information

MDPI

Article

Analyzing Social Media Data using Sentiment Mining and Bi-gram Analysis for the Recommendation of YouTube Videos

Ken McGarry[1,]

[1] School of Computer Science, University of Sunderland, St Peters Campus, Sunderland SR6 ODD, UK. ken.mcgarry@sunderland.ac.uk

Abstract: In this work we combine sentiment analysis with graph theory to analyze user posts, likes/dislikes on a variety of social media to provide recommendations for YouTube videos. We focus on the topic of climate change/global warming which has caused much alarm and controversy over recent years. Our intention is to recommend informative YouTube videos to those seeking a balanced viewpoint of this area and the key arguments/issues. To this end we analyze Twitter data; Reddit comments and posts; user comments, view statistics and likes/dislikes of YouTube videos. The combination of sentiment analysis with raw statistics and linking users with their posts gives deeper insights into their needs and quest for quality information. Sentiment analysis provides the insights into user likes and dislikes, graph theory provides the linkage patterns and relationships between users, posts and sentiment.

Keywords: recommender systems; graph theory; sentiment analysis; Twitter; Reddit, YouTube

1. Introduction

Recommender systems (RS) are intended to provide the online user with advice, reviews and opinions from previous purchasers on products and services mainly through methods such as collaborative filtering (CF)[1]. The main RS objective using CF is to persuade users to buy items or services they have not previously bought/seen before based on the buying patterns of others. This can be achieved by ranking either the item-to-item similarity or the user-to-user similarity and then predicting the top scoring product that ought to appeal to the potential buyer. Unfortunately, CF has a number of limitations such as the *cold-start problem* i.e. generating reliable recommendations for those with few ratings or items. However, this issue can be alleviated to some extent by reusing pre-trained deep learning models and/or using contextual information [2]. Since CF is generally an open process, they can be vulnerable to biased information or fake information [3,4]. Fake user profiles can easily manipulate recommendation results by giving the highest rates to targeted items and rate other items similar to regular profiles. This behavior is called a "shilling attack" [5].

Initially launched in 2005, YouTube has seen an exponential growth of submitted videos and is the most popular platform for viewing material that informs, educates and entertains it's users. YouTube is a free video sharing service allowing users to view online videos and also for them to develop and upload their own materials to share with others [6,7]. However, for many YouTube contributors the opportunity to earn money from their channels popularity is a great incentive. To earn money from YouTube, a contributor must have 1,000 subscribers and at least 4,000 watch hours in the past year. Contributors can then apply to YouTube's Partner Program and monetize their channel. However, YouTube keeps careful surveillance on any mechanism that artificially inflates the number of comments, views or likes. Unscrupulous contributers often achieve increased rankings by using bots or automatic systems or even presenting videos to unsuspecting viewers.

The objective of our work is to demonstrate that a recommendation engine can be used to provide users with reliable YouTube videos based on initial keyword searches. The

Citation: McGarry, K.. *Information* **2023**, *1*, 0. https://doi.org/

Received:
Revised:
Accepted:
Published:

