Social Network Analysis and Mining (2023) 13:46
https://doi.org/10.1007/s13278-023-01048-1

ORIGINAL ARTICLE

Social media discourse and voting decisions influence: sentiment analysis in tweets during an electoral period

Paulo Rita[1] · Nuno António[1] · Ana Patrícia Afonso[1]

Received: 25 May 2022 / Revised: 16 February 2023 / Accepted: 17 February 2023 / Published online: 7 March 2023

Abstract
In a time where social media is fundamental for any political campaign and to share a message with an electoral audience, this study searches for a conclusion of the actual persuasion capacity of social media in the electors when they need to decide whom to vote for as their next government. For this, it compares the sentiment that Social Media users demonstrated during an electoral period with the actual results of those elections. For this analysis, it was used, as a case study, tweets mentioning the two major English parties, Conservative and Labor, their respective candidates for the position of prime minister, and terms that identified their political campaign during the electoral period of the General Elections of the United Kingdom that occurred on December 12, 2019. Data were collected using R. The treatment and analysis were done with R and RapidMiner. Results show that tweets' sentiment is not a reliable election results predictor. Additionally, results also show that it is impossible to state that social media impacts voting decisions. At least not from the polarity of the sentiment of opinions on social media.

Keywords Elections · Politics · Sentiment analysis · Social media

1 Introduction

During the last 20 years, social media has been taking a very important and transforming role in the daily life of citizens worldwide. Recent statistics show that one-third of the worldwide population uses at least one socialization platform for the most diverse ends, especially among the younger generations that spend, on average, more than 4 h connected to the internet (Ortiz-Osina 2019).

The growing adoption of these technologies provoked a need for adaptation to this new virtual reality. What before only existed in an offline mode needed to be adapted to this digital generation, like finding new ways of presenting information to a new digital reader, exploring new online business opportunities, and developing strategies to get closer to a more web-connected audience.

✉ Paulo Rita
prita@novaims.unl.pt

1 NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa, Campus de Campolide, Lisboa, Portugal

Politics was one of those areas that had to adapt to this new reality. In an era where Social Media platforms are used to retain political information to help voters' decision-making process, the political parties had to enlarge their electoral strategy to these platforms (Broersma and Graham 2012). This decision came to gain new voters by persuading online information seekers with their ideologies and not lose potential voters with the fake news overload that appears during an electoral period (Anstead and O'Loughlin 2015).

However, not only politicians use social media to expose their ideologies (Weeks et al. 2015). Before an election, political parties and their representatives become a massive trend in these socialization platforms, mainly because of the user's exposition of their feelings about current news related to the parties' campaigns, with all the million users present in the network, creating a discussion around that topic (Chan and Fu 2017).

This research aims to understand if the sentiment in those discussions can predict votes and conclude if the online political persuasion created within those discussions impacts the vote decision.

Previous studies helped prove the existence of opinion polarization in social media and its effects. However, most of the studies focused on the actual prediction capacity using the content of Social Networks presented the following literature gaps: Researchers made analytical studies without using data close to the election day; Studies created a prediction model based on the count of tweets for each party, instead of their content (Tumasjan et al. 2010); Conclusions made mainly from expert views; and No consensus if Twitter should be used as an electoral predictor (Chung and Mustafaraj 2011).

This study contributes to research on social media by showing by investigating if Twitter content is a reliable election predictor by using data collected closer to the election, based on the sentiment presented on the tweets mentioning a party, if it could have affected positively or negatively the actual votes for that party. In other words, during a search for political information on social media before Election Day, the sentiment of the words that the user was in contact with could have influenced the decision to vote for a specific party.

This research used a case study of the General Elections in the United Kingdom (UK) electoral period from December 2019.

2 Literature review

2.1 Social media

At the beginning of the 2000s, Web 2.0 and online applications appeared due to digitalization and the Internet evolution. These online applications, now called social media, are platforms that promote interaction, the sharing of knowledge, and collaboration between users, without the existence of geographical barriers. It can be seen as a pedagogical tool that stimulates the discovery and sharing of information, supporting the connection between users who share common goals and interests (Usluel and Mazman 2009). In contrast with websites and mobile applications (Ramos et al. 2019), applied to brand management (Moro et al. 2016; Pinto et al. 2019; Romão et al. 2019) and used by a myriad of sectors such as hospitality and tourism (Moro and Rita 2018; Nave et al. 2018).

Due to their characteristics, social media had fast growth during the last two decades. The rise started with just one Social Network—MySpace, founded in 2003 to allow users to interact on the website and customize their profiles without any associated costs (Allgaier 2018). The audience well accepted these features. In less than a year, achieved the first million users and temporally overtook Google as the most visited website in the USA. Around one-third of the worldwide population is connected to a social network. MySpace has more than 100 million user accounts created but was surpassed by several other Social Networks that appeared years later, like Facebook, which alone, in 2019, presented 2.4 billion active users (Ortiz-Osina 2019).

The allowance of socialization without geographical barriers made Social Media platforms an everyday tool for a more online audience. It permitted users from around the world to interact with others who share the same ideologies/interests without meeting personally. The rise of these platforms brought a new phenomenon—the online polarization effect on the users (Maes and Bischofberger 2015).

The selective exposition of topics, due to personalization algorithms that tend to show the user's subjects that supposedly are preferred by them (Maes and Bischofberger 2015) and the natural proximity from users that share common interests, led to the creation of discussions around different topics in these social platforms (Weeks et al. 2015). Scholars argued that those discussions led to opinion polarization among the audience, which can be defined as "the extent to which opinions on an issue are opposed about some theoretical maximum" and polarization as a process that "refers to the increase in such opposition over time" (DiMaggio et al. 1996). Especially during heightened political conflicts, the polarization effects of the digital are more likely to arise. With and without a clear political orientation, users see themselves as applicable to this influence by contacting discussions and news exposure in digital media (Lee 2016).

The rapid adoption of social networking platforms changed the socialization between users and provoked a need to adapt from the digital to what was possible offline to be closer to a new online audience. Digital media appeared because of digitalization.

In an era where information from around the world is at the distance of a simple click, media, to survive this new reality and stay relevant to the younger generations, found its path by creating Social Media accounts were able to share in real-time news for users to read, discuss, and share content with their network and by using Social Media content to create news of trending topics and relevant discussions occurred in these platforms (Enli and Simonsen 2017).

2.2 Politics

The advances of media in the digital field and real-time content creation, without physical barriers, made Politics a more mediatized and globally non-stop trend topic in online platforms.

Due to the controversy around this subject and its importance for the audience to keep easily updated, media uses Social Media content created by political parties and their respective candidates to obtain and share political information (Broersma and Graham 2012). Social media is also used to reflect a mainstream political opinion by looking at discussions created around trend political topics as a source to produce breaking news (Anstead and O'Loughlin 2015).

Being a constant target of mass media, Politics was another area that had to adapt to this digitalization. To stay relevant in the eyes of an online audience, campaign strategies and the propagation of political messages had to adapt to this new reality while building strong image management in a time where every flaw shown in media coverage can be used as leverage by the opposition.

The beginning of using social media as a political advantage can be backed up by the U.S. Presidential Election Campaign in 2008. Barack Obama's campaign was seen as a model guide for the future in how efficient the use of social media was for political marketing during campaign season. Knowing the large percentage of the young American generation present on Social Networks, to capture their attention for this election period, the Democrats created a multichannel Social Media campaign with a high focus on the constant creation of content with the democratic political message and media coverage (Enli and Naper 2016). As a result, Barack Obama's innovative initiative made him a celebrity politician with almost 21 million more followers than his opposition by the time of the U.S. Presidential Election Day. Ultimately, he won the elections (Bimber 2014). The success of this campaign led to an increase in the adoption of these platforms and high engagement by political parties around the world. This rise was especially noticed among "underdog" politicians, and a term used to describe politicians still new and not well established in the politician environment, to express their message since the traditional media is more focused on popular politicians (Larsson and Kalsnes 2014).

