective of sentiment analysis is seen as a valuable tool for a better understanding of the dynamics of the underlying emotional state, especially in a circadian approach in certain disciplines such as in the case of COVID, advertising, internal communication in schools, and even in the prevention of bullying. A high prevalence of temporal sentiment analysis studies were found to be based on rules and lexicons. Temporal analysis of sentiment has also shown an important relationship with other non-temporal variables, such as location. This approach has made it possible to find a double dimensional domain (space–time) where regions may characterize a given emotional state.

Even further, temporal analysis reached a greater dimension with the incorporation of the concept of Causality. Granger causality has shown not only to be of great interest to the scientific community due to the significant number of publications, but also due to the potential to link different variables, as is the case of the company value prediction in relation to social media mood. In this same example, and as described earlier, an interesting conclusion worth discussing is how price and sentiment present a two-way relationship S. Bouktif et al. (2020). Instant messages may anticipate price movements, and these in turn may precede sentiment outbreaks in the news. In this second case, however, the relationship is produced in shorter time frames. These conclusions suggested the existence of a causal relationship between both variables. On the other hand, it was proven impossible to establish rigorously which of these two directions of relationship is more intense. The works published used a good number of learning methods and techniques, such as linear and non-linear methods (support vector machine, latent Dirichlet, N-grams, etc.). The conclusion regarding the studies presented does not allow us to validate in a general way a technique as the most appropriate, not even for the particular case of the valuation of the stock market and the news, demonstrating that some techniques are valid for certain companies and not for others. The main conclusion that can be drawn from the current literature is that sentiment evidence increases the ability to predict the behavior of other variables, but it is not enough to fully describe the more than complex relationships among them. Although this aspect of the discipline is already covered by an important number of publications, as it was described in detail in the corresponding section of this work, we could argue that the multiplicity of approaches with diverse results suggests the existence of a wide field of work in this area, and especially when the approach is limited to specific domains where the specifics of the applications and the semantics require an adequate in-depth treatment of the input information.

One piece of evidence about the interest and the existence of tangible results in the application of artificial intelligence to social networks can be found in the vast proliferation of both patents and commercial applications for this type of analysis. The analysis of research results carried out in this area allows us to conclude that the current state of these commercial tools can act on three main areas: (a) the management of social networks, (b) monitoring and active listening, and (c) the analysis of what is published. The first two cannot be considered strictly belonging to the field of sentiment analysis in social networks, but the third can. As mentioned, the current state of the art does not seem to be consolidated in the commercial field, due to the enormous volume of software available and the fact that the few comparative studies indicate the divergence of the results of the different tools. This fact, together with the more than 8,000 patents published in 5 years, where large companies in the sector (Google, Facebook, Microsoft, etc.) monopolize a very significant share of them, encourage us to think about the relevant room for further improvement in sentiment analysis, whether in the academic or commercial field.

In terms of domains on which the analysis of sentiment in social networks has focused, the detailed analysis of the over two thousand academic publications since 2007 has shown an uneven follow-up. Although social networks cover all areas of society, the analysis of these has not been carried out in the same way. In particular, Marketing stands out as the most analyzed field, Politics is second, then Economics and Health. The strong increase in the number of publications in essentially all domains suggests a growing interest in the use of these techniques and their potential value for the different businesses.

Another factor to consider is the diffusion of the studies of the different domains. Marketing and politics, as stated, are the most researched, and we can argue this could be related to the ease of access to data (via data extraction from Twitter or commercial web pages) and also íts likely that the composition of the teams facilitates these types of studies (mainly from the economic field, social sciences, mathematicians and engineers). We can debate here, why other domains such as health or emergencies, did not show an equivalent proliferation of studies. This might be related to the fact that specialized knowledge is required and the datasets are very much subjected to data protection, subsequently limiting the access for researchers to propose further analysis. It is also worth mentioning that the domain of business and economics has given prominence to other domains in the academic field and scientific publications. This fact could be related, on the one hand, to the relevant entry of new domains in scope that have multiplied the number of publications, and on the other hand, to the great interest shown by commercial entities to offer specific and tailored solutions and applications to companies, financial institutions and public administrations, corresponding efforts academic environments.

We want to point out that the existence of 4 times more patents in only 5 years than the number of academic publications in the last 15 may denote the existence of huge efforts by commercial entities to monopolize and protect their results in this area. This fact forces us to consider the existence of a significant number of results that are part of the current public domain.