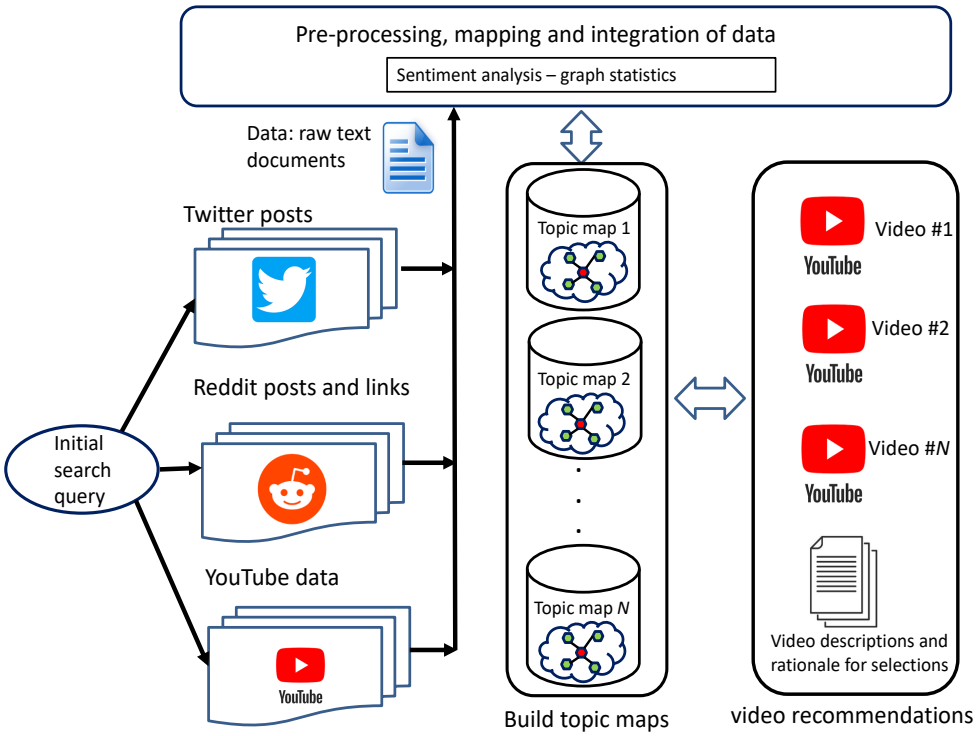


Figure 1. Overall system operation of data throughput and transformations

topic of interest is global warming/climate change but the system could be applied to any subject. The objectives are two-fold, once we can identify their sentiment/opinions on global warming we can provide users with authoritative videos with scientific credence based on their beliefs. Then, we can present users with authoritative videos representing the opposite stance. The intent is to balance out the debate with evidence they would perhaps not necessarily seek out. Out intention is not to change opinions but to help users become more aware of the issues.

To achieve these objectives, we combine sentiment analysis and graph theory to provide deeper insights into YouTube recommendations. Rather than use different software platforms, we combine several R library's into a unified system, making overall integration easier. The overall system workflow is shown in Fig 1. An initial search topic is defined and fed into the API's of the three platforms (Twitter, Reddit and YouTube). The resulting posts are preprocessed and parsed, the text data is then analysed by graph theoretic measures that provide statistical metrics of user posts and how they interact. The sentiments of user posts are used to create topic maps which reflect common themes and ideas these users have. Ratings of YouTube videos and provenance of their sources are estimated to provide some indication of their validity and integrity.

The main contribution of this work is threefold, first we integrate sentiment mining with graph theory providing statistical information on the posters and contributers, we also use up-votes and down-votes as a recommendation source, finally we create a logical structuring of the twitter, youtube and reddit data using topic maps. Topic modelling, is necessary since most topics of interest will comprise a mixture of words and sentiments which is a feature of human language. Therefore some overlapping of concepts will occur, so an unsupervised classification method is required. We use Latent Dirichlet allocation (LDA) which is commonly used for fitting a topic models [8].

The remainder of this paper is structured as follows: section two describes related work and recent advances in recommender systems, section three outlines the social media data used, we then describe in section four the computational methods used. Section five

presents the experimental results and the discussion and finally section six presents the conclusions and future work.

2. Related work

Here we discuss related work for recommender systems, sentiment analysis and graph theoretic methods.

2.1. Recommender systems

We can say that Recommender systems can be categorized into three main groups; such as content based recommender systems, collaborative recommender systems and hybrid recommender systems. One of the first and most predominant is the Amazon recommendation system which has undergone many refinements over the past 20 years [9]. The RS are generally trained from historical data and provide the customer with potentially useful feedback with products or services they may like. The details of the RS algorithm used by YouTube is unknown but it is generally believed to employ deep neural learning [10]. However, a recent study revealed it to contain biases and is a major source of misinformation on certain health related videos [11]. Another issue, which we do not tackle in this paper are the attacks on recommender systems to either down vote or up vote content [12].