Adopting social media as a form of political campaigning brought several advantages to politicians. It is a cheap, easy-to-work, and engaging tool to interact with voters without the need for intervention by traditional media coverage. It gives politicians the power to decide what and how they want to share content with the audience, to be later used as news coverage by the media (Broersma and Graham 2012).

Nevertheless, most Social Media users do not use these platforms primarily for political reasons, as the primary goal of these platforms is, as mentioned before, the socialization between users (Diehel et al. 2015). However, in some situations, online social interaction with individuals who share common ideologies and contact with political content created by the parties and digital media can make a user reconsider his political opinions (Weeks et al. 2015).

Recent studies mention social media as the ideal place for political persuasion, as its characteristics incentivize users to politically express and participate in discussions with others (Stieglitz and Dang-Xuan 2012). Users who engage in these activities are called "prosumers"—Social Network engagers who access news in digital format, interact with others, and create/share content, potentially caring influence on others. These "prosumers", by highly engaging in these activities, have a higher probability of trying to change other users' minds about a specific candidate and political causes and persuade them to vote for the candidates they believe to be the "right party." They believe that, by doing so, they can persuade other individuals' opinions. The Social Media environment also incentivizes this behavior by associating long discussions between users around a specific topic as a trend, making it easy to be found by other network users (Park 2019).

Especially, during heightened political conflicts, as discussed before, the opinion polarization effect tends to arise from these networks. With younger generations being more inclined to use social media as a source of political information (Weeks et al. 2015) instead of traditional news, they end up being in contact with emotional discussions between self-perceived opinions leaders actively attempting to change other users' perception. Being the main source of truth for these individuals, prosumers achieve the goal of influencing users with and without a clear political orientation with their ideology (Chan and Fu 2017). However, the real impact of online persuasion is yet to be proved, as previous studies on this topic were focused on applying surveys to an audience of experts in the field (Diehel et al. 2015).

Social Media, besides being used as a quick and cheap newsgathering tool by the media (Broersma and Graham 2012) and a self-advertising tool by political parties, can also be used as a satisfaction predictor for the parties. By looking at the discussions involving political parties and their candidates, it is possible to understand how positively and negatively the users react to specific topics and how the strategy of politicians needs to change to gain more support from the audience.

The Academic world has been studying this phenomenon and discussing if the use of text generated from Social Network users' interactions transmits the vote intention of the electorate and can be used to effectively predict the outcome of an election (Makazhanov and Rafiei 2013). This discussion started in 2010 with the paper "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment." The authors were able to prove that tweets reflect the electorate's political sentiment by extracting the sentiment to create candidate profiles with the results and that Twitter serves as a predictor of election results by comparing the share of attention the political parties received with the election result (Tumasjan et al. 2010). After 1 year, this study was contested for these two reasons: The researcher used data collected ten days before Election Day. It was proved that tweets collected one week before an election negatively correlate with the result. Counting tweets that mention a party is not enough to predict an election outcome as it does not measure the sentiment of electors toward a political party (Jugherr et al. 2012).

More than a decade later, this topic is still controversial as no exact conclusion was accepted by the academic community (Liu et al. 2021; Skoric et al. 2020).

From the analysis of several similar studies, researchers who defend this prediction (Makazhanov and Rafiei 2013; Sang and Bos 2012) reveal a clear preference for Twitter as a data source for this analysis due to its characteristics—it allows users to post messages with a maximum of 140 characters, which obligates to resume the feelings in two or three phrases, focusing more on what they want to express. The mentioned research was collected from Twitter and used as a data source, tweets by users referencing the political parties and their top leaders running for an electoral seat. Regarding methodology, the choice was to apply a Sentiment Analysis as it often shows better predictor results than traditional pools (León-Borges et al. 2015).

Sentiment Analysis is a technique used in Text Mining to computationally extract sentiment, opinions, and subjectivity from text data. According to the algorithm applied to the model, this type of analysis will analyze and retrieve the percentage of positive/negative/neutral sentiment present in a text (Medhat et al. 2014). This analysis is often applied to social media content to measure opinion and satisfaction with a particular topic, person, or product (Cambria 2016). In this research field, however, there is no agreement on the Sentiment Analysis's efficiency for this prediction type and which algorithm should be applied to achieve the most accurate prediction model (Hemalatha et al. 2013).

However, it is consensual between the studies which followed this approach that using Social Media content as a measure of public opinion has several limitations. Starting from the point of data collection, the Twitter API collects only 1% of the overall tweets made in a specific period. This choice of data collection may affect the balance of the dataset and induce bias (Jaidaka et al. 2018). Another concern is regarding an equal demographic representation between the tweets. Social Media platforms are mainly used by younger generations (Kemp 2020) and higher education citizens (Wei and Hindman 2011), leaving a large part of the electorate out of this sample. Besides this, the discussions in social media are often dominated by "prosumers," who represent a small portion of this web platform users and carry highly convicted opinions that may not represent most users (Stieglitz and Dang-Xuan 2012). Without being possible to retrieve personal information from Twitter to confirm an equal demographic distribution of the tweets used for analysis purposes, previous studies to face this problem simply assumed the tweet distribution to be demographically representative of the overall population (Sang and Bos 2012). This decision, however, can lead to biased estimations of the probability of a win for a certain party.

Due to these factors, a part of the Academic field defends that Social Media analysis will never replace election forecasts, as both should complement each other (Skoric et al. 2020). The election forecast, executed as a survey to be applied to individuals who can vote, transmits the vote intention of a group considered representative of the overall population. Nevertheless, it does not show what shaped the opinion of that group to make them choose a particular party (Lazarsfeld 1957). The use of social media offers the possibility to explore how public opinion was formed and changed during a period (Skoric et al. 2020). As the formation of public opinion was proved to be the result of the participation of citizens in discussions and debates (Blumer 1948), an analysis of political behavior in social media can provide deeper insights into how the influence of this socialization shapes the opinion of an audience. Even so, the impact of these online discussions in the offline world is yet to be demonstrated. Several other factors may influence an elector's decision, such as personal emotions and recent events in life (Healy et al. 2009).

3 Conceptual model and research hypotheses

3.1 Objectives of the investigation

Looking at the discoveries presented in the previous chapter, studies surrounding the use of social media in the Politics field left several open gaps in the literature.

Aside from the lack of agreement on how Twitter sentiment can be used to predict the outcome of an election, there is also doubt if it would be a better approach than the well-established forecasts. The veracity of the online discussion in social media having a tangible impact on the offline is yet to be proved.

This study appeared to contribute to investigating whether Social Media content should be used as an effective election predictor. By applying Machine Learning algorithms to retrieve the sentiment from the users' messages, along with the exploration to which one presents a better accuracy, this research attempted to fill this literature gap: Prediction based on data collected closer to an election day.

Additionally, it aimed to conclude whether political persuasion in social media impacts the vote decision. In other words, the sentiment of the words a user was in contact with days before Election Day could affect the vote decision for a specific party.

This study was confronted with two research questions:

- Is the result of an election possible to be predicted based on the sentiment of tweets mentioned during an electoral period?
- Can the positive discourse of Social Media influence electors' vote decisions?

3.2 Research hypotheses

To answer the objectives above-mentioned, there were formulated two research hypotheses (Table 1).

