Multimedia Tools and Applications (2023) 82:9277–9302
https://doi.org/10.1007/s11042-022-13438-2

Searching for associations between social media trending topics and organizations

João Henriques[1] · **João Ferreira**[1]

Received: 21 December 2020 / Revised: 11 May 2022 / Accepted: 2 July 2022 /
Published online: 19 August 2022
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract
Trending topics are the most discussed topics at the moment on social media platforms, particularly on Twitter and Facebook. While the access to trending topics are free and available to everyone, marketing specialists and specific software are more expensive, therefore there are companies that do not have the budget to support those costs. The main goal of this work is to search for associations between trending topics and companies on social media platforms and HotRivers prototype was developed to fill this gap. This approach was applied to Twitter and used text mining techniques to process tweets, train personalized models of companies and deliver a list of the matched trending topics of the target company. So, in this work were tested different pre-processing text techniques and a method to select tweets called Centroid Strategy used on trending topics to avoid unwanted tweets. Also, were tested three models, an embedding vectors approach with Doc2Vec model, a probabilistic model with Latent Dirichlet Allocation, and a classification task approach with a Convolutional Neural Network used on the final architecture. The approach was validated with real cases like Adidas, Nike and Portsmouth Hospitals University. In the results stand out that trending topic *Nike* has an association with the company Nike and *#WorldPatientSafetyDay* has an association with Portsmouth Hospitals University. This prototype, HotRivers, can be a new marketing tool that points the direction to the next campaign.

Keywords Text mining · Text similarity · Text classification · Convolutional neural network · Doc2Vec · Latent Dirichlet allocation

1 Introduction

The number of Internet users is growing exponentially, as well as the offer of new social media platforms over the course of the years, and the amount of users has also increased substantially [26, 32]. Social media has the incredible power of making information,

✉ João Henriques
jpshs@iscte-iul.pt

João Ferreira
jcafa@iscte-iul.pt

[1] Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR-IUL, 1649-026 Lisboa, Portugal

opinions and complains accessible to everyone [10]. An example that illustrates this fact is the video that captured the death of George Floyd, which gained international attention after hitting an astonishing number of comments, likes and shares across multiple social media platforms. As a result of that, protests appeared across the world [36].

The following examples given by the authors [13], in 2010, illustrate the power that social media has to destroy a marketing campaign and products reputation. The case study about Motrin, a medicine that was so criticized that climbed to the top of trending topics on Twitter and gained such visibility that reached mainstream media. One impressive fact is that all this happened throughout a period of 24 hours during a weekend. The other case study is related to a milk-based product, Raging Cow and the failed attempt to create a good reputation around it. The campaign was not well-received by the blogosphere community and bloggers attacked and boycotted the marketing campaign, making the product disappear from the market.

On the other hand, the next example is a mistake that could have damaged the image of Red Cross, but turned out to be a successful blood-donation campaign. In 2011, an employee from Red Cross made a tweet about drinking beer with an uncommon hashtag (#gettingslizzerd) from the company's account in the middle of the night. This event was noticed on Twitter getting attention from the users. In order to reverse this situation, Red Cross acknowledged the mistake and took action with humor [10]. Both companies Red Cross and Dogfish Head Brewery took advantage of the trending hashtag to their own benefit.

The given examples were triggered without intention and became a discussion topic across multiple platforms. Therefore, knowing the social media environment can be a very powerful tool to avoid harm or to improve social media metrics [9, 10].

A trending topic is defined by Twitter as an emerging discussion topic that is popular in the present moment. To be considered trending, a topic, needs to be discussed more then what it usually is. The authors [8] defined trending topics as the official Twitter description of 2010 "the hottest emerging topics (or the "most breaking" breaking news), rather than the most popular ones".

Additionally, as the authors [43] refereed in their work, in the year of 2011, trending topics became interesting to users, journalists, applications developers and social media researchers. Besides being new and relevant to people at that moment, the active time of a trending topic is limited [8]. Therefore, if it takes too long to evaluate trending topics, companies may not have time to do something effective [25].

Trending topics are relevant to companies' marketing as the authors [8] said "(...)Trending Topics present a comparable visibility to other traditional advertisement channels and thus they can be considered a useful tool in marketing and advertisement contexts.". Instead of companies spending more time and money to increase the visibility of their products or brand, they might take advantage of the already spoken topics on social media to reach their marketing goal.

There are companies already taking advantage of current events to communicate. On the 10 of June, Control a Portuguese brand, made a post on Instagram with the quote "it is day to raise the flag" (Fig. 1(a)) to take advantage of the national holiday of Portugal.

On February 8, 2019, in the 21st round of the Portuguese first league, a match between *FC Porto* and *Vitória de Guimarães*, Marega, FC Porto player, was the target of racist chants and shouts by supporters of the *Vitória de Guimarães* team. This topic was very discussed in social and traditional media. Nine days after that event, Super Bock and Sagres, two Portuguese competitor beer brands, made a post together on Facebook with the quote "Against racism, there are no rivals"(Fig. 1(b)).

(a) Control Portugal post on Instagram adapted from [29]

(b) Super Bock and Sagres post on Facebook adapted from [30]

Fig. 1 Control Portugal post on Instagram and Super Bock and Sagres post on Facebook

Finally, having an account on a social media platform is free and trending topics are easily available to everyone. Therefore, even organizations that may not have the budget to invest on specialized human resources and software to make social media marketing can take advantage of trending topics. The key is to analyse which trending topics are relevant to each company in a timely manner.

The goal of this work is to design and implement a prototype cable of use text mining techniques and train personalized models to find associations between trending topics and a social media account of an organization. Three different approaches are used a text similarity, a probabilistic and a classification task approach.

The prototype is called HotRivers and needs to be capable of:

1. **Collecting data from a social media platform:** By using the name of the company social media account (e.g. Adidas, Nike, Pull & Bear, and others) and desired location of the trending topic (e.g. United Kingdom, Lisbon, New York, and more);
2. **Preparing data:** Apply pre-processing techniques to clean and transform the data;
3. **Modeling:** Use different approaches to measure the similarity between data from trending topics and companies social media accounts.

It is important to refer that it is not the scope of this work to study the relationship between trending topics in multiple social media platforms, because each social media platform has different features, has distinct communities and can be classified into many categories [13, 15]. In Section 2.2 is explain which and why social media platform was picked.

Finally, the structure of this work is the following, in Section 1 were presented the problem, the motivations and the goals of this work. In Section 2 is defined what is social media,

the impacts of marketing and the presence of companies in social media platforms. Additionally, which social media should be used for this work and why. Furthermore, a research was conducted on works done with trending topics and similar prototypes and systems. In Section 3 is introduced the architecture of HotRivers and user requirements. In Section 4, three experiments are conducted. The first intends to demonstrate the result of all phases and to discuss what went wrong and which aspects need more testing. The second experiment continues testing the techniques and models that were not abandoned in the first experiment. The third experiments confirm that the chosen model is suitable for HotRivers and the objectives are reachable. Finally, in Section 6 is presented the conclusions, future work and limitations of the current work.

2 Literature review

In this section is discussed what is a social media, the impacts of marketing on social media and why it is beneficial to companies. Also, a systematic review was conducted on similar works to HotRivers and on related works on trending topics.

2.1 Social media and marketing

In 2010, [15] defined social media as "a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of User Generated Content.". According to [7], in the year of 2015, "Social media are Internet-based channels that allow users to opportunistically interact and selectively self-present, either in real-time or asynchronously, with both broad and narrow audiences who derive value from user-generated content and the perception of interaction with others.".

According to Tiago and Veríssimo [38], in 2014, marketers disseminated information related to the company or it's products through e-mail blasts, direct marketing, telemarketing, informational websites, television, radio, and others. Hence, if costumers are on social media then firms should be as well. Also, with social media evolution and growth, new challenges have appeared to improve social media services and user experience [16].

In the year of 2014, an online survey was conducted on the managers of the largest companies in Portugal. The authors concluded that 87% of the managers agrees that digital presence improves information gathering and feedback. Also, 85% of them acknowledged that digital presence increases knowledge and 82% admitted that digital presence promotes internal and external relationships [38].

In other study made in 2013, the authors [1], made a case study where they conducted interviews with the staff of running events. They concluded that the gain of using social media in relationship marketing in sport was getting higher acknowledgement from consumers, improved communication client-organization, better customer engagement and more efficient use of resources.