Regarding technologies, it should be noted that the traditional ones (lexicons, tokens, Bayesian methods, bag of words) are still widely used whereas the newest methods (auto-regressive and encoder-decoder transformers) are not yet in widespread use. This may be due to the fact that the classic techniques are well established and are more affordable or more approachable for multidisciplinary teams. If we compare techniques across domains, we don't see a significant difference in terms of the mix of techniques applied in each of the domains. Regarding the specific techniques applied, the use of lexicons, Word2-vec, and n-grams stands out. Transformer-based techniques emerge, but to a lesser extent, and newer and more complex models such as T5, T0++, GPT-3 or GPT-J are infrequently cited. We believe that the use of these large pre-trained models will represent a future paradigm in sentiment analysis, as in many other disciplines, especially when zero-shot, one-shot or few-shot learning models are applicable. While this appears to be the case, preliminary outcomes of the reproducibility analyses performed in this work still yield very limited results. In particular, for the simple experiments carried out using GPT-J and GPT-3 over the TweetEval dataset, we scored only 57.0 and 58.4 (macro-averaged recall), respectively, which compare poorly with the state of the art from traditional methods, and suggest that careful domain adaptation is still needed. We believe that additional efforts are needed from academics and commercial entities to assess the trade-offs involved

in the use of complex systems such as large language models. Unless those trade-offs (including computing resources, power consumption, inference time, cost, ease of development, operational complexity) are fully understood, it will remain hard to measure the suitability and applicability of these techniques to solve real-world problems across a variety of domains.

In summary, based on the present study, some of the limitations that we have found are related to the impossibility of reproducing all the experiments proposed in previous investigations with the new reproducibility techniques, such as Transformers-based systems, like BERT, T5, T0++, or GPT-2/3, and thus being able, later, to compare the results obtained with those that we would obtain with the traditional techniques. In addition to that, although the authors have made a thorough review of what the industry has done on sentiment analysis, existing patents, etc., the operation, capacity and scope of the tools built by the different companies are still not covered adequately by the scientific community. One of the areas of sentiment analysis in which there is more to be done is related to the temporal causality of sentiment. Demonstrating the causality of sentiment in the different areas of study and the development of new algorithms in this field is a huge area of research. The fundamental prospects that are shown and discussed in this study are important for future advancements and developments, as they will provide a great motivation for researchers to explore and devise new approaches that will overcome the critical issues and major new challenges in this field.

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

No data was used for the research described in the article.

Acknowledgements

This work is partly supported by research grants miHeart RisBi (PID2019-104356RB-C42), miHeart-DaBa (PID2019104356RB-C43), BigTheory (PID2019-106623RB-C41), from Agencia Estatal de Investigacion of Science and In- novation Ministry and cofunded by FEDER funding. It is also partially supported by REACT EU grants from the Community of Madrid and Rey Juan Carlos University funded by the Next Generation EU.

References

Abbasi, A., Hassan, A., & Milan, D. (2014). Benchmarking twitter sentiment analysis tools. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (p. 823–829). European Language Resources Association.

Ashima, Y., & Kumar, V. D. (2020). Sentiment analysis using deep learning architectures: A review. *Artificial Intelligence Review, 53*, 4335–4385.

Balaz, J., & Velazquez, J. (2016). Opinion mining and information fusion: A survey. *Inf. Fusion, 27*, 95–110.

Barbieri, F., Camacho-Collados, J., Espinosa Anke, L., & Neves, L. (2020). TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 1644–1650). Online: Association for Computational Linguistics.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623).

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

Blanco, R. (2021). *Comparativa de técnicas de análisis de sentimiento en contextos competitivos*. Master's thesis Universidad Complutense de Madrid, Trabajo Fin de Grado Madrid, Espana.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems* (pp. 1877–1901). Curran Associates, Inc. volume 33.

Chandio, M., & Sah, M. (2020). Brexit twitter sentiment analisys: Changing opinions about brexit and UK politicians. In *International Conference on Information, Communication and Computing Technologies: Intelligent Computing Paradigm and Cutting-edge Technologies* (pp. 1–11). Springer volume 9.