The first hypothesis assumes the sentiment of the tweets created days before Election Day reflects the actual results of an election by showing the most positive sentiment on the party that won the elections, reflecting the vote intention from the electorate.

The second hypothesis assumes that even if it is not the determinant factor to make a user choose a candidate instead of another, the strength of the sentiment in the tweets a user was in contact with days before Election Day contributes to the vote decision of an elector that uses social media to consume political information.

This sentiment strength should be proven by unveiling a positive correlation between the sentiment of tweets made days before the elections and its results. The null hypothesis will be rejected if the vote decision of the electorate was per the strength of the sentiment presented in the tweets made days before by showing the winning party of the election as the party with the highest positive sentiment in tweets and the loser party with the highest amount of negative sentiment.

3.3 Case study

To test the Research Hypotheses, it was selected as a case study of the General Elections of the United Kingdom, which occurred on December 12, 2019. This electoral period was chosen not only for being a recent event, which allowed data collection until the day before Election Day, but also for its importance and controversy.

After two consecutive general elections since 2015, around 46 million British voters met for a third time to elect the government that would most likely run the country for the next five years. This election came from a previous one that occurred in 2017. In this previous election, the two major English parties, Conservative and Labor, did not have enough votes to form a majority government and ended up losing both members in the Parliament due to colligations done with other parties to win votes in the House of the Commons.

Additionally, the United Kingdom was facing several issues that made urgent the occurrence of a new election: The National Health Service crisis and Brexit—the UK's departure from the European Union (Westbrook 2019).

The numbers divulged by Twitter (Minshall 2019) revealed a growth in engagement in this Social Network compared with the last electoral cycle. From December 6 to December 12, more than 15 million tweets mentioning this election cycle were posted by cybernauts. The volume of tweets increased by 66%, and the Conservative and Labor Party mentions were very similar until the day before Election Day when the difference was just 0.32% in tweet volume. These numbers revealed an increase in the importance of this service as a place for the electorate to receive information, as approximately 170,000 United Kingdom residents used Twitter to find where their polling station was located.

Before the analysis, it was also essential to understand the current use of digital platforms in the United Kingdom. According to the Digital 2020 United Kingdom, a report that summarizes the data, statistics, and trends of digital use, the UK had around 45 million active Social Media users in January 2020, representing a Social Media penetration of 66%. Between April 2019 and January 2020, the number of Social Media users increased by 1.3 million (+2.9%). Furthermore, the average daily internet use achieved 5h48, with 1h42 of that time spent on Social Network platforms. Twitter was the ninth most searched website of 2019, with 235,700,000 monthly traffic (Kemp 2020).

4 Methodology

The application of the methodology started with gathering data from the proposed case study. This data collection was possible after the reception of approval from Twitter, allowing the collection of data for this research and the following yield of a Developer API.

Table 1 Research hypotheses

#		Research hypotheses
1	H0	Sentimental analytics on twitter data is not enough to predict with success the outcome of an electoral result
1	H1	A sentiment analysis of twitter data can predict with success an electoral result
2	H0	The sentiment of the tweets does not reflect the voting decision of the electorate
2	H2	The sentiment of the tweets will be positively correlated with the vote decision of the electorate

The collection of data was then processed with the use of R (R [Programming Language] 1993), a statistical programming language, and the application of the package “httr” (Wickham 2020). This package allowed the interface with the Twitter Developer API to collect the data from this Social Network using the package “rtweet” (Kearney 2020). Using the Search Term of the “rtweet” package was specified the terms that should be present in a collected tweet so that the prediction could occur.

The criteria implied in the code was the collection of tweets mentioning:

- @ that defines the name of the official account from each party;
- @ that defines the name of the official account from each candidate as prime minister;
- # mentioning the name of the party;
- # mentioning the name of the candidate as prime minister;
- # mentioning a concept that characterized the party (for example, Brexit for the Conservative Party).

From this process, it was possible to collect between ten days before the election and the day before the following number of tweets mentioning the major British parties (Table 2).

Afterward, a first treatment on the data was applied as multiple tweets were collected in duplicate. This situation occurred as part of the mentions contained both one of the @ or # mentioned previously. By cleaning the data from this duplication using the Excel function to remove duplicates from a table, the datasets were cleaned, avoiding a possible bias on the training model that could compromise the results of this research. After this procedure, the dataset from each party consisted of 8375 rows from the Conservative Party and 6269 from the Labor Party.

After the treatment of the datasets was finished, this study was confronted with two possible learning applications for the prediction: Supervised Learning or Unsupervised Learning. There is no consensus on the algorithm applied in a Sentiment Analysis in this research field. Therefore, the decision was to proceed with both learning processes to discover which produces the most accurate prediction results.

Table 2 Data collected for each political party dataset

Conservative party	Labor party
@conservatives—1938	@UKLabour—1620
#conservatives—1868	@jeremycorbyn—1571
@borisjohnson—2065	#labourparty—1763
#borisjohnson—1699	#jeremycorbyn—1505
#brexit—1424	

4.1 Supervised learning

The Supervised Learning on the training of a Sentiment Analysis model, as the name refers, follows a Supervised method. A criterion is searched to allow the researcher to decide if a sample of data belongs to a certain class based on examples known as a priori (Langer et al. 2020). This method requires a dataset labeled as positive/negative, so the algorithm, during the training, can understand if a word, when used in a specific sequence, has the probability of carrying positive or negative sentiment.

This learning type is the most favored by the literature on this subject area. On the one hand, it is the preferred learning type for Sentiment Analysis (Hemalatha et al. 2013). On the other hand, because of the complexity of this theme, which obligates better support on the algorithm to identify the sentiment in Social Media text, usually filled with irony and sarcasm, a figure of speech that current sentiment analysis techniques struggle to capture (Cambria et al. 2022). Sarcasm is not the only challenge in sentiment analysis. Other challenges include polarity disambiguation (Cambria et al. 2022), filtering neutral opinions and ambivalent opinions (opinions with both positive and negative sentiments) since these opinions can influence the overall perception of the sentiment (Chan et al. 2023; Rahmani et al. 2023). These issues have been addressed by new frameworks and techniques primarily based on neural networks and graph architectures (Cambria et al. 2022; Chan et al. 2023; Dai et al. 2021; Rahmani et al. 2023).

The model's data processing, prediction, and validation were made on the software platform RapidMiner (RapidMiner [Computer Software] 2001). This data science platform has an integrated environment that allows the treatment of data, machine learning, deep learning, and the selected method to research the first hypothesis of this study, predictive analysis.

For the model training for each British party, a total of 1200 tweets from each dataset were selected. The researcher manually labeled the sentiment of these tweets as true positive or true negative sentiment. This labeling was necessary to train the model in identifying the probability of the expressions carrying a sentiment by the words present on the comment and their sequence. For this, all comments with neutral feelings had to be retrieved from each training set to avoid bias in the learning process of the models. Selecting neutral comments and removing these from each training set left 469 tweets from the Conservative Party (169 positives vs. 288 negatives) and 540 tweets from the Labor Party (277 positives vs. 263 negatives) to be used to train the predictive model.