2.2 Why Twitter?

A set of conditions were selected to choose which social media should be used in this work. One condition was that the core of publications on that social media platform was text, for example, messages, posts, micro-blogging publications. Instead of video, photography or image, which are not text-based. Even though chat messages or messaging services are text-based, they are not suitable for this work.

For this work, another condition is that legal entities have visibility in that social media application. Visibility, in this context, is not paid publicity, sponsorship or partnership, but an user account that represents a legal entity.

The language used must be English and it was required to be a worldwide application and not specific to some part of the globe. English was the language chosen to be used in this work, because many sophisticated linguistic models were developed for English.

Additionally, this work uses the topics of the day, so it is important to choose a social media platform where the hottest topics are already filtered, because that is the focus of this work. Last but not least, the appliance to access the data must be easy and fast.

Table 1 was constructed in order to compare the top fifteen most used social media platforms and to see which conditions each one check [14]. As said before, messaging service providers such as Whatsapp, Facebook Messenger, WeChat and QQ were disregarded. Then, there are social networks based on videos or images like Youtube, Instagram, TIKTOK, Kuaishou, Snapchat and Pinterest which were also disregarded. While Sina Weibo had a few companies represented there, in QZone was not clear if legal entities played a relevant role. However, both platforms were made for Mandarin speakers and were therefore disregarded. Reddit was excluded, because it is a social news media aggregation, and no legal entity has presence on it [19]. The most proper candidates were Facebook and Twitter. Facebook was excluded too, due to the difficulty to access data and the hottest topic aggregator, because it is not clear how it works. Twitter seemed to be the most accessible of all and fulfilled all requirements.

Table 1 Table with which characteristic each social media has

Top 15 most used social media platforms						
Social media platforms	Text based	Legal entities visibility	Available in english	World wide platform	Appliance conditions difficulty[3]	Hottest topics filter
Facebook	Yes	Yes	Yes	Yes	Hard	Yes
Youtube	No	–	–	–	–	–
Whatsapp	Yes	N.A.[2]	–	–	–	–
Facebook Messenger	Yes	N.A.[2]	–	–	–	–
WeChat	Yes	N.A.[2]	–	–	–	–
Instagram	No	–	–	–	–	–
TIKTOK	No	–	–	–	–	–
QQ	Yes	N.A.[2]	–	–	–	–
QZone	Yes	N.K.[1]	–	–	–	–
Sina Weibo	Yes	Yes	No	–	–	–
Reddit	Yes	No	–	–	–	–
Kuaishou	No	–	–	–	–	–
Snapchat	No	–	–	–	–	–
Twiiter	Yes	Yes	Yes	Yes	Easy	Yes
Pinterest	No	–	–	–	–	–

[1] N.K.: Not known

[2] N.A.: Not applicable

[3] Difficulty: Easy, medium, hard

Table 2 Filtering criteria for associations between trending topics and companies systematic review

Filtering Criteria	
Inclusion criteria	Exclusion criteria
Written exclusively in English	Not written exclusively in English
Work developed to English language	Work developed to other languages
Publication after 2010	Publication before 2010
Free or inside ISCTE's scientific license	Paid documents
Papers in conferences or journals	Paper or journals published in non-trust sources
Title, abstract or keywords related to association between trending topics and companies	Non-applicability to association between trending topics and companies

2.3 Systematic review of related works

In order to search for similar prototypes, systems and works, the main research database used to conduct a systematic review was Scopus. As a second source for papers it was used Google Scholar, which needs more selective search query due to the great amount of works. The query used was the following:

("trending topics" OR "trend topics" OR "hot topics") AND (association OR relationship OR relation OR connection OR link) AND (companies OR firms OR corporation OR institution OR organization)

The number of documents was enormous. In order to distinguish the relevant from the non-relevant, it was used the filtering conditions described in Table 2. On the left side of the table are the acceptance criteria and on the right side the elimination criteria.

The search for the selected query gave 7,729 results as present in Table 3. Scopus filtering features were used to search for title, abstract and keywords related to the query used which returned 330 documents. The most common words were "hot topic", while fewer instances of "trending topics" and "trends topics" appeared. Those documents were individually analysed and applied the filtering criteria, only five works remained to be fully analysed. The five studies were about spam, exploratory analysis and event classification. Unfortunately, it was not possible to identify similar studies to this work.

As no related studies were found, the next search is about works on trending topics, including the three topics identified previously. It was given preference to those studies that used Twitter, Facebook or other social media trending topics as data. Even though there are many techniques to extract topics, the focus is on studies that used trending topics given by social media algorithms.

Table 3 Filtering steps of systematic review related to association between trending topics and company

Filtering steps	Number of works
Search for query	7,729
Title, abstract or keywords related to association between trending topics to companies	330
Applied inclusion and exclusion criteria	5
Full-work analyse	0

2.4 Related works to trending topics

In a total of 9,624 works filtered in this systematic review, the title, abstract and keywords of 632 studies were examined. Of those 632, 80 were picked and after that 21 were selected to be fully analyzed. One of the conclusions of the systematic review is that none of the works found had the goal of this work.

Some authors present trending topics classification as away to assign trending topics into different categories such as sport, politics, technology, and more [20, 42]. However, trending topics classification can also be in the sphere of event classification (e.g. festival and commemorative days, news, memes, and more) [43]. The Support Vector Machines or Multinomial Naïve Bayes are suitable for this type of classifications task achieving an accuracy greater than 70% in some cases.

Although, there are other authors that tried to classify the sentiments of trending topics such as neutral, happy, sad and more [21, 33]. This type of classification tasks can be more challenge and require more complex models such as Convolutional Neural Network (CNN) [17] or Long Short-Term Memory, which may attain up to 98% of F1-score in some scenarios [21]. In Table 4 is observed the year of the analysed studies, the approach used and what were the major findings.

Another studies focus on finding sub-topics and events on trending topics or to summarized the information on those trending topics. Finding and summarized topicsby using only text can be done through probabilistic models such as latent Dirichet allocation (LDA) model [2, 6]. In order to avoid topic modeling techniques and keywords frequency models, which have disadvantages to detect events on Twitter, an alternative could be through phrase network models [22]. However, another possibility to handle this problem could be through word embedding or document embedding models [35]. Another authors, instead of only using text, they tried to use sentimental features to detect trending topics [27].

A few authors used trending topics detection as a part of their framework to generate summarization of documents [6, 27]. There is one work that were able to produce a summary of textual and visual trending topics [6]. Additionally, Doc2Vec model was capable of returning high-quality results, steaming lower the performance and tweets aggregations

Table 4 Trending topic classification approaches and findings

Authors	Year	Approach used	Findings
Zubiaga et al. [43]	2011	SVM with 15 different features	Classified current events with an accuracy of 82.9% memes with 73.1%
Lee et al. [20]	2011	- MNB with bag-of-words TF-IDF - C5.0 decision tree learner	MNB accomplished 70% and C5.0 decision tree 65% accuracy
Zhu [42]	2018	MNB with short text aggregation	Model achieved 73.33% of accuracy and build and classifies in 1.5 seconds
Shalini et al. [33]	2019	- Bag of Tricks classifier - CNN - Bi-LSTM	The best were Bag of Tricks, then slightly worse CNN and last Bi-LSTM
Liu et al. [21]	2019	CNN-LSTM (A mix of a CNN and a LSTM)	Classification binary of offensive tweets attained 98% of F1-score and 67.9% of F1-score on classified sentiments

Table 5 Approaches and findings on trending topics detection and summarization works

Authors	Year	Approach used	Findings
Bian et al. [6]	2013	Multimodal latent Dirichlet allocation	The framework output is textual and visual summaries of the trending topics
Aiello et al. [2]	2013	- LDA - Doc-p - GFeat-p - FPM - SFPM - BNgram	- The best topic recall: BNgram - The most complete topic description: SFTM and LDA - The most precise topic description: FPM - Steaming worsen the performance and tweets aggregation seems to improve topic recall - LDA is affected by noisy events
Peng et al. [27]	2015	SVM with unigram features	- Use of sentimental features - The model achieved the highest response time and accomplished 73.3% of F1-score
Sharma et al. [34]	2015	Proposed the algorithm TopicDetect	Approach effective and extensive to cover important topics
Melvin et al. [22]	2017	Phrase network model	The model accomplishes an F1-score of 54%
Singh and Shashi [35]	2019	- Bag-of-words with TF-IDF - Word2Vec - Doc2Vec - k-means	- Bag-of-words with TF-IDF had a purity score of 0.98, Doc2Vec of 0.95 and Word2Vec of 0.89 - Bag-of-words with TF-IDF had slighly better performance, but offered less options than Word2Vec and Doc2Vec - Doc2Vec delivered the highest-quality results

have a tendency to improve and noisy events affects on LDA [35]. In Table 5 is described the major findings and approaches used by the authors.

