المجلة الإلكترونية الشاملة متعددة التخصصات
العدد الثامن والخمسون شهر (٤) ٢٠٢٣

Extraction of Trending-Topics by using TF_IDF

***Roqaya Khalil Kashmola*[1*]** ***Ghayda Abdul Aziz Al-Talib*[2]**

[1,2] Department of Computer Science, College of Computer Science and Mathematics, University of Mosul, Mosul, Iraq.
Corresponding author: roqaya.20csp48@student.uomosul.edu.iq[*] , Ghaydabdulaziz@uomosul.edu.iq
[*]ORCID ID: https://orcid.org/0000-0001-7612-9888[*] , https://orcid.org/0000-0001-6008-2595
Mobile number: 07722811799
Mobile number: 07701797579

Received……..., Accepted……., Published……..

Abstract:

A trend topic is a popular event that people are looking for or posting more widely than others. It may be the name of a person, event, place, game, or any other specific thing. In this era, trends are a very important topic in the field of business, as it consists of people's opinions and words and writing about what is going on in their minds, and the topic that receives more attention than others is called the trending topic. The behavior of the trend topic is usually monitored and is important by marketers and owners of companies and factories, and is also important for those who work in the field of media and politics. To access these popular topics in our research, Twitter is used as the data source. This search gets big data (tweets) in real time for processing. Natural language processing methods were used to process the data to filter it to be ready for analysis. The TF_IDF technique was used, which is a statistical intelligent algorithm based on finding the importance of each term in all documents (tweets), in addition to using the vectorization within the results of the algorithm used in order to give a sequence of the importance of each topic from the most important to the least important, in order to extract the topic that is the trend among people belonging to a specific region and a particular language. This proposed method fulfilled the purpose of the research and gave satisfactory results.

Keywords: NLP, TF_IDF, Trending topics, Twitter API, Vectorization.

المجلة الإلكترونية الشاملة متعددة التخصصات
العدد الثامن والخمسون شهر (٤) ٢٠٢٣

استخراج الموضوع الرائج بواسطة تقنية تردد المصطلح-تردد المستند العكسي

رقية خليل كشمولة[1] **غيداء عبد العزيز الطالب[2]**

[2,1] قسم علوم الحاسوب، كلية علوم الحاسوب والرياضيات، جامعة الموصل، الموصل، العراق.

الخلاصة:

موضوع الاتجاه هو حدث شائع يبحث عنه الأشخاص أو ينشرونه على نطاق أوسع من غيرهم. قد يكون اسم شخص أو حدث أو مكان أو لعبة أو أي شيء آخر محدد. في هذا العصر تعد الاتجاهات موضوعًا مهمًا جدًا في مجال الأعمال ، حيث يتكون من آراء الناس وكلماتهم وكتابتهم عما يدور في أذهانهم ، والموضوع الذي يحظى باهتمام أكثر من غيره يسمى موضوع الاتجاه. عادة ما يتم مراقبة سلوك موضوع الاتجاه وهو مهم من قبل المسوقين وأصحاب الشركات والمصانع ، كما أنه مهم أيضًا لمن يعملون في مجال الإعلام والسياسة. للوصول إلى هذه الموضوعات الشائعة في بحثنا ، يتم استخدام tterTwi كمصدر للبيانات. يحصل هذا البحث على بيانات ضخمة (تغريدات) في الوقت الفعلي للمعالجة. تم استخدام طرق معالجة اللغة الطبيعية لمعالجة البيانات لتصفيتها لتكون جاهزة للتحليل. تم استخدام تقنية TF_IDF ، وهي خوارزمية إحصائية ذكية تعتمد على إيجاد أهمية كل مصطلح في جميع الوثائق (التغريدات) ، بالإضافة إلى استخدام التوجيه ضمن نتائج الخوارزمية المستخدمة لإعطاء تسلسل لأهمية كل مصطلح من الأهم إلى الأقل أهمية ، وذلك لاستخراج الموضوع الذي هو الاتجاه السائد بين الناس المنتمين إلى منطقة معينة ولغة معينة. حققت هذه الطريقة المقترحة الغرض من البحث وأعطت نتائج مرضية.

الكلمات المفتاحية: معالجة اللغة الطبيعية, الموضوع الرائج, واجهة برمجة تطبيقات تويتر, تردد المفردة- معكوس تردد الوثيقة, التوجيه.

Introduction:

Overview

In the past few years, social media has become the new social life. So that people share their hobbies, opinions, ideas, favorite places and almost everything around them at any time. One of the most important platforms of social media is Twitter. Twitter platform became a popular blog and the number of users of the latest Twitter statistics is estimated at 192 million daily active users at the end of 2020. And according to the stats: About half a billion tweets are sent every day, which means about 5,787 tweets per second. According to stats of (Http://www.oberlo.com). It is possible through Twitter to know the trending topics and the most interest objects to people. Therefore, this data must be used in scientific aspects, including finding popular topics for use in media, marketing, raise awareness, voting, politics and others. To reach these popular topics in our research, Twitter is used as a source of data. Twitter is one of the widespread and free social networking platforms that allow customers to post their thoughts, observations and experiences, or retweet previous posts. It is can include links to videos, photos, web pages, or something online. Twitter is one of the most used and most polarized platforms by academic students now a days[1.]

After pulling the data from Twitter and storing it in a file, the data is cleaned, then, one of the techniques of Natural Language Processing (NLP) is applied. Natural language processing is the learning of computer software that makes natural, or human, language as input to it [2].

Related Works:

Several data analysis techniques were used in references of this work, which can be summarized by the following works:

In 2018, The researchers Ahmed Rafea and Nada A. Gaballah describe how to extract the trending topic for a user that are speak Arabic who use twitter [3]. That's done by apply Hierarchical Agglomerative Clustering approach on textual data that are taken from twitter, through accounts that follows by users without needing to browse all the tweets. (HAG) defined as a combining of clusters depending on their similarity [4]. For best results, they tried different techniques of topics extraction and clustering. To check the validity of the results, they applied their approach to 12 sets of data from three different subjects and of different sizes. The accuracy showed that average recall= 0.84 and average F1_Measure= 0.71.