The remaining data collected from each party not used for training purposes were kept to the side, as it was meant to be later applied to the result of the prediction model, after

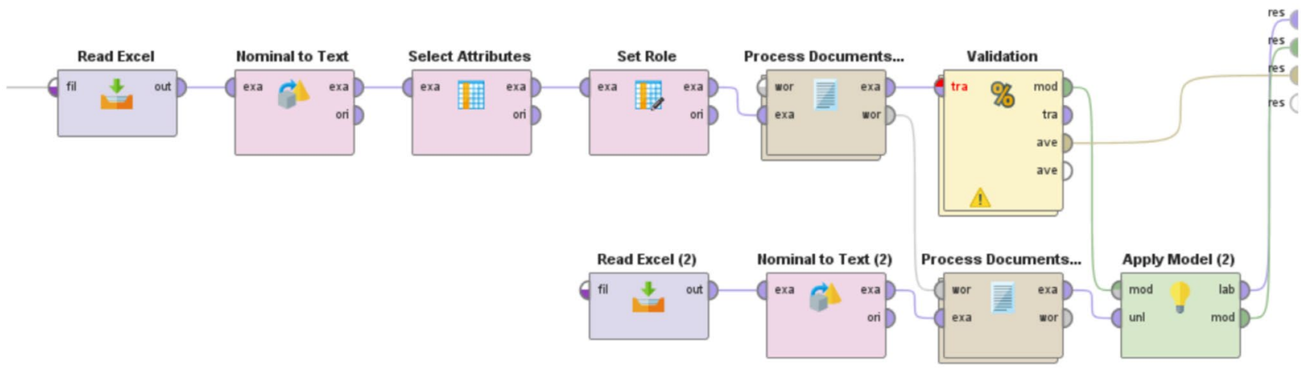


Fig. 1 Structure of the prediction model in RapidMiner

being tested and validated, to predict the amount of sentiment present in those tweets (Fig. 1).

With the training data imported into the RapidMiner platform, the column carrying the Sentiment of the dataset was identified in the target role as "label." The identification of this label indicated to the algorithm which classes existed in the model and where it should investigate the criteria that comments to carry a sentiment.

The next step was preparing the training data for the actual training. Several steps were applied to the dataset to prepare it to be interpreted by the algorithm See Figs. 3 and 4. The **Transform Cases** transformed all the dataset's content in lowercase, so the model could quickly identify the same words in the comments without having similar words in uppercase and lowercase.

Afterward, **Tokenize** was applied to divide the text into a sequence of tokens. This module divided the text into small words, allowing the model to identify the words in the dataset quickly.

Two different filters were applied to the training datasets. **Filter Stopwords (English)** removed all the tokens that did not contribute to the learning process of the model, such as connectors and pronouns. **Filter Tokens (By Content)** removed all the unnecessary content identified by the researcher from the datasets. In this case, it was specified in this field that all the "https" should be removed from the training set. As "https" was a constant in most tweets, representing hyperlinks of news, images, and videos that did not contribute to this analysis, they had to be removed from the training set to avoid bias in the learning process.

In the end, the **Stem (Porter)** grouped all tokens from the same family into a single token. This way, words like "Conservative" and "Conservator" were associated as unique words representing the same value for the analysis.

The model went to the Split Validation module with the data treated, where the test to the actual prediction capability occurred. In this step, the training set had to be divided into two phases, training, and testing, with, respectively, 70% and 30% of the data. The training phase requires the most data as the algorithm learns by understanding which words contribute the most to a positive/negative label. On the other hand, the testing phase is where the model concludes on the accuracy by which the algorithm could predict each comment's label correctly.

An algorithm is required to train a sentiment analysis model. Many algorithms existent in the Machine Learning study area were selected for this research. The criterion used for the choice was the good results presented in the previous literature on this topic and the reputation in the study field (Hemalatha et al. 2013). The algorithms used in this research were: Naïve Bayes, Support Vector Machine, Decision Tree, and Deep Learning.

The first two algorithms are considered the most popular Machine Learning algorithms to solve Sentiment Analysis problems. Starting with Naïve Bayes is known as the simplest one used algorithm for predictions. It computes the probability of each class, based on the distribution of the words in the document, following the theory of Bayes (Friedman et al. 1997).

Naïve Bayes Theorem:

$$P(label|features) = \frac{P(label) * P(features|label)}{P(features)} \tag{1}$$

This theory, as shown underneath, calculates the probability of a text carrying a particular label by multiplying the probability of a label, in this case, positive or negative, with the probability of a specific word being classified as a label and dividing this by the probability of a certain feature occurring.

The second algorithm, Support Vector Machine (SVM), is a linear regression model. The primary purpose of the SVM is to determine the linear separation that best suits the separation of different classes from each other (Hearst et al. 1998). The best suitor is the plane that offers, with more confidence, a maximum margin of difference between the data classes. This algorithm is also robust to outliers.

Another popular technique used in data mining for prediction is the Decision Tree. The algorithm breaks the dataset into tiny sets until each belongs to a particular class. As the name says, the module of decisions is built in the form of a tree, where the module decides if the set is an A or a B. The accuracy of decision trees is usually poor. Still, since it is easy to understand the relationship between the branches, makes this algorithm a very popular choice in the field (Suzuki 2020).

As an alternative to the traditional methods, this research investigated the use of the Deep Learning algorithm. The main idea of this type of technique is to learn with the most complex features extracted from the training set, using neural networks to build computational models (Pouyanfar et al. 2019). However, this well-known algorithm in the machine learning field carries a disadvantage. The model to be well-trained requires a large quantity of data.

Each of the four algorithms was applied to the models in the training phase to test which produced better accuracy See Figs. 5 and 6.

4.2 Unsupervised learning

An Unsupervised learning model follows a different approach from the Supervised learning type. The training dataset consists of a set of inputs without an assigned label, leaving the learning algorithm to discover the structure for the input. The algorithm discovers patterns in the dataset and clusters them in groups of similar examples. This type of learning is applied to achieve one of two goals: discovery of hidden patterns in the data or, as it was used for this research, as a learning feature (Greene et al. 2008).

This type of method is not the most used as a learning method for Sentiment Analysis due to the complexity of the content from social media. However, a special algorithm outperformed the accuracy of previously mentioned algorithms in this research field: The Valence Aware Dictionary for Sentiment Reasoning (Vader) (Gilbert and Hutto 2015). Vader is a lexicon rule-based sentiment analysis tool built to analyze the sentiment expressed in Social Media text and many other domains. Vader uses a set of qualitative and quantitative methods, including a list of lexical features usually found in Social Network contexts. Vader combines lexical features with grammatical and syntactical conventions to express and empathize with sentiment intensity. Vader's effectiveness can be compared to lexical-based methods such as the SentiWordNet and machine learning algorithms like Support Vector Machine and Naïve Bayes.

Gilbert and Hutto (2015) responsible for the development of this Sentiment analysis engine concluded; it performs exceptionally well in Social Media content, as Vader was able to outperform the accuracy of human-rated labeled training sets by correctly classifying the sentiment of tweets in positive, neutral, and negative classes when applied to the same dataset. It also shows more benefits than the application of traditional sentiment lexicons, as it can be used on more extensive sets of data, is more sensitive to sentiment expressions in social media contexts and can quickly be applied to a dataset without the need for training.

The Vader Package in R code was used to apply this unsupervised learning method. Using the script vader_df() on a data frame, it was possible to calculate the valence score of each text, and the weighted percentage of positive, negative, and neutral words (Roehrick 2020), for each dataset.

5 Results and discussion

In applying the supervised learning method, each algorithm used in the training of the predictive models produced a certain amount of accuracy. The accuracy can be described as measuring how correctly the predictive model can identify the label during a prediction. This measure appears in the validation step of the training process, where the remaining 30% of the training set not used to train the algorithm is now applied to confirm how well the model could predict a result. In the case of this research, the accuracy rate was measured on how correctly the models predicted the main sentiment in the comments mentioning the two political parties (Table 3 and detailed by model and party from Figs. 7, 8, 9, 10, 11, 12, 13, 14).