The exploratory analysis of trending topics are relevant, because the authors study the behavior of the trending topics, in other words, discover what drives a trending topic, how and why they become trending, what are the key features, and many other questions.It was found that Twitter follows identical pattern to media news [40], and the most important attribute is the retweet by other users and the most content shared is news from traditional media [5]. Another study found that a trending topic emerge 1.5 times in a year [4], which means that are always new topics being discussed. Also, trending topics are driven by a log-normal distribution and have a decay of a geometric distribution [5]. That seen in accordance with the time that a topics became a trending topic, which is approximately 36.2 minutes to get to top ten and 91.5 minutes to be at top one [4]. Also, in one year, 44% of trending topic on Twitter live only one day and 24% two days [3]. The Authors of the exploratory analysis works and the more relevant findings are observed in Table 6.

Another authors studied the relationship of trending topics among different social media platforms. Trending topics on Twitter and Wikipedia are faster to appear and to disappear than on Google [3]. While, Google is more specialized in sports, celebrities, entertainment and politics, Twitter is more specialized in sports, celebrities, entertainment, products and holidays, and Wikipedia is more specialized in celebrities, entertainment and incidents [3]. The capability of Twitter to predict and lead to a Google Trend arise was accomplish with a Distributed Lag model. That model could explain 80% of the variance when use both Twitter and Google information. Additionally, 43% of the times, Twitter trending topics caused an

Table 6 Finding of Trending topics exploratory analyses works

Authors	Year	Findings
Asur et al. [5]	2011	- Trending topics are driven by a log-normal distribution - Trending topics have a decay of a geometric distribution - The most important attribute is the retweet by other users - The number of followers and tweet-rate of users does not provoke trends - The most content shared is news from traditional media
Wilkinson and Thelwall [40]	2012	Twitter follows identical pattern to media news
Annamoradnejad and Habibi [4]	2019	- Half of the trending topics were a single word - On average the trending topics had 30 characters and 2 words - Approximately a trending topic needed 36.2 minutes to get to top 10 and 91.5 minutes to be at top 1 - 977 trending topics in 1 year got to rank 1 in less than 10 minutes - A trending topic emerge 1.5 times - The longer duration of a trending topic was 30 hours

identical Google Trend. Giummolè et al. [12]. The goal of the studies found in Table 7 is forecasting the behavior of trending topics, in which are used more than one social media platforms.

3 HotRivers prototype overview

As discussed in Section 1, this prototype was built to find an association between the social media account of an organization and trending topics. It was pointed that most of the trending topics have a life span of 24h to 48h, but they can survive longer [3]. It is important

Table 7 Approaches used for forecasting trending topics behavior and finding on works that studied more than one social media platform trending topics

Authors	Year	Approach used	Findings
Althoff et al. [3]	2013	- Nearest Neighnor - Forecaste approach	The duration of the lifetime of a trending topic: - on Twitter was 1 day in 44% and 2 days in 24% of the trending topics - on Wikipedia, 50% of trend stayed 1 day and 16% 2 days - on Google, 17% lived seven days, 14% four days and 13% six days - The model forecast 9,000-48,000 views up front to 14 days with an error of 19-45%
Giummolè et al. [12]	2013	- TBG - Autoregressive model - DL model - Autoregressive - Distributed Lag model	- The DL model explained approximately 75% of the variance - When Google was the dependent variable and Twitter the explanatory variable the DL model was significant 60% the time - Twitter trending topics caused an identical Google Trend 43% of the times

to clarify that only the primarily top 10 trending topics were used. That is the window of opportunity to act.

Additionally, for this work, retweets were not used. When someone makes a tweet using an hashtag related to a subject is almost guaranteed that the user is talking about that subject. Even though that is not always true as showed in Fig. 3. However, retweets are answers to tweets and are more challenging to guarantee that the discussion is still about the subject related to the hashtag.

The HotRivers architecture was designed to be capable of collecting data from Twitter, then able to perform clean and transform processes and finally train models. The division of the architecture was made in three parts based on CRISP-DM methodology [41], as it is observed in the Fig. 2.

Phase 1 is data collection where HotRivers collects data from Twitter. There are two components, one collects tweets from specific accounts related to a company and the other collects tweets from trending topics. Even though they seem similar, they are two distinct API methods and have separate collecting approaches.

Then, Phase 2 is data preparation, where the prototype cleans and transforms tweets into more suitable text objects for the next phase.

Phase 3 is modeling. In this phase the models are trained using the data collected and prepared for each model. The modeling phase has three components, one for each approach.

The architecture has two inputs. The input one is the company's social media account name. Input two is the name of the desire location if available. It is possible to keep both collecting processes running independently. The prototype was designed in a way to improve performance and to mitigate the impact of API rate limits. Although separating the inputs was a necessity imposed by Twitter Application Programming Interface (API) architecture, it also, allows a better organization, gives more structure to the development work, and grants independence of the processes.

3.1 Data collection phase

Data understanding was important to explore and verify the quality of data. Phases 1 and 2 benefit from the information gathered during this exploration because many of the filters and clean processes were identified on that moment. Word counts distribution was analysed,

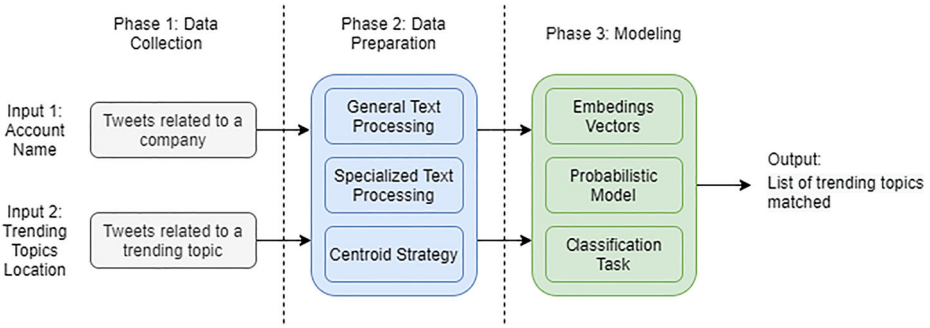

Fig. 2 HotRivers has three phases based on CRISP-DM methodology [41], which are Data Collection, Data Preparation and Modelling, and each phase has many modules. While on Phase 2 every module can be used, on Phase 3 only one module can be used. As an example, one possible HotRivers architecture combination could be Phase 1 modules, Phase 2 General Text Processing module and Phase 3 Embedding vectors module

as well as, tweets and words length, smiles, punctuation, content and style. This exploration was done in order to have an understating of the data collected.

There are two components in Phase 1, as mention before they are two distinct API methods, so they have separate collecting approaches. In order to make Phase 2 and data understating easier, both methods filtered the JSON fields needed into dictionaries to be analysed and saved. Data Collection was run every day to collect trending topics.

3.2 Data preparation phase

The general text processing (GTP) is composed by a set of methods to delete emojis, hashtags, usernames, other symbols and noise such as paragraphs (\n), Unicodes, retweet symbol, and many others, numbers, punctuation and remove stop words. Also, a method to lowercase and to tokenization. Additionally, three techniques were tested steaming, lemmatization and with no treatment. For this work, it was used mainly *nltk* package.

The specialized text processing (STP) is more focus in specific task such as produce statistical information such as count the number of times each document is in the corpus, count the number of words and the length of the documents in the corpus. Additionally, STP removes small words given length, removes small document given number of terms and duplicated document removal.

The Centroid Strategy (CS) is one of the three components of Data Preparation Phase. The trending topics are a set of tweets grouped by Twitter, in other words, trending topics could be seen as clusters of tweets. A centroid is the central point of a cluster, i.e. the tweet that best represents the trending topic [11]. We found that some trending topics tweets do not say explicitly anything about the discussion topic as shown in Fig. 3. The CS was created to filter those unintended tweets, as well as, to help to reduce the amount of tweets to extract, leading to significant less periods of waiting time.