Our system can be classed as a hybrid, similar work to ours include Kim and Shim who proposed a recommender system based on Latent Dirichlet Allocation (LDA) using probabilistic modelling for Twitter [13]. The top-K tweets for a user to read along with the top-K users that should be followed are identified based on LDA. The Expectation–Maximization (EM) algorithm was used to learn model parameters. Abolghasemi investigated the issues around human personality in decision-making as it is plays a role when individuals discuss to reach a group decision when deciding which movie to watch [14]. They devised a three-stage approach to decision making, they used binary matrix factorization methods in conjunction with an influence graph that includes assertiveness and cooperativeness as personality traits, they then applied opinion mining to reach a common goal. We use similar metrics to judge personalities based on tenor/tone of language used and their likes dislikes.

A similar approach was taken by Leng et al who were researching social influence and interest evolution for group recommendations [15]. The system they developed (DASIIE), is designed to dynamically aggregate social influence diffusion and interest evolution learning, they used Graph Neural Networks as the basis of their recommendation system. Th neural network approach allowed them to integrate the group members role weights and expertise weights enabling the decision-making process to be modeled simultaneously. Wu et al. have examined the technique of data fusion for increasing the efficiency of item recommender systems. It employed a hybrid linear combination model and used a collaborative tagging system [16].

2.2. Sentiment analysis

Over the past 10 years or so sentiment analysis has seen massive expansion both in practical applications and research theory [17–19]. The process of sentiment mining involves the preprocessing of text using either simple text analytics or the more complex NLP such as the Stanford system [20]. The text data can be organised by individual words or at the sentence and paragraph level by the positive or negative words it is comprised of [21]. Words are deemed to be either neutral, negative or positive based on the assessment of a lexicon [22,23]. Sentiment analysis is employed in many different areas from finance [24,25] to mining student feedback in educational domains [26,27]. It has also been used to automatically create ontologys from text [28]. Sentiment analysis has been used to examine the satisfaction within the computer gaming community, observing features in games they liked/disliked [29]. We have seen commercial applications for the automated mining of customer emails/feedback/reviews for improving satisfaction with products or services

that has seen the tremendous growth [30,31]. Twitter is often used as a source of data for sentiment mining on many topics [32], however it is with tweets collected over long time periods that tend to reveal interesting trends and patterns [33]. For example sentiment analysis has been applied to monitoring mental health issues based on Tweets [34].

Work by Kavitha is similar to ours as it considers YouTube user comments based on their relevance to the video content given by the description [35]. They build a classifier that analyses heavily liked and disliked videos, similarly we use counts to help rate the videos. They also consider spam and malicious content, we also filter out posts that contain sarcastic and profane content as they are unlikely to contain much cogent information. [11]. A more serious issue was considered by Abul-fottouh in the search for bias in YouTube vaccination videos. They discovered that pro-vaccine videos (64.75%) outnumbered anti-vaccine (19.98%) videos with perhaps 15.27% of videos being neutral in sentiment. It is unsurprising that YouTube tended to recommend neutral and pro-vaccine videos than anti-vaccine videos. This implies YouTube's recommender algorithm will recommend similar content to users with similar viewing habits and similar comments. This is related to the sentiment work of Alhabash who investigated cyber-bullying on YouTube, this involved examining comments, virality and arousal levels on civic behavioral patterns [36]. The findings concluded that people are more committed/interested in topics or comments that have negative sentiments, hence cyber-bulling videos appear to have disproportionate effect on users. Further work by Shiryaeva et al investigated the negative sentiment (anti-values) in YouTube videos, here the viewpoint was taken from the lens of linguistics to reveal grammar and style indicative of certain behaviors and intentions [37]. Although, the work was not automated the authors were able to identify 12 anti-values that were characteristic of bad behavior.