In 2017, the researchers Minor Eduardo Quesada Grosso, Edgar Casasola, et.al extract the topics and finding trending from those topics [5], that's done by a comparison between single Bursty Biterm Topic Model (BBTM) and the modified BBTM with term discrimination while the data taken from Twitter (tweets). (BBTM) used to model the obstetric operation of the word co-occurrence styles in short textual data such as tweets [6]. This work aims to reduce the required processing and got equally well results. The technique used to show on average time of processing 8.090 minutes and average coherence is 0.71.

In 2011, the researchers Arkaitz Zubiaga, Damiano Spina, et.al revolves around Classifying Topics to organize topics by type [7]. They used the Support Vector Machines (SVM) algorithm and applied for tweets from Twitter. (SVM) is effective prediction and modeling tool for a different application [8]. This proposed method, in addition to requiring no external data, is an accurate way to organize trending topics in real time. The accuracy for Bag-of-words is 0.752, and Accuracy for Twitter features 0.784.

In 2013, the researchers James Benhardus and Jugal Kalita proposed method to use techniques of NLP on streaming of data from streaming API platform [9], that collected from Twitter to analyze a large amount of streaming data for identifying the trending topics. the techniques they used are: Exp1: Term frequency-inverse document frequency (TF-IDF) analysis, (TF_IDF) is one of the most common criterions which is used in text mining, search reconciliation, and information retrieval [10]. Exp2: relative normalized term frequency analysis was performed on the documents to identify the trending topics. Relative Normalized Term Frequency used to reduce the impact of long vs. short documents and reverse the real importance of a keyword to a document [9]. The accuracy for both experiments were as follows: Max. precision=0.5, Max. recall= 0.58 and Max. F-measure =0.53.

In 2020, the researchers Jack Hughes, Seth Aycock, et.al a method to determine the actually trending topics by relation to a previous known topic [11], by apply preprocessing textual data (approach similar to "two-point trends") and Bayesian approach on posts (textual data) that are taken from Underground hacking forum in English-language extends over more than ten years of activity, with posts containing orthographic variation, misspellings, slang and acronyms. They stated a case of new use for the log-odds tool. Their approach of statistics supports the analysis over time of discussion topics and linguistic changeable, without a need to train model in each time interval. Bayesian techniques depends on the specification of a Previous possibility and the probability (from the evolution models and data) to

specify the next probability of hypotheses [12]. To evaluate their approach, they use TF-IDF to compare results with it. Finally, the accuracy results for log-odds tool 0.979.

In 2020, the researchers Meysam Asgari-Chenaghlu, Narjes Nikzad-Khasmakhi, et.al proposed framework to detect trending topics that are related to COVID-19 from tweets in the form of sentences that are easy to read and understand by human. They needed to collect 1.6 million tweets related to COVID-19 [13]. Then, by Transformer they extracted the sentences embedding. After that, they used K Means Clustering Algorithm to group tweets those similar, K Means Clustering is a data mining technique that can be used for unsupervised machine learning. Generally, it is stratified to continuous and numeric data [4]. Finally, they applied summarization on all clusters to get a short summary. The accuracy results for this work were: Precision= 0.67, Recall= 0.58 and F1= 0.62.

In 2013, the researchers Nargis Pervin, Fang Fang, et.al describe a way to execute and extract the trending topics from the high-rate stream in real-time of microblogging posts [14]. They applied Trend Miner method to find word clusters then cluster generation for trending topics (The two-word clusters) that taken from popular microblogging social networks, such as Twitter. Trend Miner platform, self-service and its conjectural web-based for fast-fire visualization of asset date and time string-based operation [15]. The results of research shows that their system can find stories that are “hot” from high velocity Twitter scale data streams. The accuracy results of this research are: Max. Recall 0.83 and Max. precision 0.42.

In 2017, The researchers Syafruddin Syarif, Anwar, et.al aimed to assisting the Makassar City government to the trending topic prediction which happen by analyze the historical stack in data mining [16]. They applied Application Programming Interface (API) which have been used to collect the data from tribunnews.com, rakyatku.com, rakyatsulsel.com, and pojoksulsel.com and twitter, then applying K-Nearest Neighbor (KNN) to analyze data. (KNN) is non parametric mechanism it well known in the statistical manner ranking, owing to its effectiveness, simplicity and intuitiveness [17]. The accuracy was follows: Max. accuracy= 0.8113, Max. precision= 0.8235 and Max. recall 0.7037.

In 2015, The researchers Jose Hurtado, Shihong Huang, et.al aimed to use association analysis and ensemble forecasting to discover topics automatically from set of data and forecast their evolving trend in the near future [18]. They collected scientific papers from ACM-KDD, IEEEICDM, SIAM-SDM, and ICML conferences to apply sentence-level association rule mining, temporal correlation analysis and Clique Percolation Method (CPM) on it, (CPM) in this process, a performed of an operation of enumerating highly coherent maximal document cliques in a random diagram, while those highly neighboring cliques are mixed to form naturally mixed clusters [19]. The accuracy was measured by precision and it reaches 0.923.

In 2016, the researchers Halima Banu S and S Chitrakala proposed a new way called Volume Foreground Dynamic Topic Modelling (VF-DTM) which is applied on tweets [20], where (VF-DTM) uses sentiment classification and tweet summarization then distills the noisy content and extracts the foreground tweets from the corpus in order to get the trending topics. (VF-DTM) is topic models are prepared for definite data (categorical data) [21], It is pick up the growth of topics and trends in time series data [22]. The result of the accuracy of this research reches 0.8.

In 2021, the researchers Poonam Vijay Tijare, Jhansi Rani P. aimed to find trending event in social media by use sentiment analysis, K means and a second order derivative algorithm [23]. The Suggested

framework done on four different. data set in different topics of tweets from twitter api, include (Political, Social, Scientific and Sport). The result of accuracy was positive and in range of 0.88 to 0.95.