The first step of the CS is average all word embeddings to create a vector to each document. This also could be made by the sum or concatenate the word embeddings. The next step is to calculate the cosine distance of all documents versus all documents. The output is a matrix of distances. Average is applied to each column to find which column have higher values. For this work GloVe [28] Twitter embeddings were used, but Fasttext [24] were also a possibility to use.

This method is sensitive to duplicate documents. Because if the centroid strategy finds high similarity with a set of duplicate documents, it might put them at the top, which may not represent that set of tweets.

Fig. 3 A tweet selected from the trending topic #SackWhitty

3.3 Modeling phase

The idea of document embedding vectors approach [35] is to use the model Doc2Vec [18] from Gensim [31], in other words, transform documents into embedding vectors.

The model was trained with a vector size of 200, minimum word frequency was 5, the number of epochs was 250, the algorithm was distributed bag-of-words with train word-vectors simultaneous with DBOW doc-vector training, an alpha of 0.025 and the default minimum alpha. In the sanity test of the model, the vectors were inferred with a configuration of alpha 0.025, default minimum alpha, and 275 epochs. The trending topics vectors were inferred with a configuration of alpha 0.025, default minimum alpha and 500 epochs.

The probabilistic model approach [2] was implemented by first converting the corpus into a dictionary, then transforming the corpus into a bag-of-word representation, then words that appear in less than five documents and more than 75% of the documents were deleted and finally the model was trained. The package used was Gensim [31].For LDA it was used the Hellinger distance, because quantifies the similarity between probability distributions.

The LDA model was trained with 100 topics, a chunk size of 2000, the number of passes was 20, the value of iterations was 400, the alpha and eta values were set as auto and the default minimum probability.

The classification task approach used a CNN model [21, 33]. The architecture used was proposed by Yoon Kim [17]. The CNN is made by an embedding layer and three convolutional layers. Then, it is applied max pooling on each convolution layer output and concatenate the layer. Finally, a dropout layer and softmax output. CNN is trained with the companies' tweets. This is a problem of binary classification. One of the labels is the target company and the other label is called Others, it is composed by tweets from different companies. The output of each trending topic is a value between 0 and 1.

For this CNN the embedding layer used was the GloVe embeddings [28] and set with 200 dimensions. The data was split into 75% for training and 25% for testing. The maximum number of each word based on frequency was set to 15,000. The model was set with the number of filters 100, document maximum length 27, number of classes 2, batch size of 64, number of epochs of 350, and a dropout of 0,30. The activation function was softmax, the loss was binary cross-entropy and the optimizer was adam.

4 Experiments and results

In this section, HotRivers was tested phase by phase. The companies used for the test were chosen based on the number of tweets published, the number of followers and area of business (e.g. health care, sport and fashion, technology and more). A pair of companies were selected to analyze the similarity of the results, i.e, companies that work in the same area of business. At the end of this section it was decided which treatment from the data preparation phase to use and which model is more suitable to integrate HotRivers.

4.1 HotRivers minimum operating requirements to work

This architecture has some requirements to work properly. These requirements were constructed during implementation and after testing the phases. They were created to insure quality in results. In Table 8 are summarized the minimum operating requirements to HotRivers work.

Table 8 HotRivers minimum operating requirements to work and affected phases, i.e., before initiate Collecting and Modeling phase those requirements must be checked

Affected phase	HotRivers requirements
Collecting phase	Own a Twitter developer account
	Minimum of 1,500 posted tweets in an account[1]
	Target account cannot be private
	Minimum of 100 tweets per topic
	Only hot topics' tweet from official English speakers countries
	Maximum extraction time 4 hours
	Only words, abbreviations and different word spelling
Modeling Phase	Number of tweets repeated between 2 to 5[2]
	Minimal of 1,000 tweets after cleaning and transformation

While some of them are verified by HotRivers, such as minimum of tweets, others are user responsibility, e.g. a minimum of 1,500 posted tweets in an account

[1]It is recommend accounts with at least 3,200 tweets posted

[2]Only when the CS is used

The first condition is to have a developer account to have access to Twitter API. Since the scope of this work is to analyze the trending topics of the day, the free account is enough to extract the newest tweets.

The second requirement is the minimal number of tweets for the modeling phase, which is 1,000 tweets. This value was set to guarantee that after cleaning and transforming the data it is possible to have enough tweets to properly train the models.

This leads to the third requisite, which is to choose only accounts that have more than 1,500 tweets published. It is recommended to choose an account with near 3,200 tweets posted.

The fourth requirement was found during extraction. The accounts cannot be private, otherwise, it is impossible to collect the data.

The fifth requirement is to have only words, abbreviations e.g. *"I love u"* (I love you) and different word spelling e.g. *"It's a raccooooooooooon!!!"* (It's a raccoon) after cleaning and transforming the data.

The sixth requirement is to have between two and five repetitions of each repeated tweet. This requirement is only mandatory when the CS is used.

The seventh requirement is to extract only trend topic tweets from official English speaking countries. The countries are listed on the website of the University of Northampton and The University of Sheffield [37, 39].

The eighth requirement is to collect at least 100 tweets of each topic because Twitter already categorized those tweets as topics. It is possible to extract more, but it is not advised to collect more than 250.

The ninth requirement is to limit the maximum time of extraction of trending topics to 4 hours. Since time is a crucial feature for this work, the extracting phase should not take longer than half a working day.

4.2 Data

The data used to test the prototype were only tweets from companies and trending topics. The countries collected were Australia, Canada, Ireland, New Zealand, the United Kingdom (UK) and the United States. Only the top ten trending topics were collected for each country.

In total, 28 days between 31st of August and 9st of November were collected, in a total of 1,680 trending topics and approximately 450,000 tweets. A sample were used, ten random days were picked from the UK.

The companies used were Adidas, Nike, and Portsmouth Hospitals University (PHU). Those accounts were selected, because of the distinct areas of business and number of followers on their social media account. In Table 9 are observed the number of tweets collected without any kind of treatment, and the number of tweets published and followers of each company. Adidas and Nike are both in sport, fashion, tennis, clothes areas and both seem to be competitors in the same business market. PHU is a hospital, representing the medical area. PHU have a smaller account comparing with Adidas and Nike.

The second reason to pick those companies was due to trending topics selected for testing. There were a few trending topics that stood out, which were *#PepsiMaxTasteOneStop, Nike, Amazon UK, argos, #XboxSeriesX, #Covid_19, #NHSCOVID19app* and *iOS 13*. Those were trending topics that were immediately associate to a company, except for *#Covid_19* and *#NHSCOVID19app* that were related to the current pandemic. So, the company Nike were picked, because of the trending topic *Nike*. Also, this trending topic *Nike* serve to evaluate the model performance. It is expected that Nike have a high similarity with the company Nike. Additionally, the trending topic *#Covid_19, #NHSCOVID19app* and *#WorldPatientSafetyDay* served the same purpose as trending topic *Nike*, but for the company PHU. It is expect that PHU has high similarity with those trending topics.

4.3 Experiment 1: Adidas

The phase one and two of HotRivers prototype were successfully implemented and tested. Data Collection were able to collect the trending topics of the day in a round forty-five minutes. Relating to Data Preparation all modules were tested and worked as indented. It was found that to use the STP the better configuration was documents bigger than three words, maximum of two document repetitions and words bigger than two characters.

The first model to test was Doc2Vec. A sanity test was run to check if the algorithm is learning correctly. It utilizes the same data that was used to train the algorithm. That data is given to the algorithm as new unseen data.