Chatterjee, A., & Das, A. (2020). *Temporal Sentiment Analysis of the Data from Social Media to Early Detection of Cyberbullicide Ideation of a Victim by Using Graph-Based Approach and Data Mining Tools, 1109*. Intelligence Enabled Research.

Chaturvedi, I., Cambria, E., Welsch, R., & Herrera, F. (2018). Distinguishing between facts and opinions for sentiment analysis: survey and challenges. *Inf. Fusion*, (pp. 65–77).

Cho, K., van Merrienboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation* (pp. 103–111). Doha, Qatar: Association for Computational Linguistics.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning, 20*, 273–297.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Computer Science*, (p. 4171–4186).

Ecemis, A., Dokuz, A., & Celik, M. (2021). *Temporal Sentiment Analysis of Socially Important Locations of Social Media Users* volume 4. *Innovations in Smart Cities Applications*.

Elsner, J. B. (2007). Granger causality and atlantic hurricanes. *Tellus A Dyn. Meteorol. Oceanography, 59*, 476–485.

G2.com (2022). Best Social Media Suites Software, *https://www.g2.com/categories/social-media-suites*. Visited: January, 2022.

Gao, Q., Duan, X., & Chen, H. (2011). Evaluation of effective connectivity of motor areas during motor imagery and execution using conditional granger causality. *NeuroImage, 54*, 1280–1288.

Gartner Group Inc. (2021). *A Game Changer In The Way Organizations Deal With Data*. Stamford, USA: Technical Report Gartner Group Inc.

Granger, C. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica, 37*, 424–438.

Gujarati, D. N., & Porter, D. C. (2009). *Causality in economics: The Granger causality test in Basic Econometrics*. McGraw-Hill.

Gupta, S., & Sandhane, R. (2022). Use of sentiment analysis in social media campaign design and analysis. *Cardiometry, 22*, 351–363.

Hemmatian, F., & Sohrabi, M. (2019). A survey on classification techniques for opinion mining and sentiment analysis. *Artificial Intelligence Review, 52*, 1495–1545.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*, 1735–1780.

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 328–339). Melbourne, Australia: Association for Computational Linguistics. doi:10.18653/v1/P18-1031.

Hu, T., She, B., Duan, L., Yue, H., & Clunis, J. (2020). A systematic spatial and temporal sentiment analysis on geo-tweets. *Ieee Access, 8*, 8658–8667.

J. Park, H. Leung, & K. Ma (2017). Information fusion of stock prices and sentiment in social media using granger causality. In *IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems* (pp. 614–619).

J. Park, K. Ma, & H. Leung (2019). Prediction of stock prices with sentiment fusion and SVM granger causality. In *IEEE International Conferance on Dependable, Autonomic and Secure Computing* (pp. 207–214).

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, .

Kaur, S., & Mohana, R. (2019). Temporality based sentiment analysis using linguistic rules and meta-data. *Proceedings of the National Academy of Sciences India Section A - Physical Sciences, 89*, 331–339.

Kraaijeveld, O., & Smedt, J. (2020). The predictive power of public twitter sentiment for forecasting cryptocurrency prices. *J. Int. Financ. Mark. Insitutions Money, 65.*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692, abs/1907.11692*.

M. Chvostekova (2019). Granger causality inference and time reversal. Iń *12th International Conference on Measurement* (pp. 110–113).

Marinazzo, D., Pellicoro, M., & Stramaglia, S. (2008). Kernel-granger causality and the analysis of dynamical networks. *Physical Review E., 77*, Article 056215.

McKinsey & Co. (2018a). *Analytics comes of age*. USA: Technical Report McKinsey & Co New York.

McKinsey & Co. (2018b). *Winning in digital ecosystems*. USA: Technical Report McKinsey & Co New York.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Nazir, A., Rao, Y., Wu, L., & Sun, L. (2022). Issues and challenges of aspect-based sentiment analysis: A comprehensive survey. *IEEE Transactions on Affective Computing*, (pp. 845–863).

Nguyen, D. Q., Vu, T., & Tuan Nguyen, A. (2020). BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 9–14). Online: Association for Computational Linguistics.

Park, S., Strover, S., Choi, J., & Schnell, M. (2021). Mind games: A temporal sentiment analysis of the political messages of the internet research agency on facebook and twitter. *New Media and Society*, (pp. 1–22).

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). Doha, Qatar: Association for Computational Linguistics.