In 2021, the researchers Michael Charnine, Alexey Tishchenko, et.al shows the of a way of prediction of trending topics by the visualization of the long-term [24]. They used method of machine learning methods called Cat Boost. Cat Boost classifier is efficient in predicting categorical feature. Cat Boost is an implementation of gradient boosting, which makes use of binary decision trees as base predictors [25]. The proposed method done on textual dataset in scientific range included 5 million publications, was taken from best conferences in data mining and artificial intelligence range. The result of accuracy was 0.60. ,

In 2021, Ali Daud, Faizan Abbas, et.al aimed noting the scientific growth by finding junior researchers, who share a post containing a trending topic, therefore they extracted trends from A Miner data set and by applied Top Topics Rising Star Rank (HTRS-Rank) method with (TF_IDF) and with WordNet on the textual data (publications) [26]. Word Net method is an on-line lexical reference system developed at Princeton University. WordNet can also be seen as an ontology for natural language terms [27]. They found the result of HTRS-Rank with (TF_IDF) better than HTRS-Rank with WordNet.

In 2020, the researchers Syed Tanzeel Rabani, Qamar Rayees Khan Akib, et.al, aimed to find the feasibility in detecting and differentiating the suicidal tweets from no suicidal tweets by analyzing the data from social media platforms [28]. They applied WEKA tool was used to implement various machine and ensemble learning algorithms on textual data (tweets) taken from Twitter API in real time. The five machine learning methods which were implemented in weka are Naïve Bayes (NB), Multinomial Naïve Bayes (MNB), Decision tree (REPTree and J48), Logistic Regression (LR) and Support Vector Machine (SMO). The SMO algorithm scored the highest results and it is as follows: Accuracy = 93.5%, Precision = 94.6% and Recall = 92.3%.

In 2022, Sarah Sameer and Suhad Faisal Behadili, aimed to analyze and simulate biochemical real test data for uncovering the relationships among the tests, and how each of them impacts others [29]. They used textual data from Iraqi private biochemical laboratory. They applied many experiments on these data, it was the preprocessing step performed, to make the dataset analyzable by supervised techniques such as Linear Discriminant Analysis (LDA), Classification and Regression Tree (CART), Logistic Regression (LR), K-Nearest Neighbor (K-NN), Naïve Bays (NB), and Support Vector Machine (SVM) techniques. CART gives clear results with high accuracy between the six supervised algorithms. the highest accuracy was 97%.

Trending Topic:

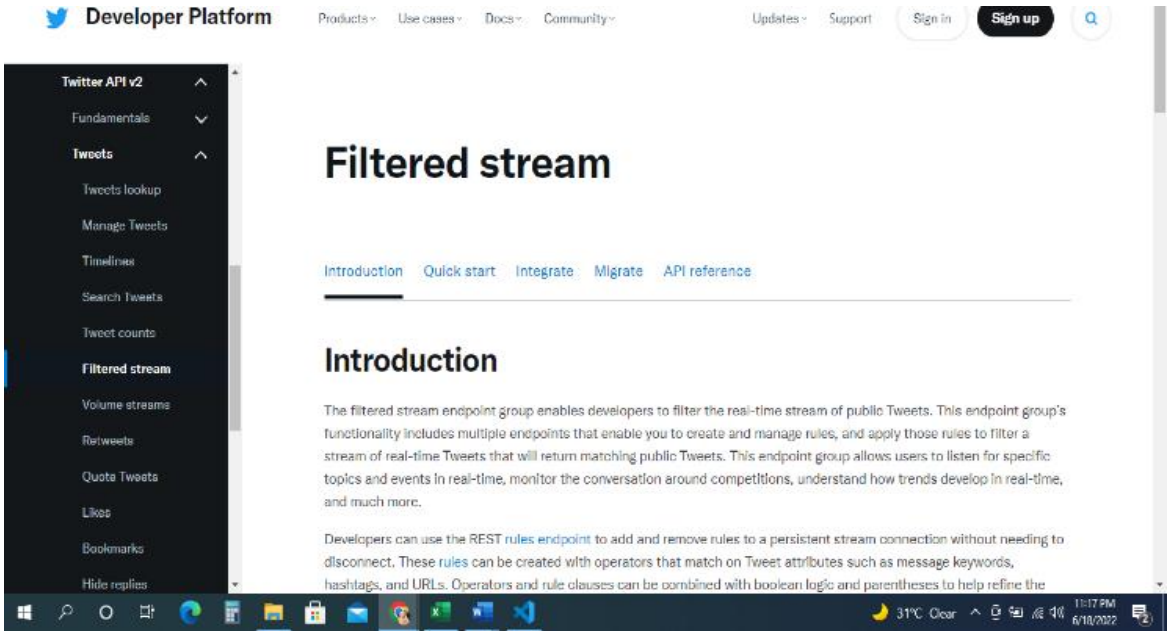
A trending topic on Twitter consists of a term or a clip and a time period. When the volume of tweets exceeds the expected level for a particular topic in period time, the topic is in response to an external event and the tweets are related to that event [30]. A trending topic is usually indicated with a hashtag "#" and a trending topic may be a word or a phrase. A trending topic is the name of a person, place, game, event, or other specific thing [31].

Samples of topics that were trend at previous several times are the following:

- Maradona
- Real Madrid
- Russia
- #Okrania
- #Suprt league
- #FIFA_World_Cup
- #COVID_19
- #Corona_virus
- #2022
- #FCBarcelona
- others.

Data Collection:

Big data were collected from the online Twitter API. Which allows access to Twitter programmatically, in unique and in progress ways. The main factors that can be used are: Tweets, Direct Messages, Retweets, Spaces, users, Lists and others[1].Twitter Developer platform is a platform that provides access to the Twitter portal, which allows the use of Twitter API features. This is done by obtaining the developer's account by communicating with the official website of Twitter and being informed of the type of data required and stating the reasons. When the request is accepted, secret keys are provided to each customer to give them access to the data provided by Twitter, which is to provide a certain number of data per month for each project depending on the type of request as show in Fig.1.



[1] https://developer.twitter.com/en/docs/twitter-api

Figure 1. Interface of a personal account in Twitter developer platform.

Using a program that supports Tweepy library, and using the powers given by Twitter through the keys that allow to withdraw tweets from Twitter. Filters can be used to specify the details to be withdrawn, such as the date and time of the tweet, the username, the text of the tweet, the location of the tweet and more, it is also possible to specify the language of the text of the tweets or specify more than one language and also can specify the coordinates of the locations of users whose tweets are pulled and other features provided by the Twitter API.

For example: By giving the coordinates of Arab world with English language, a request was done to get the text of tweets, username, date, and time for ten tweets, which are shown as an Example in Table 1 .