The results were not satisfactory. The sanity test of Adidas' tweets is observed in Table 10. The average and median values are close to each other, which means that the data is distributed around the average. The metric Rank explain how tweets are most similar to them self than with others. Rank 0 means that a tweet is the most similar to itself, Rank 1 means that there is one tweet that is more similar than himself and so one. So, with STP and using no treatment 92.1% of the tweets are similar to themselves, with lemmatization is 93.7% and with steaming is around 93.2%. That indicates that only about 8% of the tweets

Table 9 Total of tweets extracted, tweets published and number of followers of each company selected

Name of the company	Number of tweets collected	Number of tweets published	Number of followers
Adidas	3,104	13,800	3,800,000
Nike	2,889	36,800	8,200,000
Portsmouth Hospitals University	2,261	17,000	8,790

Table 10 Results of Doc2Vec sanity test trained with Adidas tweets with the GTP and STP techniques

Techniques		Average	Median	Rank	Number of Tweets
No STP	No treatment	0.357	0.360	[0: 1875, 1: 348, 2: 181, 3: 108, 4: 75, ...]	3,081
	Lemmatization	0.358	0.359	[0: 1928, 1: 355, 2: 168, 3: 104, 4: 77, ...]	
	Steamming	0.359	0.361	[0: 1923, 1: 367, 2: 180, 3: 105, 4: 64, ...]	
With STP	No treatment	0.356	0.354	[0: 1082, 1: 69, 2: 14, 3: 2, 831: 1, ...]	1,175
	Lemmatization	0.358	0.349	[0: 1078, 1: 59, 2: 6, 3: 4, 4: 1, ...]	1,151
	Steamming	0.357	0.349	[0: 1082, 1: 61, 2: 13, 3: 1, 4: 2, ...]	1,161

The STP configurations was the same present before, documents bigger than three words, maximum of two repetitions per document and words bigger than two characters

are more similar to other tweets. Even though, the ranks are better with STP, the average of the cosine distance is close to zero, which means that there are no associations.

Relating the model LDA the results of the experiment with GTP were not satisfactory. Even though the number of documents was a significant figure, this did not turn into a sizable number of unique words. In Table 11 it is observed that the number of unique tokens is not higher than 513, which is a low number. Additionally, the Hellinger distance indicates low similarity.

The CNN results with the training dataset were satisfactory. The model was tested with GTP techniques and STP. In Table 12, is observed that the average and F1-score present values higher than 90%. The STP results were always worse than without the STP. It seems that CNN learns better with more data, and duplication did not seem to have a negative impact. Lemmatization had in both techniques, with and without STP, better results than the rest of the techniques. The CS was tested, but it did not improve the model.

4.4 Adidas results with trending topics and analyse

The results of using GTP with lemmatization were consistently better than steaming and no treatment with CNN. Relating to STP technique and the CS with the CNN model, it did not help to increase the model learning and did not boost the performance.

The UK results on trending topics of days one, twenty, twenty-two, and twenty-four of September are described in Table 17 (see Appendix A). Only two out of eighty topics were selected by the HotRivers prototype. The selected topic with the highest similarity was *Ed Sheeran* and the lowest similarity was *#WorldPatientSafetyDay*. The majority of the topics are classified as Others. The disparity of values and the number of trending topics selected is satisfactory. The number of trending topics labeled as Others does not matter to Adidas. It indicates that those topics do not have any associations with the company.

Table 11 Results of LDA sanity test

Technique	Number of unique tokens	Number of documents	Average	Standard deviation similarity	Median similarity
No treatment	513	3,082	0.800	0.092	0.816
Lemmatization	502		0.814	0.089	0.831
Steaming	507		0.811	0.087	0.820

Table 12 Performance metrics results of the CNN model trained with GTP, STP techniques on Adidas data

Techniques		Loss	Accuracy	Recall	Precision	F-1 score	Quantity of Adidas tweets
No STP	No treatment	0.421	0.918	0.920	0.920	0.920	3,077
	Lemmatization	0.466	0.920	0.922	0.922	0.922	
	Steaming	0.543	0.910	0.912	0.912	0.912	
With STP	No treatment	0.479	0.901	0.903	0.903	0.903	2,886
	Lemmatization	0.447	0.903	0.903	0.903	0.903	2,874
	Steaming	0.543	0.900	0.901	0.901	0.901	2,886

The trending topic *Ed Sheeran* refers to the singer Ed Sheeran, who, like other celebrities wear Adidas products, this fact might explain the high average. The model was not clear of which label for the trending topic *Nike*. There is two points worth of stress, first was good that the prototype did not classify it as Adidas, because they are different companies, but on the other hand, having a higher average would not be a surprise, since both companies sell similar products and seem to be direct competitors. Additionally, the trending topic *#PepsiMaxTasteOneStopwere* was classified as Others. This happened with other companies and products such as *Amazon UK*, *argos*, Xbox, Playstation 5, and *iOS 13*. It is important to point out that Amazon, Pepsi, Microsoft, Apple, or Sony were not introduced to the training dataset. In other words, the model was enough well trained to label those topics as Others. The trending topics *#WorldPatientSafetyDay*, *#Covid_19* and *#NHSCOVID19app* were labeled as Others with a high average, due to medical and pharmaceutic companies used on the training dataset. In other words, hospitals and pharmacies made multiple publications about the Covid-19 subject. Which made the model learn that trending topics related to Covid-19 are not related to Adidas. Regarding the trending topic *Leighton Buzzard*, it was not found obvious connections.

4.5 Experiment 2: Nike

In this experiment the target company is Nike. The Data Collection phase occur without problems and extract a total of 2,889 tweets. Relating to Data Preparation phase, the STP techniques and the CS were not applied to companies tweets, and the GTP used was with lemmatization. For the Modeling phase only the classification task with the CNN model were used.

The results of training the model with Nike dataset is observed in Table 13. The accuracy and the F-1 score values are satisfactory, however slightly worse than in experiment 1.

4.6 Nike results with trending topics and analyse

Regarding the UK trending topics, presented in Table 18, the results were interesting, particularly to the trending topic *Nike*, which gave almost the maximum average. The lowest

Table 13 Performance metrics results of CNN model trained with Nike data

Loss	Accuracy	Recall	Precision	F-1 score	Quantity of Nike tweets
0.535	0.906	0.906	0.906	0.906	2,880

trending topic was *#AutumnEquinox*. The model was able to recognize Nike very well. The color scheme is equal to the previous experiment, color green for the picked trending topics and color yellow for the standard deviation of the highest value of average. Two trending topics were selected out of eighty. The disparity of the values seems satisfactory and in two days there were not any associations with Nike.

The dataset for training the model was the same used in the last experiment, except for the targeted company. Again, the trending topic *#WorldPatientSafetyDay* were classified with one of the highest average values for the label Others. The trending topics *#Covid_19* and *#NHSCOVID19app* were also labeled as Others. Additionally, the trending topic *#AutumnEquinox* was classified as Others with a high average too. The main reason seems to be the data used in the training. Trending topics such as *iOS 13, #PepsiMaxTasteOneStop, argos*, and *Amazon UK* were also classified as Others and no association was found to the trending topic *Harold Evans*. The high standard deviation and the low average made a lot of topics being picked.

4.7 Experiment 3: Portsmouth Hospitals University

The target company in this experiment was the Portsmouth Hospitals University (PHU). Similar to previous experiments the Data Collection phase did not have issues and extract 2,261 tweets. The Data Preparation used GTP with lemmatization and the CS only for the tranding topics. The Modeling phase used the CNN model.

The same dataset from the last experiment was used on the training dataset, except for the target company the PHU. The results are observed in Table 14. The results of the F1-score and accuracy are satisfactory, however the results were slightly worse than Nike and Adidas.

4.8 Portsmouth Hospitals University Results with trending topics and analyse

Concerning the UK trending topics results showed, in Table 19, only day seventeen of September had associations with PHU. In a total of eighty trending topics, only three were picked. The trending topic with the lowest value of average was *Ed Sheeran* and the highest was *#WorldPatientSafetyDay*. The disparity of values is satisfactory and the chosen standard deviations were close to zero, which means that the model classified the trending topics with confidence.

The trending topic *#WorldPatientSafetyDay* is related to the hospital affairs, but *#SackWhitty* and *#northeastlockdown* are associate to Covid19 and political issues, apart of Covid-19 subject, the political matter may not be the interest of the hospital. On the opposite side classified as Others, the trending topic *Ed Sheeran* was correctly classified since in PHU training dataset Adidas was labeled as Others. Again, it is worth to point that other companies trending tropics *#PepsiMaxTasteOneStop* and *Nike* had no association with PHU. As a trending topic *argos*, which is a subsidiary of Sainsbury's, was another company labeled as Others on the training dataset. On day twenty-two of September, the trending topic *#Covid_19* did not have enough value of average, because *Leighton Buzzard*

Table 14 Performance metrics results of CNN model trained with PHU data

Loss	Accuracy	Recall	Precision	F1-score	Quantity of PHU tweets
0.638	0.872	0.872	0.872	0.872	2,259

had a high association with label Others. The lack of an association with *#Covid_19* and *#NHSCOVID19app* might be explained by many retweets about Covid-19 and not many tweets related to Covid-19, because only tweets were used, and another point contributing to that situation is that companies such as Nike, Adidas, Tesco and others, that were in training dataset label as Others, made publications about their Covid-19 policies and giving to support to everyone suffering from the pandemic.