Preethi, P., Uma, V., & kumar, A.. (2015). Temporal sentiment analysis and causal rules extraction from tweets for event prediction. *Procedia Computer Science, 48*, 84–89.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog, 1*, 9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research, 21*, 1–67.

Rajalakshmi, S., S.Asha, & Pazhaniraja, N. (2017). A comprehensive survey on sentiment analysis. In *Fourth Int. Conf. Signal Process. Commun. Netw.* (pp. 1–5). IEEE.

Ravi, K., & Ravi, V. (2015). A survey on opininon mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems, 89*, 14–46.

Rodriguez-Ibanez, M., Gimeno-Blanes, F., Cuenca-Jimenez, P., Munoz-Romero, S., Soguero-Ruiz, C., & RojoAlvarez, J. . On the statistical and temporal dynamics of sentiment analysis. *IEEE ACCESS*, *8*, 87994–88013.

Rodríguez-Ibánez, M., Gimeno-Blanes, F.-J., Cuenca-Jiménez, P. M., Soguero-Ruiz, C., & Rojo-Alvarez, J. L. (2021). Sentiment analysis of political tweets from the, spanish elections. *IEEE Access, 9*, 101847–101862.

Rosenthal, S., Farra, N., & Nakov, P. (2017). SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)* (pp. 502–518). Vancouver, Canada: Association for Computational Linguistics.

Bouktif, S., Fiaz, A., & Awad, M. (2020). Augmented textual features-based stock market prediction. *IEEE Access, 8*, 40269–40282.

S. Garg, D. Panwar, A. Gupta, & R. Katarya (2020). A literature review on sentiment analysis techniques involving social media platforms. In *Sixth International Conference on Parallel, Distributed and Grid Computing* (pp. 254–259).

S. Zhao, Y. Tong, X. Liu, & S. Tan(2016). Correlating twitter with the stock market through non-gaussian SVAR. In *Eighth International Conference on Advanced Computational Intelligence* (pp. 257–264).

Sammut, C., & Webb, G. I. (2010). TF–IDF. In *Encyclopedia of Machine Learning* (pp. 986–987). Boston, MA: Springer, US.

Sanh, V., Webson, A., Raffel, C., Bach, S. H., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Scao, T. L., Raja, A. et al. (2022). Multitask Prompted Training Enables Zero-Shot Task Generalization. In *ICLR 2022 - Tenth International Conference on Learning Representations*. Online, Unknown Region.

Sharuee, M., Liu, F., & Pratama, M. (2021). Sentiment analysis: Dynamic and temporal clustering of product reviews. *Applied Intelligence, 51*, 51–70.

Singh, N. A. (2020). Sentiment analysis on motor vehicles amendment act, 2019 an initiative by government of india to follow traffic rule. In *In 2020 International Conference on Computer Communication and Informatics (ICCCI)* (pp. 1–5). IEEE.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems*. Curran Associates, Inc. volume 30.

Wang, B., & Komatsuzaki, A. (2021). GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax.

Yu, J., Aduragba, O. T., Sun, Z., Black, S., Stewart, C., Shi, L., & Cristea, A. (2020). Temporal sentiment analysis of learners: Public versus private social media communication channels in a women-in-tech conversion course. In *15th International Conference on Computer Science and Education* (pp. 182–187).

Yu, S., Eisenman, D., & Han, Z. (2021). Temporal dynamics of public emotions during the COVID-19 pandemic at the epicenter of the outbreak: Sentiment analysis of weibo posts from wuhan. *Journal of Medical Internet Research, 23*.

Zhang, Y.-X., Liu, X.-H., Wang, W.-J., & Liu, Y.-J. (2020). A study of relationship between investor sentiment and stock price: Realization of investor sentiment classification based on bayesian model. In *In 2020 International Symposium on Computer Engineering and Intelligent Communications (ISCEIC)* (pp. 34–37). IEEE.

Zheng, C., & He, T. (2010). Investor sentiment and stock index: A test of causality based on vector error correction model. In *The 2nd International Conference on Information Science and Engineering* (pp. 1–4).

Zhu, J., Chen, Y., Leonardson, A. S., Wang, K., Lamb, J. R., Emilsson, V., et al. (2010). Characterizing dynamic changes in the human blood transcriptional network. *Computational Biology, 6*, 10–67.