2.3. Graph Theory

This area of computer science uses statistical measures to gather information about the connectivity patterns between the nodes (which can be people, objects or communications) which can reveal useful insights into the dynamics, structure and relationships that may exist [38,39]. Numerous areas have benefited from graph theory such as computational biology and especially social media which has received a great deal of attention from researchers [40]. The most notorious incident was the FaceBook/Cambridge Analytica scandal which involved the misuse of personal data [41]. However, this particular case served to highlight the power of machine learning and interconnected data to influence individuals. In social media analysis, individuals are connected to friends, colleagues, political, financial and personal web interests all of which can analyzed by organizations to improve services, products or detect trends and opinions [42].

Graph theory was used by Cai to examine the in-degree of posters, the intention was to identify if Schilling attacks were occurring in user posts [43]. Each user was assigned a "suspicion" rating based on their in-degree and their behavior characteristics such as diversity of interests, long-term memory of interest, and memory of rating preference). The graph information was fed to a density clustering method and malicious users were generally identified. A similar approach was taken by Cruickshank to use a combination of graph theory and clustering on Twitter hash-tags [44]. The method investigated the application of multiple different data types that can be used to describe how users interact with hashtags on the COVID-19 Twitter debate. They discovered that certain topical clusters of hashtags shifted over the course of the pandemic, while others were more persistent. The same effect (homophilly) likely to be true of climate change debate, for example the HarVis system of Ahmed uses graph theory to untangle frequent from infrequent posters to assist a better understanding of the authors/posters ranking [45]. This is an important point as it is best to weigh authoritative heads instead of just counting them.

The use of graph-like structures such as Graph Convolutional Networks (GCN) is becoming more popular, this approach has the flexibility and power to model many social media problems. These are more powerful than standard graph theoretic methods but

come with a computational burden and requirement for more data. The use of GCN is also receiving attention for identifying Schilling attacks in recommender systems [46]. Another issue is the informal language used in posts and other characteristics of this type of data, for example, Keramatfor et al understood that short posts such as Tweets have dependencies upon previous posts [47,48]. To model Tweet dependencies requires the combination of data such as textual similarity, hashtag usage, sentiment similarity and friends in common.

In table 1 we provide a short qualitative comparison with the most similar recommendation systems to ours. The difference is that our system uses a greater variety of social media data and uses profiling and a wider variety of computational methods

Author	System Name	Date	Methods	Social Media
McGarry			Graph theory, sentiment analysis, bigrams, profiling	Twitter, YouTube, reddit
Keramatfar [47]	MHLSTM	2021	LSTM, profiling, sentiment analysis	Twitter
Cruickshank [44]	MVMC	2020	Hash-tags, sentiment analysis	Twitter
Ahmad[45]	HarVis	2017	Graph theory	YouTube
Kavitha[35]		2020	Bag of Words, NLP	YouTube
Kim[13]	TWLITE	2014	LDA, probability	Twitter
Nilashi [49]		2023	LDA, EM, clustering	TripAdvisor

Table 1. Qualitative comparison with other systems

3. Data

Here we describe our data sources, how they are pre-processed and integrated prior to building machine learning models and implementing the recommendation system. Twitter, Reddit and Youtube posts are searched based on climate change keywords, then downloaded using the appropriate APIs, the posts are cleaned of stopwords, stemming, punctuation and emojis. A separate corpus, consisting of term-document-matrix is created for each data source. We then build topic maps for each corpus, the optimum number is generated from a range of 10-100 potential topics. The most optimum number is selected by calculating the harmonic mean for each number. We did not analyze the social media data to determine if any content was generated by bots. The social media companies are well aware of the issues and have developed bot detection software [50,51], for a comprehensive recent survey see Hayawi et al [52].

In table 2 the sources of the data are presented, showing the number of the records, the approximate date of collection and where collected from.