In the validation results on the accuracy rate of the four algorithms, it is visible that the Naïve Bayes and the Support Vector Machine were the algorithms achieving a higher precision in their predictive capabilities on both training sets. These results follow the findings of studies in this field, where the same algorithms have proven to produce more reliable prediction results than the rest.

Table 3 Accuracy rate from each algorithm of the supervised learning method

	Conservative party (%)	Labor party (%)
Naïve Bayes	78.57	74.07
Support vector machine	78.57	77.31
Decision tree	77.14	71.30
Deep learning	76.43	65.74

Observing the above table, while in the Labor Party predictive model, the higher accuracy was present in the model trained by the SVM algorithm, the Conservative Party predictive model achieved the same results while using this algorithm and Naïve Bayes. With these results, additional prediction measurements, such as precision and recall, were overlooked to understand if one of the algorithms was more reliable than the other. In the case of the use case with the SVM algorithm, the training set of the Conservative Party was able to produce a 9% higher precision on the positive class and a 4.41% higher recall of true negatives compared with the Naïve Bayes application. As the difference was not substantial, both models were accepted as capable models of predicting the sentiment for the Conservative Party.

With the predictive models validated and with a 78% accuracy rate on sentiment prediction, the models were now able to receive the rest of the dataset not used for training or validation purposes and identify the amount of true positive and true negative sentiment present in those tweets mentions.

The remaining 7375 tweets from the training set were applied to both predictive models starting with the Conservative Party dataset. Regardless of the algorithm used, it was possible to determine the same conclusion: The percentage of negative sentiment in tweets mentioning the Conservative Party was significantly higher than the positive (Table 4).

The application of the same process in the Labor Party predictive model, on the other hand, produced an opposite result. Applying the remaining 5036 tweets of this party dataset to the prediction model learned with the Support Vector Machine algorithm predicted 63% of positive sentiment present in this party's tweets before Election Day (Table 5).

To be considered a successful election predictor, the predictive model of the Conservative Party, as a representation of the actual winner of this electoral moment with an absolute majority, needed to indicate most of the positive sentiment on the tweets mentioning this party when compared with its opposition.

The results from applying the Supervised method revealed a negative correlation between the sentiments transmitted by the Social Media users on both parties. While the Conservative Party prediction model verified a higher negative sentiment in its tweets, the Labor model estimated 63% of positive sentiment. With these results, it was possible to conclude that these predictive models could not successfully predict the outcome of this electoral period, as the winner of this electoral cycle failed to show an estimation of positive sentiment superior to the opposition.

Table 4 Results of the supervised learning method on the conservative party dataset

	Positive sentiment	Negative sentiment
Naïve Bayes	2277	5098
Support vector machine	1122	6253

Table 5 Results of the supervised learning method on the labor party dataset

	Positive sentiment	Negative sentiment
Support vector machine	3183	1853

Accordingly, it was also possible to observe that the results of this estimation did not reflect the vote decision of the electorate. This observation means the supervised learning method could not verify the second hypothesis of this research. This leads us to conclude that Social Media content was not enough, in this case, study, to influence the electors' decision, and other factors, such as personal life events, may have had a higher impact on this process.

Moving on to the results of applying the unsupervised learning method to this study, it did not reveal clear results as the supervised learning method. By learning from the data, the Vader algorithm retrieved an average amount of positive/negative/neutral feelings present in the dataset from each political party (Table 6).

Analyzing the Sentiment Analysis results, it was impossible to make a concrete conclusion on this analysis. The average amount of neutral feeling occupied more than 80% of the content from each dataset. As the rest of the 20% was formed by a similar amount of positive and negative tweets, it was impossible to conclude if this model had efficient predictor capabilities with such small data.

From the obtained results in the estimations of both learning methods, it was possible to conclude that this model failed to predict the electoral results of the 2019 General Elections in the United Kingdom. These results may be explained by the limitations shown previously, such as the lack of representation from the overall population in this Social Network.

Table 6 Results of the unsupervised learning method on the parties dataset

	Conservative party (%)	Labor party (%)
Average positive	8.05	9.2
Average negative	8.02	7.52
Average neutral	83.93	83.29

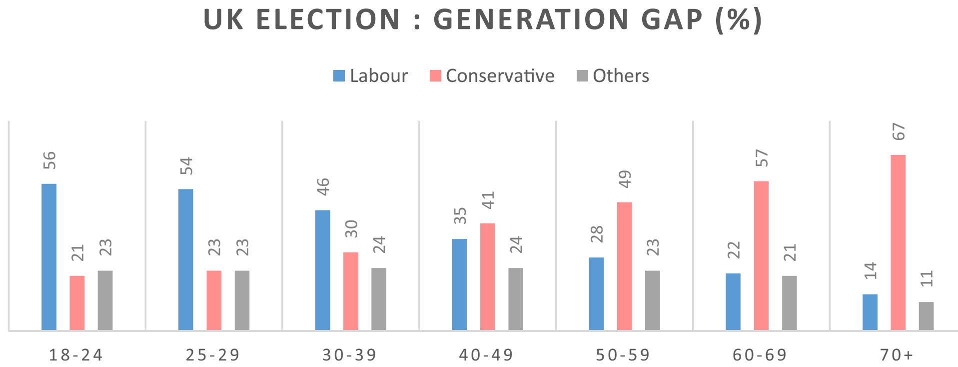


Fig. 2 Post-Election Survey on the vote decision of the British electors. *Source*: YouGov

To corroborate this affirmation, a survey was conducted after this electoral period by a market research and data analytics firm, YouGov, to 41,995 British adults to understand their voting behavior in these past elections. The results of this survey unveiled a clear preference by older generations (+ 50 years old) for Conservative politics, while the youngest voters (18–39) were more apologetic for Labor politics (Curtis 2019) (Fig. 2).

The Digital Report 2020 for the United Kingdom showed the same age range that presented a preference for Labor politics and was the most active generation on Social Networks (Kemp 2020). Looking at both factors together, it is evident the reason why most of the positive sentiment was concentrated in tweets mentioning Labor politics and why the amount of negative sentiment for the Conservatives did not seem to demote the electorate from voting for this party: the target audience most likely to vote Conservative was not present in Twitter.

Therefore, this estimation corroborates the discussion that Social Media data are essential to understand the opinion polarization effect and the sentiment of individuals toward a topic. However, all by itself, it cannot predict the winner of an election due to the limitation mentioned above.

Furthermore, the second learning approach results reiterate the conclusions made with the supervised learning method regarding the second hypothesis. With the amount of positive and negative sentiment representing less than 20% of the content from the datasets, the probability of a Social Media user being in contact with a positive/negative comment about a party was much lower than a feeling-less comment. This means comments made by "prosumers" that could influence others' vote decisions were restricted to less than 20% of the million tweets published in that period. This research failed to conclude whether online persuasion can influence the decision of an electorate, leaving this literature gap still to be filled.

6 Conclusion, limitations, and recommendations

This study had as its primary purpose to collaborate in the research on the assumption that Social Media content could be used as an election predictor and conclude whether political persuasion in these Social Networks had an actual impact on the voting decision of the electorate.

According to the literature review, it was possible to understand that the first topic was not a novelty in this academic area as multiple researchers contributed to it throughout the years. However, during that time, other researchers have refuted some studies for not following the correct procedures, leading to biased results. This circumstance led to several open gaps in the literature, by which this study purposed to collaborate on the fill. For the second topic, the literature review helped prove the existence of persuasion in social media made by "prosumers", making it an environment propitious for the rise of the opinion polarization effect in these networks. Nevertheless, this effect's actual impact on election results was yet to be proved.