TABLE 1. THE RESPONSE FROM TWITTER TO OUR REQUEST

Date/Time	User name	Tweets
2022-07-03 14:44:30+00:00	_sseuwitttie	BACK IT UP/HIT + # SECTOR 17
2022-07-03 14:44:32+00:00	MediaVidi	RT @ToolsTipsNews: Digital Advertising in the Age of Automation # # #MicrosoftAds # #MicrosoftAdvertising # #DigitalMarketing # #OnlineMark…
2022-07-03 14:44:33+00:00	MediaVidi	RT @ToolsTipsNews: Grow Your Scrap Car Company #PPCAdvertising # #PPCTips # #GoogleAdwords # #GoogleAds # [Video] https://t.co/mUB2oCThOd
2022-07-03 14:44:36+00:00	itboypjjm	RT @PJM_data: "Lie" has surpassed 180 Million streams on Spotify Jimin now sets the record as the First and Only Korean soloist to hav…
2022-07-03 14:44:39+00:00	jiwonskart	RT @grIszne: lf kahati will get wonyoung and unsealed albums # wts lfb kep1er yeseo soundwave doublast ive gaeul withmuu pob photocard pc…
2022-07-03 14:44:39+00:00	ClickbySBhamidi	@HYDTP pillar # 210 attapur road t/r Silver swift desire driven by a minor causing nuisance and honking
2022-07-03 14:44:41+00:00	Newshawkerug	This how people look ugly in their professional. Uganda zzabu. # struggle continues. https://t.co/ravpaFOON8
2022-07-03 14:44:42+00:00	Hoonpawz	# tags for #ENHYPEN !
2022-07-03 14:44:42+00:00	AlexKittoe	RT @gbergphoto: GM y'all. Two secondary sales on ASOTT over the last 48 hours. A SIGN OF THE TIMES token # 85 just got picked up by @_coope…
2022-07-03	MarioKrenn6240	@rguha @KohulanRajan @egonwillighagen import selfies,

14:44:44+00:00		random alphabet=selfies.get_semantic_robust_alphabet() rnd_selfies=''.join(random.sample(list(alphabet), 9)) rnd_smiles=selfies.decoder(rnd_selfies) print(rnd_smiles) # Gives crazy molecules. You can adjust alphabet for simpler structures (e.g. no ions etc)

Data Preprocessing:

This stage includes many steps to prepare the data in a suitable form for processing, which include the following steps as shown in Fig. 2.

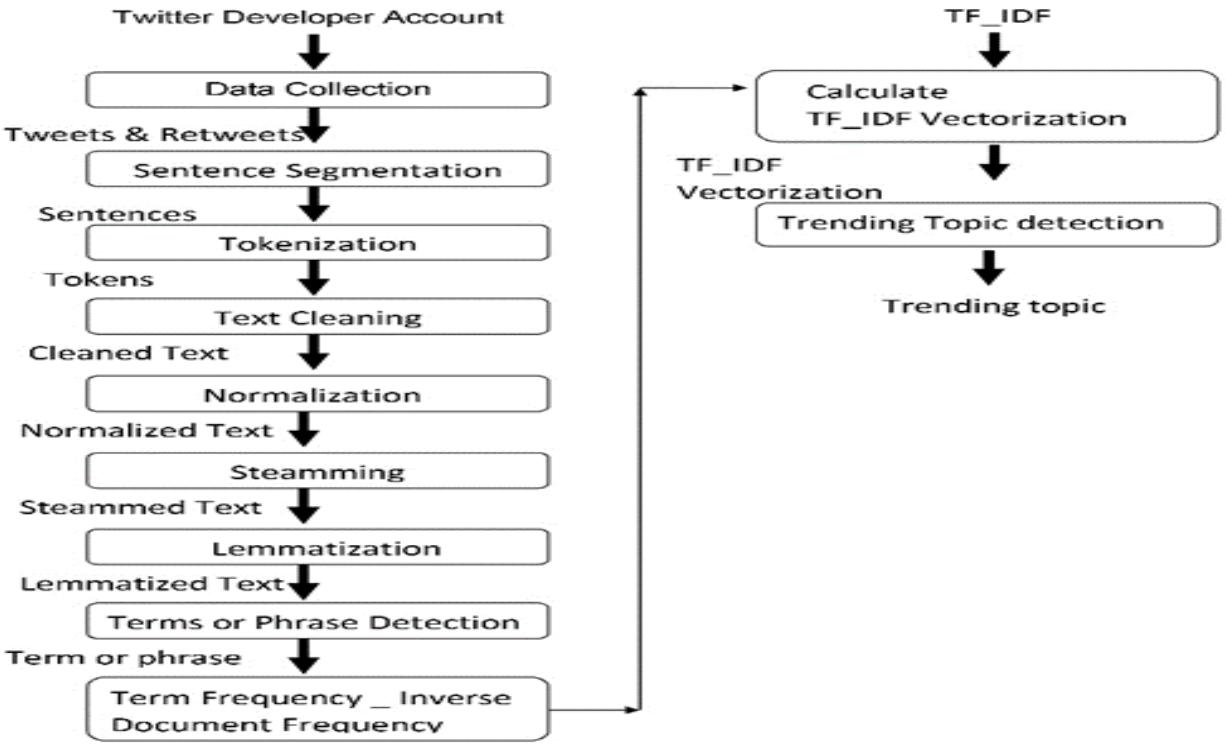


Figure 2. The Pipeline of processing to finding the Trending Topic

- **Sentences segmentation:**

Includes cutting the data into sentences (Tweets) are presented in Fig .3.