5 Results discussion

Relating to the models tested in the three experiments, it was concluded that tweets made by companies to marketing purposes are written in a way that results in low number of unique words and short tweets, as illustrated in Table 11. Also, it was found that during Data Preparation on Adidas' tweets the most frequent length of tweets is between two and three words per tweet, contrary to journalistic texts, reviews or books [23, 34], which are extensive and detailed. This is one of the reasons to justify the mediocre performance of models such as Doc2Vec and LDA. Tweet aggregation methods could have helped [2, 42] the models learning, however the variety of words would not change with tweet aggregation. On the other hand, the model CNN was able to learn the data well and to classify the tweets assertively. A few points worth of stressing is that repetitions did not seem to have impact on learning. Also, it seems that when training the model with lower number of tweets the results are worse. This issue can be fixed by continually extracting tweets from target companies. Therefore, the CNN was the selected model for HotRivers prototype.

The final scheme of HotRivers, as observed in Fig. 4, consists in data collection phase as described in Section 3, data preparation phase with GTP with lemmatization and CS, only for trending topics, and modeling phase with the CNN. This scheme complies with all minimal operating requirements Section 4.1 and goals of this work.

Relating to the trending topics identified in Section 4.2, they were successful classified by the model. It is important to highlight that Sony, Apple, Microsoft, Amazon and Pepsi were not in the training dataset. In Table 15 are ten trending topics that were identified to have an association or not with the companies.

In Adidas case test, the comparison case of Nike case test, trending topic *Nike* were pinpoint as one of the trending topics that could deceive the model, due to the fact both Adidas and Nike companies contest the same market. Even though it gave a considerable

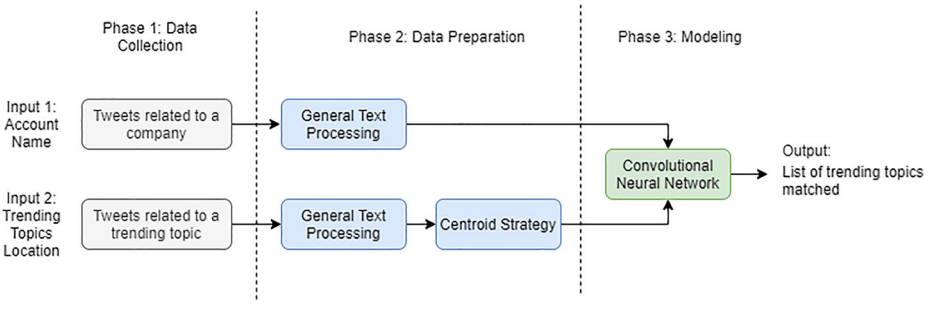

Fig. 4 Final Scheme of HotRivers, for company tweets GTP with lemmatization, and for trending topics GTP with lemmatization and the CS, and the picked model is a CNN

Table 15 Trending topics that may have an association with test case companies

Trending Topic Name	Average		
	Adidas	Nike	PHU
#PepsiMaxTasteOneStop	0.263	0.323	0.234
Nike	0.560	0.986	0.382
#WorldPatientSafetyDay	0.013	0.036	0.816
#ps5preoder	0.222	0.566	0.373
#northeastlockdown	0.202	0.351	0.722
Amazon Uk	0.326	0.400	0.369
argos	0.382	0.514	0.297
#XboxSeriesX	0.251	0.603	0.392
#Covid_19	0.121	0.237	0.689
#NHSCOVID19app	0.054	0.200	0.683
iOS 13	0.130	0.284	0.526

high average, it was not enough. For the rest of the trending topics the average value were very low, which were satisfactory results.

In Nike case test, the most notorious trending topic was *Nike*, because a low average value indicates that the model did not learn correctly the data. Fortunately, the model learned correctly the data and predicted with high confidence that trending topic. The rest of the trending topics had a low average value as expected, except for trending topics *#XboxSeriesX, argos,* and *#ps5preoder* that had a slightly higher average value comparing with Adidas and PHU.

Finally, in the last case test, the most important trending topics were *#WorldPatientSafetyDay, #northeastlockdown, #Covid_19* and *#NHSCOVID10app*. While *#WorldPatientSafetyDay* and *#northeastlockdown* were classify by the model with higher confidence, the trending topics *#Covid_19* and *#NHSCOVID10app* had slightly lower average. Relating to the rest of the trending topics, they had a low value of average as expected, except for trending topic *iOS 13* that has a slightly higher value.

Final appreciation is that the most critical trending topics were correctly classified and that is demonstrate that HotRivers is able to extract, clean, transform and learn the data as supposed.

6 Conclusion

One of the main goals of this work was to search for associations between companies and trending topics. The most evident result was the high association between the company Nike and the trending topic *Nike*, which had a value of 0.986, almost the maximum possible value. Another outcome worth of reference was that with the company Adidas, all trending topics selected had an association value higher than 0.826, which shows high confidence from the model. Also, between the hospital PHU and *#WorldPatientSafetyDay* and *#northeastlockdown*, there was high association, which was expected since both the company and trending topics are related to medical and health affairs. Unfortunately, the trending topics *#Covid_19*

Table 16 Summary of objectives, conclusions and contributions

Objectives	Conclusion	Contributions
Building a solution that is cable of use text mining techniques to process text data, train personalized models based on companies social media account and delivered a list of matched trending topics with the target company	- The CNN were the model that was better adapted of the models - The lemmatization were the techniques that more improved the models and the CS were capable of avoid unwanted tweets. - Results in less than one hour - Only requires two inputs from the user - Fully automated and the output is easy to understand	- A detailed explanation how to implement a prototype cable of finding associations between trending topics and companies - Three different approaches and pre-processing techniques were tested on three different companies - First work that tried to search for associations between companies and trending topics

and *#NHSCOVID19app* were not picked by the prototype as associated with PHU, however both trending topics had an average value of approximately 0.68. Even though, it was possible to conclude that the connection between companies and trending topics exists.

The final appreciation is that HotRivers can be a powerful tool to point out the direction of the marketing campaign. It was possible to observe that HotRivers prototype was capable of finding associations between companies and trending topics. Also, HotRivers is capable of deliver fast results, does not need specialized staff to operate and is automated. This qualities make HotRivers prototype suitable to be used in a larger and complex marketing platform. In Table 16 is illustrated a summary of the objectives, conclusions and contributions of this work.

6.1 Future work and limitations

For future work, the hashtags, mentions, and emojis need to be properly handled to be used to improve model learning, and the use of other feature than text should be considered such as timeline of the events, location, sentiments and others.

It would be interesting to use groups of companies to represent markets and to compare the results with personalized models for each company. This approach could handle some HotRivers requirement limitations. Additionally, measure how companies feel that our models can identify correctly useful trending topics for them.

One of the limitations of this work is the difficulty in confirming the associations between companies and trending topics, it was made by resorting the news and queries on google and by analyzing the content of tweets from some companies and trending topics. Even though it is possible to state there are reasonable associations, more work needs to be done.

Another limitation is that those associations in trending topics that apparently did not have a clear connection in these work may have when analysed by a marketing specialist or the employees of the targeted company. Also, trending topics with not a solid connection might still have potential for marketing purposes. That is why judging the model decision veracity is a complicated task and further studies with companies are required. Nevertheless, it presents interesting results.