Data	Source	Date	No Records
Twitter	API	Jan 2020 to Mar 2020	2K
Twitter	Kaggle	Apr 2015 to Feb 2018	44K
Reddit	API	Dec 2022 to Feb 2023	100K
YouTube	API	Dec 2022 to Feb 2023	26K

Table 2. Data sources, number of records and approximate date of collection

3.1. Reddit Data

Reddit is a social news aggregation platform and discussion forum, users can post comments, web links, images, and videos. Other users can up/down vote these posts and engage in dialog, the site is well known for it's open and diverse nature. User posts are organized by subject into specific boards called *communities* or *subreddits*. The communities are moderated by volunteers who set and enforce rules specific to a given community, they can remove posts and comments that are offensive or that break the rules, they also keep discussions on subject topic [53–55]. Reddit is becoming very popular as statistics show from the SemRush web traffic system which estimates Reddit to be the 6th most visited site in the USA [56]. We text mine Reddit for posts and sentiment pertaining to the issues surrounding the climate change debate [57–61]. The reddit data was collected between December 2022 and February 2023, the reddit API limited extraction with rate limits, we used the R interface (RedditExtractoR) [62].

The reddit data consists of two structures, the comments and the threads. The comments data consists of the following variables: url, author, date, timestamp, score, upvotes,

downvotes, golds, comment and the comment-id. The threads data has further information pertaining to other users actions on the posts such as total-awards-received, golds, cross-posts, and other user comments. Fig 2 shows a list of the types of data residing in the posts.

	url	author	date
1	https://www.reddit.com/r/change/comments/3d9zzr/texas_state_commission_on_judicial_conduct_remove/	Stinger267	2015-07-14
2	https://www.reddit.com/r/change/comments/3csq4m/petition_for_the_city_council_of_killeen_tx/	Justin-Robert	2015-07-10
3	https://www.reddit.com/r/change/comments/3b9h3d/4_language_tips_to_positively_change_your_psyche/	tadziowv	2015-06-27
4	https://www.reddit.com/r/change/comments/39yuvx/petition_remove_ellen_pao_from_her_position_as/	1shomofo	2015-06-15
5	https://www.reddit.com/r/change/comments/34v0yh/gta_5_modding_dont_worry_we_can_change_it/	Dr_Slenderman	2015-05-04
6	https://www.reddit.com/r/change/comments/15pqy7/we_need_a_change_anyone_agree/	konradturin	2013-01-16

	timestamp	score	upvotes	downvotes	golds
1	1436895895	1	1	0	0
2	1436551458	1	1	0	0
3	1435370209	1	1	0	0
4	1434406721	0	0	0	0
5	1430770254	1	1	0	0
6	1358310115	2	2	0	0

Figure 2. Preprocessed Reddit data structure

3.2. YouTube Data

Youtube provides data pertaining to users opinions on the various videos they find of interest. Fig 3 displays a data structure showing the first ten records of Youtube data. The *Comment* column is the user post, other columns identify the user ID, authors image profile URL, author channel ID, author channel URL, Reply count, Like count, post published date, when post updated, post parent ID, post ID and Video ID. The Youtube data was collected December 2022 and February 2023 using the R vosonSML package [63].