BACK IT UP/HIT + # SECTOR 17 end_Tweet RT @ToolsTipsNews: Digital Advertising in the Age of Automation # # #MicrosoftAds # #MicrosoftAdvertising # #DigitalMarketing # #OnlineMark… end_Tweet

RT @ToolsTipsNews: Grow Your Scrap Car Company #PPCAdvertising # #PPCTips # #GoogleAdwords # #GoogleAds # [Video] https://t.co/mUB2oCThOd end_Tweet
RT @PJM_data: "Lie" has surpassed 180 Million streams on Spotify

Jimin now sets the record as the First and Only Korean soloist to hav… end_Tweet
RT @grIszne: lf kahati will get wonyoung and unsealed albums

wts lfb kep1er yeseo soundwave doublast ive gaeul withmuu pob photocard pc… end_Tweet
@HYDTP pillar # 210 attapur road t/r Silver swift desire driven by a minor causing nuisance and honking end_Tweet
This how people look ugly in their professional. Uganda zzabu.
struggle continues. https://t.co/ravpaFOON8 end_Tweet
tags for #ENHYPEN ! end_Tweet
RT @gbergphoto: GM y'all. Two secondary sales on ASOTT over the last 48 hours. A SIGN OF THE TIMES token # 85 just got picked up by @_coope… end_Tweet
@rguha @KohulanRajan @egonwillighagen import selfies, random
alphabet=selfies.get_semantic_robust_alphabet()
rnd_selfies=''.join(random.sample(list(alphabet), 9))
rnd_smiles=selfies.decoder(rnd_selfies)
print(rnd_smiles)

Gives crazy molecules. You can adjust alphabet for simpler structures (e.g. no ions etc) end_Tweet

Figure 3. The Data Segmented into Sentences

- **Tokenization:**

Include tokenize sentences into terms and phrases, while keeping the token related to the hashtag symbol uncut as shown in Fig.4.

[['BACK', 'IT', 'UPHIT', '', '', 'SECTOR', '17'], ['RT', 'ToolsTipsNews', 'Digital', 'Advertising', 'in', 'the', 'Age', 'of', 'Automation', '', '', '#MicrosoftAds', '', '#MicrosoftAdvertising', '', '#DigitalMarketing', '', '#OnlineMark…'], ['RT', 'ToolsTipsNews', 'Grow', 'Your', 'Scrap', 'Car', 'Company', '#PPCAdvertising', '', '#PPCTips', '', '#GoogleAdwords', '', '#GoogleAds', '', 'Video', 'httpst.comUB2oCThOd'], ['RT', 'PJMdata', 'Lie', 'has', 'surpassed', '180', 'Million', 'streams', 'on', 'Spotify', 'Jimin', 'now', 'sets', 'the', 'record', 'as', 'the', 'First', 'and', 'Only', 'Korean', 'soloist', 'to', 'hav…'], ['RT', 'grIszne', 'lf', 'kahati', 'will', 'get', 'wonyoung', 'and', 'unsealed', 'albums', '', 'wts', 'lfb', 'kep1er', 'yeseo', 'soundwave', 'doublast', 'ive', 'gaeul', 'withmuu', 'pob', 'photocard', 'pc…'], ['HYDTP', 'pillar', '', '210', 'attapur', 'road', 'tr', 'Silver', 'swift', 'desire', 'driven', 'by', 'a', 'minor', 'causing', 'nuisance', 'and', 'honking'], ['This', 'how', 'people', 'look', 'ugly', 'in', 'their', 'professional', 'Uganda', 'zzabu', '', 'struggle', 'continues', 'httpst.coravpaFOON8'], ['', 'tags', 'for', '#ENHYPEN', ''], ['RT', 'gbergphoto', 'GM', 'y'all', 'Two', 'secondary', 'sales', 'on', 'ASOTT', 'over', 'the', 'last', '48', 'hours', 'A', 'SIGN', 'OF', 'THE', 'TIMES', 'token', '', '85', 'just', 'got', 'picked', 'up', 'by', '_coope…'], ['rguha', 'KohulanRajan', 'egonwillighagen', 'import', 'selfies', 'random', 'alphabetselfiesgetsemanticrobustalphabet)',

'rndselfiesjoinrandomsamplelistalphabet,', '9', 'rndsmilesselfiesdecoderrndselfies', 'printrndsmiles', '', 'Gives', 'crazy', 'molecules', 'You', 'can', 'adjust', 'alphabet', 'for', 'simpler', 'structures', 'eg', 'no', 'ions', 'etc']]

Figure 4. Terms or Phrases Produced by Tokenization

- **Text Cleaning:**

Includes removing punctuation except for the "#" symbol, removing emojis, URLs, stop words for the chosen language and removing the "RT" retweet symbol, because these repetitions do not often affect the meaning of the words, but when they remain, they will be the most frequent in the document as shown in Fig .5.

[['BACK', 'UP HIT', '', '', 'SECTOR', '17'], ['ToolsTipsNews', 'Digital', 'Advertising', 'Age', 'Automation', '', '#MicrosoftAds', '#MicrosoftAdvertising', '#DigitalMarketing', '#OnlineMark…'], ['ToolsTipsNews', 'Grow', 'Scrap', 'Car', 'Company', '#PPCAdvertising', '', '#PPCTips', '#GoogleAdwords', '#GoogleAds', 'Video', ' '], ['PJM data', 'Lie', 'surpassed', '180', 'Million', 'streams', 'Spotify', 'Jimin', 'sets', 'record', 'First', 'Korean', 'soloist', 'hav'], ['grIszne', 'lf', 'kahati', 'get', 'wonyoung', 'unsealed', 'albums', '', 'wts', 'lfb', 'kep1er', 'yeseo', 'soundwave', 'doublast', 'ive', 'gaeul', 'withmuu', 'pob', 'photocard', 'pc'], ['HYDTP', 'pillar', '', '210', 'attapur', 'road', 't r', 'Silver', 'swift', 'desire', 'driven', 'minor', 'causing', 'nuisance', 'honking'], ['people', 'look', 'ugly', 'professional', 'Uganda', 'zzabu', '', 'struggle', 'continues', ' '], ['', 'tags', '#ENHYPEN', ''], ['gbergphoto', 'GM', 'y'all', 'Two', 'secondary', 'sales', 'ASOTT', 'last', '48', 'hours', 'SIGN', 'TIMES', 'token',
'', '85', 'got', 'picked', 'coope'], ['rguha', 'KohulanRajan', 'egonwillighagen', 'import', 'selfies', 'random', 'alphabet selfies get semantic robust alphabet', 'rnd selfies join random sample list alphabet', '9', 'rnd smiles selfies decoder rnd selfies', 'print rnd smiles', '', 'Gives', 'crazy', 'molecules', 'adjust', 'alphabet', 'simpler', 'structures', 'e g', 'ions', 'etc']]

Figure 5. Cleaned Text

- **Normalization:**

Includes converting all letters to lowercase as shown in Fig .6.