Appendix A: HotRivers experiments results

Table 17 Table of Adidas CNN model of days one, seventeen, twenty-two and twenty-four of September UK trending topics results

Day	Trending Topic Name	Average - Adidas	Average - Others	Standard Devitation - Adidas	Standard Devitation - Others
01/09/2020	#September1st	0.212	0.788	0.159	0.159
	#BackToSchool	0.269	0.731	0.154	0.154
	Ritchie	0.367	0.633	0.194	0.194
	Marcus Rashford	0.506	0.494	0.248	0.248
	Ed Sheeran	0.838	0.162	0.049	0.049
	#TuesdayMorning	0.416	0.584	0.197	0.197
	#ThisMorning	0.402	0.598	0.180	0.180
	#PepsiMaxTasteOneStop	0.263	0.737	0.216	0.216
	Jim Davidson	0.629	0.371	0.209	0.209
	Nike	0.560	0.440	0.329	0.329
17/09/2020	Thiago	0.532	0.468	0.251	0.251
	Alex Scott	0.308	0.692	0.168	0.168
	#ThursdayThoughts	0.332	0.668	0.137	0.137
	#WorldPatientSafetyDay	0.013	0.987	0.020	0.020
	#SackWhitty	0.130	0.870	0.073	0.073
	#ps5preorder	0.222	0.778	0.101	0.101
	#northeastlockdown	0.202	0.798	0.163	0.163
	Amazon UK	0.326	0.674	0.183	0.183
	Neil Warnock	0.414	0.586	0.288	0.288
	argos	0.382	0.618	0.232	0.232
22/09/2020	#XboxSeriesX	0.251	0.749	0.156	0.156
	Starmer	0.214	0.786	0.153	0.153
	#TuesdayThoughts	0.298	0.702	0.214	0.214
	#AutumnEquinox	0.191	0.809	0.069	0.069
	Britain First	0.304	0.696	0.164	0.164
	Michael Gove	0.440	0.560	0.259	0.259
	#GBBO	0.265	0.735	0.104	0.104
	#Covid_19	0.121	0.879	0.166	0.166
	Leighton Buzzard	0.826	0.174	0.022	0.022
	Bake Off	0.345	0.655	0.202	0.202
24/09/2020	#NHSCOVID19app	0.054	0.946	0.065	0.065
	#ThursdayThoughts	0.323	0.677	0.205	0.205
	Harold Evans	0.673	0.327	0.296	0.296
	#Magic	0.334	0.666	0.161	0.161
	iOS 13	0.130	0.870	0.070	0.070
	Kent	0.207	0.793	0.193	0.193
	Gigi	0.677	0.323	0.185	0.185
	#thursdayvibes	0.408	0.592	0.193	0.193
	Downloaded	0.355	0.645	0.145	0.145
	#bbcbreakfast	0.228	0.772	0.115	0.115

Table 18 Table of Nike model results of days one, seventeen, twenty-two and twenty-four of September UK trending topics

Day	Trending Topic Name	Average - Nike	Average - Others	Standard Devitation - Nike	Standard Devitation - Others
01/09/2020	#September1st	0.220	0.780	0.199	0.199
	#BackToSchool	0.251	0.749	0.144	0.144
	Ritchie	0.525	0.475	0.135	0.135
	Marcus Rashford	0.500	0.500	0.163	0.163
	Ed Sheeran	0.352	0.648	0.134	0.134
	#TuesdayMorning	0.371	0.629	0.151	0.151
	#ThisMorning	0.362	0.638	0.186	0.186
	#PepsiMaxTasteOneStop	0.323	0.677	0.170	0.170
	Jim Davidson	0.435	0.565	0.204	0.204
	Nike	0.986	0.014	0.017	0.017
17/09/2020	Thiago	0.579	0.421	0.211	0.211
	Alex Scott	0.265	0.735	0.133	0.133
	#ThursdayThoughts	0.462	0.538	0.180	0.180
	#WorldPatientSafetyDay	0.036	0.964	0.024	0.024
	#SackWhitty	0.215	0.785	0.149	0.149
	#ps5preorder	0.566	0.434	0.173	0.173
	#northeastlockdown	0.351	0.649	0.191	0.191
	Amazon UK	0.400	0.600	0.236	0.236
	Neil Warnock	0.491	0.509	0.171	0.171
	argos	0.514	0.486	0.206	0.206
22/09/2020	#XboxSeriesX	0.603	0.397	0.210	0.210
	Starmer	0.268	0.732	0.202	0.202
	#TuesdayThoughts	0.345	0.655	0.234	0.234
	#AutumnEquinox	0.077	0.923	0.032	0.032
	Britain First	0.428	0.572	0.195	0.195
	Michael Gove	0.393	0.607	0.238	0.238
	#GBBO	0.229	0.771	0.132	0.132
	#Covid_19	0.237	0.763	0.201	0.201
	Leighton Buzzard	0.735	0.265	0.048	0.048
	Bake Off	0.286	0.714	0.189	0.189
24/09/2020	#NHSCOVID19app	0.200	0.800	0.191	0.191
	#ThursdayThoughts	0.459	0.541	0.230	0.230
	Harold Evans	0.616	0.384	0.229	0.229
	#Magic	0.303	0.697	0.191	0.191
	iOS 13	0.284	0.716	0.245	0.245
	Kent	0.328	0.672	0.231	0.231
	Gigi	0.378	0.622	0.077	0.077
	#thursdayvibes	0.510	0.490	0.200	0.200
	Downloaded	0.451	0.549	0.232	0.232
	#bbcbreakfast	0.457	0.543	0.258	0.258

Table 19 Table of PHU model results of days one, seventeen, twenty-two and twenty-four of September UK trending topics

Day	Trending Topic Name	Average - PHU	Average - Others	Standard Deviation - PHU	Standard Deviation - Others
01/09/2020	#September1st	0,486	0,514	0,152	0,152
	#BackToSchool	0.441	0.559	0.167	0.167
	Ritchie	0.457	0.543	0.184	0.184
	Marcus Rashford	0.506	0.494	0.238	0.238
	Ed Sheeran	0.123	0.877	0.020	0,020
	#TuesdayMorning	0.303	0.697	0.110	0.110
	#ThisMorning	0.375	0.625	0.129	0.129
	#PepsiMaxTasteOneStop	0.234	0.766	0.074	0.074
	Jim Davidson	0.201	0.799	0.105	0.105
	Nike	0.382	0.618	0.213	0.213
17/09/2020	Thiago	0.422	0.578	0.225	0.225
	Alex Scott	0.432	0.568	0.138	0.138
	#ThursdayThoughts	0.462	0.538	0.086	0.086
	#WorldPatientSafetyDay	0.816	0.184	0.168	0.168
	#SackWhitty	0.733	0.267	0.170	0.170
	#ps5preorder	0.373	0.627	0.143	0.143
	#northeastlockdown	0.722	0.278	0.139	0.139
	Amazon UK	0.369	0.631	0.169	0.169
	Neil Warnock	0.372	0.628	0.168	0.168
	argos	0.297	0.703	0.111	0.111
22/09/2020	#XboxSeriesX	0.392	0.608	0.099	0.099
	Starmer	0.610	0.390	0.189	0.189
	#TuesdayThoughts	0.458	0.542	0.132	0.132
	#AutumnEquinox	0.405	0.595	0.119	0.119
	Britain First	0.387	0.613	0.131	0.131
	Michael Gove	0.356	0.644	0.252	0.252
	#GBBO	0.484	0.516	0.154	0.154
	#Covid_19	0.689	0.311	0.193	0.193
	Leighton Buzzard	0.186	0.814	0.028	0.028
	Bake Off	0.481	0.519	0.193	0.193
24/09/2020	#NHSCOVID19app	0.683	0.317	0.154	0.154
	#ThursdayThoughts	0.390	0.610	0.173	0.173
	Harold Evans	0.180	0.820	0.102	0.102
	#Magic	0.401	0.599	0.191	0.191
	iOS 13	0.526	0.474	0.199	0.199
	Kent	0.523	0.477	0.148	0.148
	Gigi	0.137	0.863	0.036	0.036
	#thursdayvibes	0.402	0.598	0.152	0.152
	Downloaded	0.390	0.610	0.189	0.189
	#bbcbreakfast	0.513	0.487	0.157	0.157