From these problematics, the two hypotheses this research aimed to investigate were defined: A Sentiment Analysis on Twitter data can predict with success an electoral result; The Sentiment of the tweets will be positively correlated with the vote decision of the electorate.

To prove the veracity of these hypotheses was applied as a methodology, a Sentiment Analysis of tweets mentioning the United Kingdom General Election of December 2019 was used as a case study for this research. This text mining application followed two approaches: a Supervised and an unsupervised learning method.

Both learning methods were applied to the Sentiment Analysis models to test the prediction capabilities of Twitter data. Looking at the results, the Supervised method revealed a negative correlation between the sentiments of both parties in social media: The Conservatives in this electoral period showed a majority of 69% of negative sentiment on the tweets mentioning this party. On the other hand, the model estimated 63% of positive opinions in the comments regarding the Labor Party. The Unsupervised Learning type, however, could not achieve a similar conclusion. The feeling-less comments occupied 80% of the datasets, making it impractical to conclude with this estimation. With only the first method able to reflect on the first hypothesis of this research, as the estimation wrongly predicted the winner of this election, the hypothesis was refuted.

The results of the Supervised Learning estimation also showed that the sentiment of the tweets was not positively correlated with the vote decision of the electorate. Since the Conservative Party was the winner of this electoral period, for the social media content to impact the vote decision, it was expected that the number of tweets mentioning this party would show positive feelings toward it. The outcome from the Unsupervised method estimation explained these results: Considering the amount of sentiment data represented less than 20% of the content from the datasets, this means the probability of a Social Media user being in contact with a true positive/negative comment was much lower than a feeling-less comment. Accordingly, it was not possible to prove or refute the second hypothesis of this research, leaving an open gap in the literature still to be filled.

In conclusion, this research methodology failed to prove the two proposed hypotheses' veracity. This outcome can be explained by the several limitations unveiled in this research.

Starting from the already mentioned lack of representation from the overall population in this prediction, the strong presence of younger generations in Social Networks is well known. Crossing this information with the survey results conducted by YouGov after the elections, which revealed a clear political preference from this generation, helped explain the results: the audience most likely to vote for the winning party was not present in Social Networks. For this reason, this study corroborates the assumption that social media should be used as a complement to electoral forecasts instead of replacing them to unveil the reason behind the voting decision of the electors.

Another limitation of this research was the amount of data collected. The sentiment of tweets classified as neutral in the unsupervised learning method represented around 80% of the overall dataset, significantly reducing the amount of data that could be used to predict the outcome of an election. With only 20% of tweets classified with true positive or negative sentiment, it was impossible to conclude with this estimation's results. This limitation resulted from the collection being accomplished only with the mentions of the parties in this Social Network, leaving out of this analysis all the tweets about this topic without mentioning the official accounts of the politicians.

Still, on the topic of data collection, another constraint was the limitations of the Twitter API itself. This API allows the capture of a representative sample of tweets and a maximum of 100,000 requests per day, which may not be enough for more in-depth research (at the moment, this research was done).

As a recommendation for future research, collecting a more considerable amount of data would be beneficial to achieve better accuracy in estimating the prediction models. This data collection can be achieved by enlarging the data collection scope by including tweets mentioning the election itself and trend topics involving the political parties. The application of an econometric regression would help the researcher understand which factors may be positively correlated with the vote decision of an elector and whether online persuasion in Social Media is one of those factors.