[['back', 'up hit', '', '', 'sector', '17'], ['toolstipsnew ', 'digital', 'advertising', 'age', 'automation', '', '#microsoftads', '#microsoftadvertising', '#digitalmarketing', '#onlinemark…'], ['toolstipsnews', 'grow', 'scrap', 'car', 'company', '#ppcadvertising', '', '#ppctips', '#googleadwords', '#googleads', 'video', ' '], ['pjm data', 'lie', 'surpassed', '180', 'million', 'streams', 'spotify', 'Jimin', 'sets', 'record', 'first', 'korean', 'soloist', 'hav'], ['grIszne', 'lf', 'kahati', 'get', 'wonyoung', 'unsealed', 'albums', '', 'wts', 'lfb', 'kep1er', 'yeseo', 'soundwave', 'doublast', 'ive', 'gaeul', 'withmuu', 'pob', 'photocard', 'pc'], ['hydtp', 'pillar', '', '210', 'attapur', 'road', 't r', 'silver', 'swift', 'desire', 'driven', 'minor', 'causing', 'nuisance', 'honking'], ['people', 'look', 'ugly', 'professional', 'uganda', 'zzabu', '', 'struggle', 'continues', ' '], ['', 'tags', '#enhypen', ''], ['gbergphoto', 'gm', 'y'all', 'two', 'secondary', 'sales', 'asott', 'last', '48', 'hours', 'sign', 'times', 'token',
'', '85', 'got', 'picked', 'coope'], ['rguha', 'kohulanrajan', 'egonwillighagen', 'import', 'selfies', 'random', 'alphabet selfies get semantic robust alphabet', 'rnd selfies join random sample list alphabet', '9', 'rnd smiles selfies decoder rnd selfies', 'print rnd smiles', '', 'gives', 'crazy', 'molecules', 'adjust', 'alphabet', 'simpler', 'structures', 'e g', 'ions', 'etc']]

Figure 6. Normalized Text

- **Steamming:**

Usually means to a heuristic procedure that removes prefixes and suffixes from the word to get the root of it, and sometimes includes the remove of affixes which is be derivational [32] as shown in Fig.7.

[['back', 'up hit', '', '', 'sector', '17'], ['toolstipsnew', 'digit', 'advertis', 'age', 'autom', '', '#microsoftad', '#microsoftadvertis', '#digitalmarket', '#onlinemark…'],
['toolstipsnew', 'grow', 'scrap', 'car', 'compani', '#ppcadvertis', '', '#ppctip', '#googleadword', '#googlead', 'video', ' '], ['pjm data', 'lie', 'surpass', '180', 'million', 'stream', 'spotifi', 'jimin', 'set', 'record', 'first', 'korean', 'soloist', 'hav'], ['griszn', 'lf', 'kahati', 'get', 'wonyoung', 'unseal', 'album', '', 'wt', 'lfb', 'kep1er', 'yeseo', 'soundwav', 'doublast', 'ive', 'gaeul', 'withmuu', 'pob', 'photocard', 'pc'], ['hydtp', 'pillar', '', '210', 'attapur', 'road', 't r', 'silver', 'swift', 'desir', 'driven', 'minor', 'caus', 'nuisanc', 'honk'], ['peopl', 'look', 'ugli', 'profession', 'uganda', 'zzabu', '', 'struggl', 'continu', ' '], ['', 'tag', '#enhypen', ''], ['gbergphoto', 'gm', 'y'all', 'two', 'secondari', 'sale', 'asott', 'last', '48', 'hour', 'sign', 'time', 'token', '', '85', 'got', 'pick', 'coop'], ['rguha', 'kohulanrajan', 'egonwillighagen', 'import', 'selfi', 'random', 'alphabet selfies get semantic robust alphabet', 'rnd selfies join random sample list alphabet', '9', 'rnd smiles selfies decoder rnd selfi', 'print rnd smil', '', 'give', 'crazi', 'molecul', 'adjust', 'alphabet', 'simpler', 'structur', 'e g', 'ion', 'etc']

Figure 7. Steammed Terms or Phrases (Roots)

- **Lemmatization:**

The target of both stemming and lemmatization is to decrease inflectional forms and occasionally derivationally attached forms of a word to a common standard form. However, the two mechanisms different. Lemmatization usually means to access to the root by use of a vocabulary and morphological analysis of words, usually goaling to remove inflectional ends of words only and access to root which is called the lemma [32] as shown in Fig .8.

[['back', 'up hit', '', '', 'sector', '17'], ['toolstipsnew', 'digit', 'advertis', 'age', 'autom', '', '#microsoftad', '#microsoftadvertis', '#digitalmarket', '#onlinemark…'], ['toolstipsnew', 'grow', 'scrap', 'car', 'compani', '#ppcadvertis', '', '#ppctip', '#googleadword', '#googlead', 'video', ' '], ['pjm data', 'lie', 'surpass', '180', 'million', 'stream', 'spotifi', 'jimin', 'set', 'record', 'first', 'korean', 'soloist', 'hav'], ['griszn', 'lf', 'kahati', 'get', 'wonyoung', 'unseal', 'album', '', 'wt', 'lfb', 'kep1er', 'yeseo', 'soundwav', 'doublast', 'ive', 'gaeul', 'withmuu', 'pob', 'photocard', 'pc'], ['hydtp', 'pillar', '', '210', 'attapur', 'road', 't r', 'silver', 'swift', 'desir', 'driven', 'minor', 'caus', 'nuisanc', 'honk'], ['peopl', 'look', 'ugli', 'profession', 'uganda', 'zzabu', '', 'struggl', 'continu', ' '], ['', 'tag', '#enhypen', ''], ['gbergphoto', 'gm', 'y'all', 'two', 'secondari', 'sale', 'asott', 'last', '48', 'hour', 'sign', 'time', 'token', '', '85', 'got', 'pick', 'coop'], ['rguha', 'kohulanrajan', 'egonwillighagen', 'import', 'selfi', 'random', 'alphabet selfies get semantic robust alphabet', 'rnd selfies join random sample list alphabet', '9', 'rnd smiles selfies decoder rnd selfi', 'print rnd smil', '', 'give', 'crazi', 'molecul', 'adjust', 'alphabet', 'simpler', 'structur', 'e g', 'ion', 'etc']]

Figure 8. Terms or Phrasees as Cleaned Roots

TF-IDF:

Term frequency–inverse document frequency, is a numeral statistic which is aimed to mirror the importance a word in a document in a corpus or collection [33].