Declarations

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

References

1. Abeza G, O'Reilly N, Reid I (2013) Relationship marketing and social media in sport. Int J Sport Commun 6(2):120–142. https://doi.org/https://doi.org/10.1123/ijsc.6.2.120, https://journals.humankinetics.com/view/journals/ijsc/6/2/article-p120.xml
2. Aiello LM, Petkos G, Martin C, Corney D, Papadopoulos S, Skraba R, Göker A, Kompatsiaris I, Jaimes A (2013) Sensing trending topics in twitter. IEEE Trans Multimed 15(6):1268–1282. https://doi.org/10.1109/TMM.2013.2265080
3. Althoff T, Borth D, Hees J, Dengel A (2013) Analysis and forecasting of trending topics in online media streams. In: Proceedings of the 21st ACM international conference on multimedia, pp 907–916
4. Annamoradnejad I, Habibi J (2019) A comprehensive analysis of twitter trending topics. In: 2019 5th International Conference on Web Research (ICWR), pp 22–27. https://doi.org/10.1109/ICWR.2019.8765252
5. Asur S, Huberman B, Szabó G, Wang C (2011) Trends in social media : persistence and decay. In: 5th international AAAI conference on weblogs and social media. https://doi.org/10.2139/ssrn.1755748
6. Bian J, Yang Y, Chua TS (2013) Multimedia summarization for trending topics in microblogs. In: Proceedings of the 22nd ACM international conference on information & knowledge management, association for computing machinery, New York, NY, USA, CIKM '13, p 1807–1812. https://doi.org/10.1145/2505515.2505652. https://doi.org/10.1145/2505515.2505652
7. Carr CT, Hayes RA (2015) Social media: defining, developing, and divining. Atl J Commun 23(1):46–65. https://doi.org/10.1080/15456870.2015.972282
8. Carrascosa JM, González R, Cuevas R, Azcorra A (2013) Are trending topics useful for marketing? visibility of trending topics vs traditional advertisement. In: Proceedings of the first ACM conference on Online social networks, pp 165–176
9. Deepa N, Deshmukh S (2013) Social media marketing: the next generation of business engagement. Int J Manag Res Rev 3(2):2461
10. Fan W, Gordon MD (2014) The power of social media analytics. Commun ACM 57(6):74–81
11. Giorgis S, Rousas A, Pavlopoulos J, Malakasiotis P, Androutsopoulos I (2016) aueb.twitter.sentiment at SemEval-2016 task 4: A weighted ensemble of SVMs for Twitter sentiment analysis. In: Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016), association for computational linguistics, San Diego, California, pp 96–99. https://doi.org/10.18653/v1/S16-1012, https://www.aclweb.org/anthology/S16-1012
12. Giummolè F, Orlando S, Tolomei G (2013) Trending topics on twitter improve the prediction of google hot queries. In: 2013 International Conference on Social Computing, pp 39–44. https://doi.org/10.1109/SocialCom.2013.12
13. Hoffman DL, Fodor M (2010) Can you measure the roi of your social media marketing? MIT Sloan Manag Rev 52(1):41
14. Hootsuite and We Are Social (2020) Global social media overview. Available at https://datareportal.com/social-media-users Accessed 21 Sept 2020
15. Kaplan AM, Haenlein M (2010) Users of the world, unite! The challenges and opportunities of social media. Bus Horiz 53(1):59–68. https://doi.org/10.1016/j.bushor.2009.09.003
16. Kim HN, Rawashdeh M, Alghamdi A, El Saddik A (2012) Folksonomy-based personalized search and ranking in social media services. Inf Syst 37(1):61–76. https://doi.org/10.1016/j.is.2011.07.002, https://www.sciencedirect.com/science/article/pii/S0306437911000858
17. Kim Y (2014) Convolutional neural networks for sentence classification. arXiv:1408.5882
18. Le QV, Mikolov T (2014) Distributed representations of sentences and documents. arXiv:1405.4053
19. Leavitt A, Robinson JJ (2017) The role of information visibility in network gatekeeping: information aggregation on reddit during crisis events. In: Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing, association for computing machinery, New York, CSCW '17, p 1246–1261. https://doi.org/10.1145/2998181.2998299

20. Lee K, Palsetia D, Narayanan R, Patwary MMA, Agrawal A, Choudhary A (2011) Twitter trending topic classification. In: 2011 IEEE 11th international conference on data mining workshops, pp 251–258. https://doi.org/10.1109/ICDMW.2011.171
21. Liu L, Huang X, Xu J, Song Y (2019) Oasis: online analytic system for incivility detection and sentiment classification. In: 2019 international conference on data mining workshops (ICDMW), pp 1098–1101. https://doi.org/10.1109/ICDMW.2019.00162
22. Melvin S, Yu W, Ju P, Young S, Wang W (2017) Event detection and summarization using phrase network. In: Altun Y, Das K, Mielikäinen T, Malerba D, Stefanowski J, Read J, Žitnik M, Ceci M, Džeroski S (eds) Machine learning and knowledge discovery in databases. Springer International Publishing, Cham, pp 89–101
23. Mikolov T, Chen K, Corrado GS, Dean J (2013) Efficient estimation of word representations in vector space. arXiv:1301.3781
24. Mikolov T, Grave E, Bojanowski P, Puhrsch C, Joulin A (2018) Advances in pre-training distributed word representations. In: Proceedings of the international conference on language resources and evaluation (LREC)
25. Mosley Jr RC (2012) Social media analytics: data mining applied to insurance twitter posts. In: Casualty actuarial society e-forum, vol 2. Citeseer, p 1
26. Ortiz-Ospina E (2019) The rise of social media. Available at https://ourworldindata.org/rise-of-social-media Accessed 28 May 2020
27. Peng B, Li J, Chen J, Han X, Xu R, Wong KF (2015) Trending sentiment-topic detection on twitter. In: Gelbukh A (ed) Computational linguistics and intelligent text processing. Springer International Publishing, Cham, pp 66–77
28. Pennington J, Socher R, Manning CD (2014) Glove: global vectors for word representation. In: Empirical methods in natural language processing (EMNLP), pp 1532–1543. http://www.aclweb.org/anthology/D14-1162 Accessed 8 June 2020
29. Portugal C (2020) Control portugal social media publication. Available at https://www.instagram.com/controlportugal/?hl=pt Accessed 03 July 2020
30. Publico (2020) Super bock e sagres aliam-se em campanha contra o racismo. Available at https://www.publico.pt/2020/02/17/fugas/noticia/super-bock-sagres-aliamse-campanha-racismo-1904547 Accessed 11 Sept 2020
31. Řehůřek R, Sojka P (2010) Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks, ELRA, Valletta, Malta, pp 45–50. http://is.muni.cz/publication/884893/en Accessed 6 June 2020
32. Roser M, Ritchie H, Ortiz-Ospina E (2015) Internet. Available at https://ourworldindata.org/internet Accessed 28 May 2020
33. Shalini K, Kumar MA, Soman K (2019) Deep-learning-based stance detection for indian social media text. In: Emerging research in electronics, computer science and technology, Springer, pp 57–67. https://doi.org/10.1007/978-981-13-5802-9_6
34. Sharma S, Aggarwal K, Papneja P, Singh S (2015) Extraction, summariz ation and sentiment analysis of trending topics on twitter. In: 2015 Eighth International Conference on Contemporary Computing (IC3), pp 295–301. https://doi.org/10.1109/IC3.2015.7346696
35. Singh AK, Shashi M (2019) Vectorization of text documents for identifying unifiable news articles. Int J Adv Comput Sci Appl 10(7):. https://doi.org/10.14569/IJACSA.2019.0100742
36. Smith S, Wu J, Murphy J (2020) Map: George Floyd protests around the world. Available at https://www.nbcnews.com/news/world/map-george-floyd-protests-countries-worldwide-n1228391 Accessed 18 June 2020
37. The University of Sheffield (2020) List of majority native english speaking countries. Available at https://www.sheffield.ac.uk/international/english-speaking-countries Accessed 29 Mar 2020
38. Tiago MTPMB, Veríssimo JMC (2014) Digital marketing and social media:, Why bother? Bus Horiz 57(6):703–708. https://doi.org/10.1016/j.bushor.2014.07.002
39. University of Northampton (2020) Majority native english speaking countries. Available at https://www.northampton.ac.uk/international/english-language-requirements/majority-native-english-speaking-countries/ Accessed 12 Feb 2020
40. Wilkinson D, Thelwall M (2012) Trending twitter topics in english: an international comparison. J Am Soc Inf Sci Technol 63(8):1631–1646. https://doi.org/10.1002/asi.22713. https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.22713
41. Wirth R (2000) Crisp-dm: towards a standard process model for data mining. In: Proceedings of the fourth international conference on the practical application of knowledge discovery and data mining, pp 29–39

42. Zhu Q (2018) Classification of trending topics in twitter. In: 2018 international conference on computational science and computational intelligence (CSCI), pp 274–277. https://doi.org/10.1109/CSCI46756.2018.00060
43. Zubiaga A, Spina D, Fresno V, Martínez R (2011) Classifying trending topics: a typology of conversation triggers on twitter. In: Proceedings of the 20th ACM international conference on information and knowledge management, association for computing machinery, New York, CIKM '11, pp 2461–2464. https://doi.org/10.1145/2063576.2063992

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

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.