Appendix

A.1 Rapidminer

See Figs. 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, and 14

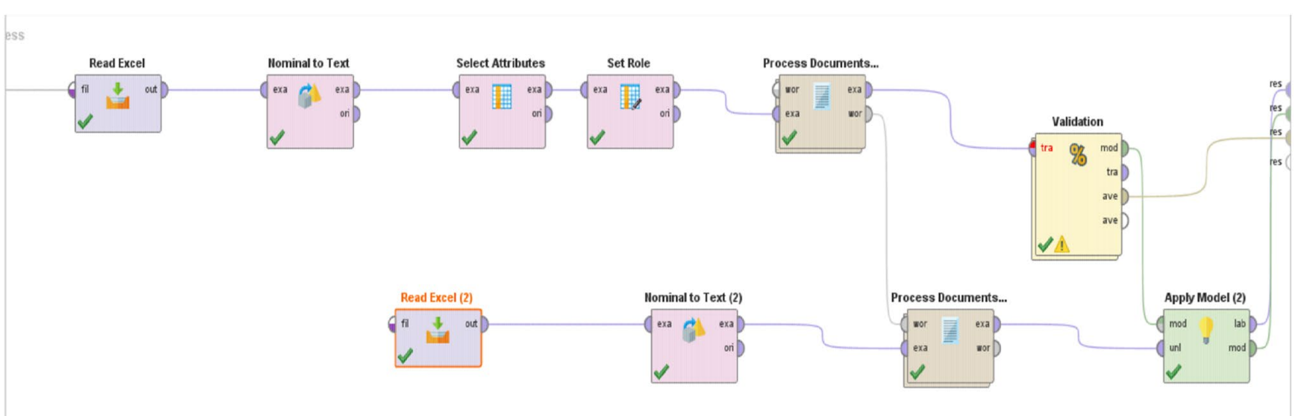


Fig. 3 Supervised Training (RapidMiner)—Overall prediction model

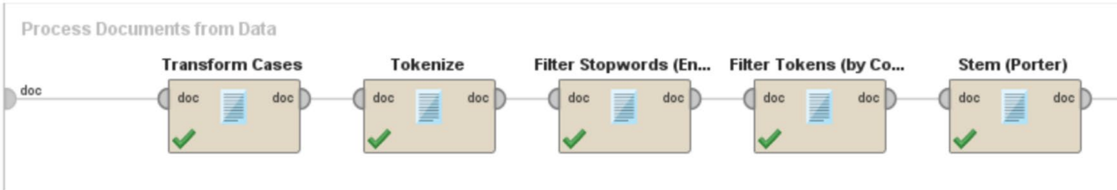


Fig. 4 Supervised Training (RapidMiner)—Process Documents from Data

Fig. 5 Supervised Training (RapidMiner)—Applied algorithms in the training model

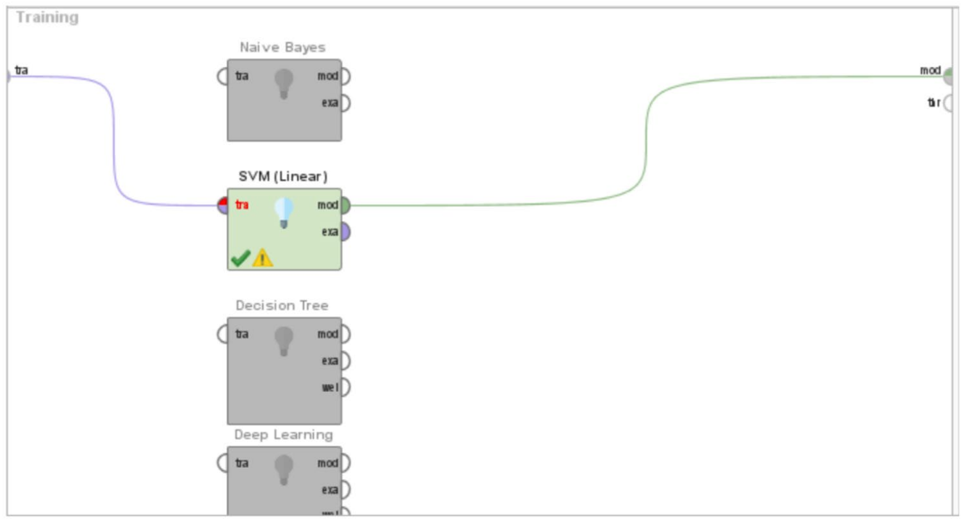
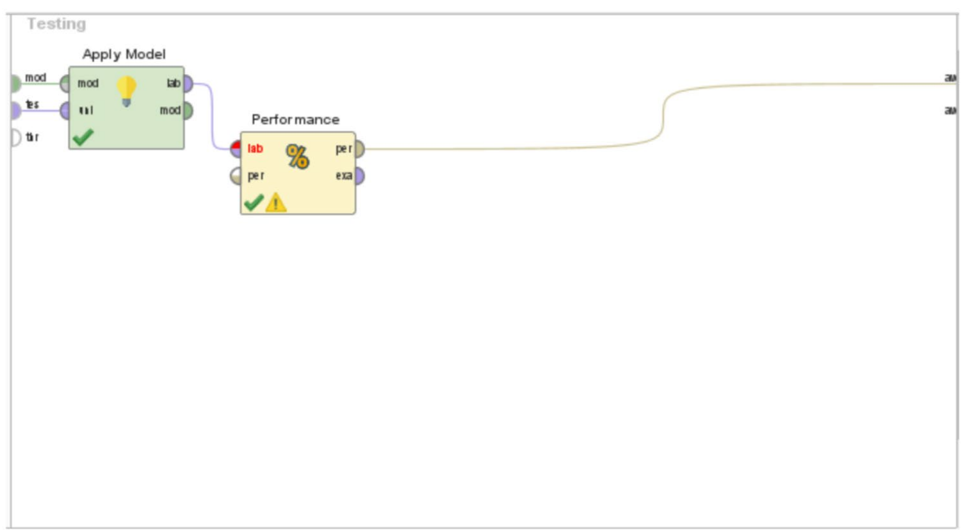


Fig. 6 Supervised Training (RapidMiner)—Testing the training model performance



accuracy: 78.57%

	true negative	true positive	class precision
pred. negative	79	23	77.45%
pred. positive	7	31	81.58%
class recall	91.86%	57.41%	

Prediction			Least	Most	Values
prediction(Sentiment)	Text	0	positive (2277)	negative (5098)	negative (5098), positive (2277)

Fig. 7 Results of the Conservative Party training model applied Naïve Bayes

accuracy: 78.57%

	true negative	true positive	class precision
pred. negative	83	27	75.45%
pred. positive	3	27	90.00%
class recall	96.51%	50.00%	

Prediction			Least	Most	Values
prediction(Sentiment)	Text	0	positive (1122)	negative (6253)	negative (6253), positive (1122)

Fig. 8 Results of the Conservative Party training model when applied SVM

accuracy: 77.14%

	true negative	true positive	class precision
pred. negative	81	27	75.00%
pred. positive	5	27	84.38%
class recall	94.19%	50.00%	

Prediction			Least	Most	Values
prediction(Sentiment)	Text	0	negative (2822)	positive (4553)	positive (4553), negative (2822)

Fig. 9 Results of the Conservative Party training model when applied Decision Trees

accuracy: 76.43%

	true negative	true positive	class precision
pred. negative	77	24	76.24%
pred. positive	9	30	76.92%
class recall	89.53%	55.56%	

Prediction			Least	Most	Values
prediction(Sentiment)	Text	0	positive (2266)	negative (5109)	negative (5109), positive (2266)

Fig. 10 Results of the Conservative Party training model when applied Deep Learning

accuracy: 74.07%

	true Negative	true Positive	class precision
pred. Negative	81	32	71.68%
pred. Positive	24	79	76.70%
class recall	77.14%	71.17%	

Prediction			Least	Most	Values
prediction(Sentiment)	Text	0	Negative (2192)	Positive (2844)	Positive (2844), Negative (2192)

Fig. 11 The Labor Party training model results when Naïve Bayes is applied

accuracy: 77.31%

	true Negative	true Positive	class precision
pred. Negative	83	27	75.45%
pred. Positive	22	84	79.25%
class recall	79.05%	75.68%	

Prediction			Least	Most	Values
prediction(Sentiment)	Text	0	Negative (1853)	Positive (3183)	Positive (3183), Negative (1853)

Fig. 12 Results of the Labor Party training model when applied SVM

accuracy: 71.30%

	true Negative	true Positive	class precision
pred. Negative	98	55	64.05%
pred. Positive	7	56	88.89%
class recall	93.33%	50.45%	

Prediction			Least	Most	Values
prediction(Sentiment)	Text	0	Positive (1011)	Negative (4025)	Negative (4025), Positive (1011)

Fig. 13 Results of the Labor Party training model when applied Decision Trees

accuracy: 65.74%

	true Negative	true Positive	class precision
pred. Negative	104	73	58.76%
pred. Positive	1	38	97.44%
class recall	99.05%	34.23%	

Prediction			Least	Most	Values
prediction(Sentiment)	Text	0	Positive (2161)	Negative (2875)	Negative (2875), Positive (2161)

Fig. 14 Results of the Labor Party training model when applied Deep Learning

Author contributions All authors were involved in writing the main manuscript as well as in reviewing it before submission.

Funding Open access funding provided by FCT|FCCN (b-on). This work was supported by national funds through FCT (Fundação para a Ciência e a Tecnologia), under the project - UIDB/04152/2020 - Centro de Investigação em Gestão de Informação (MagIC)/NOVA IMS.

Declarations

Competing interests The authors declare no competing interests.

References

Allgaier J (2018) Myspace. SAGE Int Encycl Mass Media Soc 42(1):1156–1158

Anstead N, O'Loughlin B (2015) Social media analysis and public opinion: the 2010 UK general election. J Comput Mediat Commun 20:204–220

Bimber B (2014) Digital media in the Obama campaigns of 2008 and 2012: adaptation to the personalized political communication environment. J Inform Technol Politics 11(2):130–135

Blumer H (1948) Public opinion and public opinion polling. Am Sociol Rev 13(5):542–549

Broersma M, Graham T (2012) Social media as beat. Journal Pract 6(3):403–419

Cambria E (2016) Affective computing and sentiment analysis. IEEE Intell Syst 31(2):102–107. https://doi.org/10.1109/MIS.2016.31

Cambria E, Liu Q, Decherchi S, Xing F, Kwok K (2022) SenticNet 7: a commonsense-based neurosymbolic AI framework for explainable sentiment analysis. In: 2022 language resources and evaluation conference, LREC 2022, pp 3829–3839

Chan C-H, Fu K-WT (2017) The relationship between cyberbalkanization and opinion polarization: time-series analysis on Facebook pages and opinion polls during the Hong Kong occupy movement and the associated debate on political reform. J Comput Med Commun 22(5):266–283

Chan JYL, Bea KT, Leow SMH et al (2023) State of the art: a review of sentiment analysis based on sequential transfer learning. Artif Intell Rev 56:749–780. https://doi.org/10.1007/s10462-022-10183-8

Chung J, Mustafaraj E (2011) Can collective sentiment expressed on Twitter predict political elections? In: Twenty fifth AAAI conference on artificial intelligence. Wellesley, San Francisco, pp 1170–1071

Curtis C (17 de December de 2019) How Britain voted in the 2019 general election. https://yougov.co.uk/topics/politics/articles-reports/2019/12/17/how-britain-voted-2019-general-election

Dai Z, Liu Y, Di S, Fan Q (2021) Aspect-level sentiment analysis merged with knowledge graph and graph convolutional neural network. J Phys Conf Ser. https://doi.org/10.1088/1742-6596/2083/4/042044