*TF-IDF score = TF * IDF*

- **Term frequency:**

Is the number of repetitions of a term, phrase, or sentence in a part of content [34].

TF (t, d) = (Number of occurrences of term t in document d)/ (Total number of terms in the document d)

- **Inverse document frequency:**

Is a mechanism which is minimize the value of the most frequent keywords and raise the value of less frequent words or unique words and phrases in the content. generally, IDF formula gives an idea on terms which is have more weight and value between the terms [34].

IDF(t) = loge ((Total number of documents in the corpus)/ (Number of documents with term t in them))

Vectorization:

Vectorization Is used in natural language processing applications and necessary for dealings with text data. Vectorization cement the machines to realize the contents of the text data by transmutation them to numeral representations with meaning. The result of TF IDF Vectorization matrix as shown in Fig .9, term that has the highest value is the Trending Topic [35].

	17	180	210	48	85	adjust	advertis
0	0.076029	0.076029	0.076029	0.076029	0.076029	0.076029	0.076029

	age	album	all	alphabet	asott	attapur	autom
0	0.076029	0.076029	0.076029	0.304114	0.076029	0.076029	0.076029

	back	car	caus	compani	continu	coop	crazi
0	0.076029	0.076029	0.076029	0.076029	0.076029	0.076029	0.076029

	data	decoder	desir	digit	digitalmarket	doublast	driven
0	0.076029	0.076029	0.076029	0.076029	0.076029	0.076029	0.076029

	egonwillighagen	enhypen	etc	first	gaeul	gbergphoto
0	0.076029	0.076029	0.076029	0.076029	0.076029	0.076029

	get	give	gm	googlead	googleadword	got	griszn
0	0.152057	0.076029	0.076029	0.076029	0.076029	0.076029	0.076029

	grow	hav	hit	honk	hour	hydtp	import
0	0.076029	0.076029	0.076029	0.076029	0.076029	0.076029	0.076029

	ion	ive	jimin	join	kahati	kep1er	kohulanrajan
0	0.076029	0.076029	0.076029	0.076029	0.076029	0.076029	0.076029

	korean	last	lf	lfb	lie	list	look
0	0.076029	0.076029	0.076029	0.076029	0.076029	0.076029	0.076029

	microsoftad	microsoftadvertis	million	minor	molecul	nuisanc
0	0.076029	0.076029	0.076029	0.076029	0.076029	0.076029

	onlinemark	pc	peopl	photocard	pick	pillar	pjm
0	0.076029	0.076029	0.076029	0.076029	0.076029	0.076029	0.076029

	pob	ppcadvertis	ppctip	print	profession	random	record
0	0.076029	0.076029	0.076029	0.076029	0.076029	0.152057	0.076029

	rguha	rnd	road	robust	sale	sample	scrap
0	0.076029	0.304114	0.076029	0.076029	0.076029	0.076029	0.076029

	secondari	sector	selfi	selfies	semantic	set	sign
0	0.076029	0.076029	0.152057	0.228086	0.076029	0.076029	0.076029

	silver	simpler	smil	smiles	soloist	soundwav	spotifi
0	0.076029	0.076029	0.076029	0.076029	0.076029	0.076029	0.076029

	stream	structur	struggl	surpass	swift	tag	time
0	0.076029	0.076029	0.076029	0.076029	0.076029	0.076029	0.076029

	token	toolstipsnew	two	uganda	ugli	unseal	up
0	0.076029	0.152057	0.076029	0.076029	0.076029	0.076029	0.076029

	video	withmuu	wonyoung	wt	yeseo	zzabu
0	0.076029	0.076029	0.076029	0.076029	0.076029	0.076029

Figure 9. TF IDF Vectorization results

Accuracy Metrics:

In comparison with the trending topics on Twitter generated by Twitter method, and considering that the true positive is matching of our results with the results of Twitter method, and considering that the false positive is the mismatch of our results with the results of Twitter method, the average accuracy according to the precision scale was not less than 0.60, This was accomplished by comparing the results from proposed method by running it several times with the results obtained from twitter itself.

Conclusion:

When collecting data and requesting tweets, it was found that the tweets were incomplete, which necessitated a request to obtain the "full text". URLs have been canceled because the link is usually an image, video, web page, application or anything else, which increases the burden on the technology

used. The TF_IDF technique helped to discover the trending topic because it is based on the repetition of the term or phrase in the document and the corpus.

Authors' Contributions: Al-Talib G. was the Conception owner, also she did the revision and proofreading. And Kashmola R. designed the idea and she did acquisition the data, interpretation, and drafting the MS. The article was finally approved by two writers after thorough consideration.

References:

1. Mahdikhani M. Predicting the popularity of tweets by analyzing public opinion and emotions in different stages of Covid-19 pandemic. l of Information Management Data International Journa .Insights. 2022 Apr;2(1):100053, Doi: 10.1016/j.jjimei.2021.100053
2. Lamurias A, Couto FM. Text Mining for Bioinformatics Using Biomedical Literature. In: .11–er; 2019. p. 602Encyclopedia of Bioinformatics and Computational Biology. Elsevi, Doi: 10.1016/B978-0-12-809633-8.20409-3.
3. Rafea A, Gaballah NA. Trending Topic Extraction from Twitter for an Arabic Speaking User. .In: ISCA, CATA. Las Vegas, Nevada, USA; 2018, Link: https://www.researchgate.net/publication/324452007_Trending_Topic_Extraction_from_Twitter_for_an_Arabic_Speaking_User
4. Wei Z, Liang C, Tang H. A Cross-Regional Scheduling Strategy of Waste Collection and Transportation Based on an Improved Hierarchical Agglomerative Clustering Algorithm. .17–Comput Intell Neurosci. 2022 Mar 30;2022:
5. Eduardo Quesada Grosso M, Casasola E, Antonio Leoni de León J. Trending Topic Extraction using Topic Models and Biterm Discrimination. .Clei Electronic Journal Vol. 20. 2017.
6. Chen W, Wang J, Zhang Y, Yan H, Li X. User Based Aggregation for Biterm Topic Model. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). .94–Stroudsburg, PA, USA: Association for Computational Linguistics; 2015. p. 489
7. Zubiaga A, Spina D, Fresno V, Martínez R. Classifying trending topics. In: Proceedings of the 20th ACM international conference on Information and knowledge management - CIKM '11. .New York, New York, USA: ACM Press; 2011. p. 2461
8. Tanveer M, Rajani T, Rastogi R, Shao YH, Ganaie MA. Comprehensive review on twin support vector machines. Ann Oper Res. 2022 Mar 8.
9. Benhardus J, Kalita J. Streaming trend detection in Twitter [Internet]. Vol. 9, Int. J. Web Based Communities. 2013. Available from: http://Twitter.com.
10. Kanber B. Hands-on Machine Learning with JavaScript. First Edition. Packt Publishing; 2018.
11. Hughes J, Aycock S, Caines A, Buttery P, Hutchings A. Detecting Trending Terms in Cybersecurity Forum Discussions [Internet]. Online. 2020. Available from: https://hackforums.net