Diehel T, Weeks BE, Zúñiga HG (2015) Political persuasion on social media: tracing direct and indirect effects of news use and social interaction. SAGE J 34(5):302–315

DiMaggio P, Evans J, Bryson B (1996) Have American's social attitudes become more polarized? Am J Sociol 102:690–755

Enli G, Naper AA (2016) Social media incumbent advantage: Barack Obama's and Mitt Romney's tweets in the 2012 U.S. presidential election campaign. In: Bruns A, Enli G, Skogerbø E (eds) The Routledge companion to social media and politics. Routledge Handbooks Online, New York, pp 364–378

Enli G, Simonsen C-A (2017) 'Social media logic' meets professional norms: Twitter hashtags usage by journalists and politicians. Inf Commun Soc 21(8):1081–1096

Friedman N, Geiger D, Goldszmidt M (1997) Bayesian network classifiers. Mach Learn 29:31–16

Gilbert E, Hutto CJ (2015) VADER: a parsimonious rule-based model for sentiment analysis of social media text. In: Proceedings of eighth international AAAI conference on weblogs and social media. Georgia Institute of Technology, Atlanta, pp 1–10

Greene D, Cunningham P, Mayer R (2008) Unsupervised learning and clustering. In: Cord M, Cunningham P (eds) Machine learning techniques for multimedia. Springer, Berlin, pp 51–90

Healy AJ, Malhotra N, Mo CH (2009) Personal emotions and political decision making: implications for voter competence. Stanford Business, Stanford, pp 1–48

Hearst MA, Dumais ST, Platt J, Scholkopf B (1998) Support vector machines. IEEE Intell Syst Appl 13:18–28

Hemalatha I, Saradhi Varma GP, Govardhan A (2013) Sentiment Analysis tool using machine learning algorithms. Elixir Int J 58:14791–14794

Jugherr A, Juergens P, Schoen H (2012) Why the pirate party won the german election of 2009 or the trouble with predictions: a response to Tumasjan, A., Sprenger, T. O., Sander, P. G., & Welpe, I. M. Predicting elections with Twitter: what 140 characters reveal about political sentiment. Soc Sci Comput Rev 30(2):229–234

Kearney MW (9 de January de 2020) Package 'rtweet'. Retrieved from The Comprehensive R Archive Network: https://cran.r-project.org/web/packages/rtweet/rtweet.pdf

Kemp S (February 12, 2020) Digital 2020: the United Kingdom. https://datareportal.com/reports/digital-2020-united-kingdom?rq=united%20kingdom

Langer H, Falsaperla S, Hammer C (2020a) Chapter 2: Supervised learning. In: Langer EHF (ed) Advantages and pitfalls of pattern recognition. Elsevier, Amsterdam, pp 33–85

Larsson AO, Kalsnes B (2014) 'Of course, we are on Facebook': use and non-use of social media among Swedish and Norwegian politicians. Eur J Commun 29(6):653–667

Lazarsfeld P (1957) Public opinion and the classical tradition. Public Opin Q 21(1):39–53

Lee FL (2016) Impact of social media on opinion polarization in varying times. Commun Pub 1(1):56–71

León-Borges AJ, Balam RI, Gómez LR, Strand MP (2015) The machine learning in the prediction of elections. ReCIBE 4(2):1

Maes M, Bischofberger L (2015) Will the personalization of online social networks foster opinion polarization? SSRN, 1–36

Makazhanov A, Rafiei D (2013) Predicting political preference of twitter users. In: IEEE/ACM international conference on advances in social networks analysis and Mining. Niagara Falls: IEEE, pp 298–305

Medhat W, Hassan A, Korashy H (2014) Sentiment analysis algorithms and applications: a survey. Ain Shams Eng J 5:1093–1113

Minshall K (19 de December de 2019) UK election conversation attracts over 15m Tweets. Retrieved from Twitter: https://blog.twitter.com/en_gb/topics/company/2019/uk-election-conversation-attracts-over-fifteen-million-tweets.html

Moro S, Rita P (2018) Brand strategies in social media in hospitality and tourism. Int J Contemp Hosp Manag 30(1):343–364

Moro S, Rita P, Vala B (2016) Predicting social media performance metrics and evaluation of the impact on brand building: a data mining approach. J Bus Res 69(9):3341–3351

Nave M, Rita P, Guerreiro J (2018) A decision support system framework to track consumers sentiments in social media. J Hosp Market Manag 27(6):693–710

Ortiz-Osina E (18 de September de 2019) The rise of social media. Retrieved from Our World in Data: https://ourworldindata.org/rise-of-social-media

Park CS (2019) Learning politics from social media: interconnection of social media use for political news and political issue and process knowledge. Commun Stud 70(3):253–276

Pinto L, Loureiro S, Rita P, Sarmento E (2019) Fostering online relationships with brands through websites and social media brand pages. J Promot Manag 25(3):379–393

Pouyanfar S, Sadiq S, Yan Y, Tian H, Tao Y, Reyes MP, Iyengar SS (2019) A survey on deep learning: algorithms, techniques, and applications. ACM Comput Surv 51(5):1–36

R [Programming Language] (1993) Retrieved from The R Project for Statistical Computing: https://www.r-project.org/

Rahmani S, Hosseini S, Zall R, Kangavari MR, Kamran S, Hua W (2023) Transfer-based adaptive tree for multimodal sentiment analysis based on user latent aspects. Knowl Based Syst 261:110219. https://doi.org/10.1016/j.knosys.2022.110219

Ramos R, Rita P, Moro S (2019) From institutional websites to social media and mobile applications: a usability perspective. Eur Res Manag Bus Econ 25(3):138–143

Rapidminer [Computer Software] (2001). https://rapidminer.com/

Roehrick K (7 de September de 2020) Package 'vader'. Retrieved from The Comprehensive R Archive Network: https://cran.r-project.org/web/packages/vader/vader.pdf

Romão T, Moro S, Rita P, Ramos P (2019) Leveraging a luxury fashion brand through social media. Eur Res Manag Bus Econ 25(1):15–22

Sang ET, Bos J (2012) Predicting the 2011 Dutch senate election results with Twitter. In: Workshop on semantic analysis in social media. Association for Computational Linguistics, Stroudsburg, pp 53–60

Skoric MM, Liu J, Jaidka K (2020) Electoral and public opinion forecasts with social media data: a meta-analysis. Information 11(4):187

Stieglitz S, Dang-Xuan L (2012) Social media and political communication: a social media analytics framework. Soc Netw Anal Min 3(4):1272–1291

Suzuki J (2020) Decision trees. Statistical learning with math and R. Springer, Berlin, pp 147–170

Tumasjan A, Sprenger T, Sander PG, Welpe IM (2010) Predicting elections with twitter: what 140 characters reveal about political sentiment. Soc Sci Comput Rev 30(2):229–234

Usluel YK, Mazman GM (2009) Adoption of web 2.0 tools in distance education. Procedia Soc Behav Sci 1:818–823

Weeks BE, Ardèvol-Abreu A, Zúñiga HG (2015) Online influence? Social media use, opinion leadership, and political persuasion. Int J Public Opin Res 29(2):214–239

Wei L, Hindman DB (2011) Does the digital divide matter more? Comparing the effects of new media and old media use on the education-based knowledge gap. Mass Commun Soc 14(2):216–235

Westbrook I (2019) General election 2019: a really simple guide. Retrieved from BBC: https://www.bbc.com/news/uk-politics-49826655

Wickham H (2020) Package 'httr'. Retrieved from The Comprehensive R Archive Network: https://cran.r-project.org/web/packages/httr/httr.pdf

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.