12. Vanitha CN, Malathy S, Dhanaraj RK, Nayyar A. Optimized pollard route deviation and route selection using Bayesian machine learning techniques in wireless sensor networks. Computer Networks. 2022 Oct;216:109228.
13. Asgari-Chenaghlu M, Nikzad-Khasmakhi N, Minaee S. Covid-Transformer: Detecting COVID-19 Trending Topics on Twitter Using Universal Sentence Encoder. 2020 Sep 8; Available from: http://arxiv.org/abs/2009.03947
14. Pervin N, Fang F, Datta A, Dutta K, Vandermeer D. Fast, scalable, and context-sensitive detection of trending topics in microblog post streams. ACM Trans Manag Inf Syst. 2013 Jan;3(4).
15. https://www.trendminer.com/. Advanced Analytics for the Process Manufacturing Industry. 2020.
16. Syarif S, Anwar, Dewiani. Trending topic prediction by optimizing K-nearest neighbor algorithm. In: 2017 4th International Conference on Computer Applications and Information Processing Technology (CAIPT). IEEE; 2017. p. 1–4.
17. Lin G, Lin A, Gu D. Using support vector regression and K-nearest neighbors for short-term traffic flow prediction based on maximal information coefficient. Inf Sci (N Y). 2022 Aug;608:517–31.
18. Hurtado J, Huang S, Zhu X. Topic Discovery and Future Trend Prediction Using Association Analysis and Ensemble Forecasting. In: Proceedings - 2015 IEEE 16th International Conference on Information Reuse and Integration, IRI 2015. Institute of Electrical and Electronics Engineers Inc.; 2015. p. 203–6.
19. Baudin A, Danisch M, Kirgizov S, Magnien C, Ghanem M. Clique Percolation Method: Memory Efficient Almost Exact Communities. In 2022. p. 113–27.
20. Halima Banu S, Chitrakala S. Trending Topic Analysis using novel sub topic detection model. In: 2016 2nd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB). IEEE; 2016. p. 157–61.
21. Blei DM, Lafferty JD. Dynamic topic models. In: Proceedings of the 23rd international conference on Machine learning - ICML '06. New York, New York, USA: ACM Press; 2006. p. 113–20.
22. Zosa E, Granroth-Wilding M. Multilingual Dynamic Topic Model. In: Proceedings - Natural Language Processing in a Deep Learning World. Incoma Ltd., Shoumen, Bulgaria; 2019. p. 1388–96.
23. Tijare PV, Rani P. J. Detecting Trending Event Topics Using Sentiment Driven Derivatives Method On Twitter. Indian Journal of Computer Science and Engineering. 2021 Aug 20;12(4):818–26.
24. Charnine M, Tishchenko A, Kochiev L. Visualization of Research Trending Topic Prediction: Intelligent Method for Data Analysis. In: Proceedings of the 31th International Conference on Computer Graphics and Vision Volume 2. Keldysh Institute of Applied Mathematics; 2021. p. 1028–37.

25. Ibrahim AA, L. R, M. M, O. R, A. G. Comparison of the CatBoost Classifier with other Machine Learning Methods. International Journal of Advanced Computer Science and Applications. 2020;11(11).
26. Daud A, Abbas F, Amjad T, Alshdadi AA, Alowibdi JS. Finding rising stars through hot topics detection. Future Generation Computer Systems. 2021 Feb;115:798–813.
27. Khan A, Gomez F, Gonzalez R, Diakoff H, Vera J, McCrae J. Towards the Construction of a WordNet for Old English. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association; 2022. p. 3934–41.
28. Sameer S, Behadili S. Data Mining Techniques for Iraqi Biochemical Dataset Analysis. Baghdad Science Journal. 2022 Apr 1;19(2).
29. Rabani ST, Khan QR, Khanday AMUD. Detection of Suicidal Ideation on Twitter using Machine Learning & Ensemble Approaches. Baghdad Science Journal. 2020 Dec 1;17(4):1328.
30. Wilkinson D, Thelwall M. Trending Twitter topics in English: An international comparison. Journal of the American Society for Information Science and Technology. 2012 .46–Aug;63(8):1631
31. de França FO, di Genova DVB, Penteado CLC, Kamienski CA. Understanding conflict origin and dynamics on Twitter: A real-time detection system. 2022 .Expert Syst Appl .212:118748;.Sep
32. Balakrishnan V, Ethel LY. Stemming and Lemmatization: A Comparison of Retrieval Performances. .7–Lecture Notes on Software Engineering. 2014;2(3):262
33. Vajjala S, Majumder B, Gupta A, Surana H. Practical Natural Language Processing. st ed. 1 .O'Reilly Media; 2020
34. Carneiro A, Matos MJ, Uriarte E, Santana L. Trending Topics on Coumarin and Its Derivatives in 2020. .Molecules. 2021 Jan 19;26(2):501
35. Singh AK, Shashi M. Vectorization of Text Documents for Identifying Unifiable News Articles. International Journal of Advanced Computer Science and Applications. 2019;10(7).

.