	Comment	Autho~1	Autho~2	Autho~3	Autho~4	Reply~5	LikeC~6	Publi~7	Updat~8	Comme~9	Paren~*	VideoID
	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>
1	"Damage done, I'm affraid"	febbra2	https:~	http:/~	UCZiJ-~	0	1362	2013-0~	2013-0~	UgyeaS~	NA	oJAbAT~
2	"The way I was predicting so far, about the~	Levon ~	https:~	http:/~	UCRaam~	0	162	2013-0~	2013-0~	Ugxp-A~	NA	oJAbAT~
3	"I dare to any body who call them self sci~	Levon ~	https:~	http:/~	UCRaam~	0	71	2013-0~	2013-0~	UgwIzs~	NA	oJAbAT~
4	"Whatever it is, I'm not fucking paying for~	justfi~	https:~	http:/~	UCZWpY~	0	102	2012-0~	2012-0~	UgxXdk~	NA	oJAbAT~
5	"I just can't endorse it."	True M~	https:~	http:/~	UCEeU8~	0	60	2012-0~	2012-0~	UgylvJ~	NA	oJAbAT~
6	"Please have a look at the Youtube Video of~	Tokyor~	https:~	http:/~	UCPgWW~	0	19	2011-0~	2011-0~	Ugwaxr~	NA	oJAbAT~
7	"anyone who thinks its a hoax its to be sla~	kevin ~	https:~	http:/~	UC-c2Y~	0	555	2010-0~	2010-0~	Ugwmr8~	NA	oJAbAT~
8	"i cant believe theres still nim rods out t~	kevin ~	https:~	http:/~	UC-c2Y~	0	222	2010-0~	2010-0~	Ugw79m~	NA	oJAbAT~
9	"@me\nMaybe a requirement for equal spendin~	bflatk~	https:~	http:/~	UCPwCw~	0	7	2010-0~	2010-0~	Ugydez~	NA	oJAbAT~
10	"I'm not so sure the fossilsaurs need to w~	bflatk~	https:~	http:/~	UCPwCw~	0	8	2010-0~	2010-0~	UgxiRU~	NA	oJAbAT~

Figure 3. Preprocessed Youtube data structure

3.3. Twitter Data

The data downloaded and preprocessed from Twitter consisted of two sources. Twitter was more problematic as difficulties encountered recently with the API access. The first set of data was collected by API by the authors in 2020 and then from a datasource in the public domain from Kaggle, this consisted of 44,000 tweets collected by the University of Waterloo in 2015-2018 [64]. Twitter contains variables similar to Reddit and Youtube they describe the UserScreenName, the UserName, the Timestamp, Text, Embedded-text, Emojis, Comments, Like counts, number of Retweets, Image.links and the Tweet.URL. This is displayed in fig 4.

4. Methods

In fig 5 the basic flow of social media searching, storage of data and flow information is presented. This example is for global warming/climate change but the process would be similar for any topic of interest e.g. war in Ukraine or Covid pandemic. Keywords are selected and used to search the social media, YouTube videos are downloaded and observed for validity. The data is saved in RDATA format (R programming language

UserScreenName	UserName	Timestamp	Embedded_text	Emojis	Comment	Likes	Retweets	Image.link	Tweet.URL
user_screen1	user1	2022-01-17T2:	The only solution Iâ€™ve ever heard the Left		1,683	2,259	11.7K	[]	https://twitter.com/laurenboebert/status/1483220748487569409
user_screen2	user2	2022-01-17T2:	Climate change doesnâ€™t cause volcanic eru		158	64	762	[]	https://twitter.com/catherine__c/status/1483211036463603713
user_screen3	user3	2022-01-17T2:	Vaccinated tennis ball boy collapses in the ten		24	118	159	['https://pl	https://twitter.com/KaConfessor/status/1483225542824505347
user_screen4	user4	2022-01-17T2:	North America has experienced an average w		15	50	158	[]	https://twitter.com/climate_parent/status/1483192925152587777
user_screen5	user5	2022-01-17T2:	They're gonna do the same with Cli	ðŸ…¾	4	24	127	['https://pl	https://twitter.com/Thomas_Sp8/status/1483185023066902528
user_screen6	user6	2022-01-17T2:	HELLO AMERICA,Who would have ever thoug		1	12	22	[]	https://twitter.com/ruggiere_l/status/1483192848086233089

Figure 4. Preprocessed Twitter data structure

format), data structures are formed from sentiment, graph analysis of bi-grams and user profiles consisting of likes/dislikes and overall estimated stance on global warming. Data are split for training 90% and 10% for test. The recommendation engine is constructed and tested and compared with other methods.

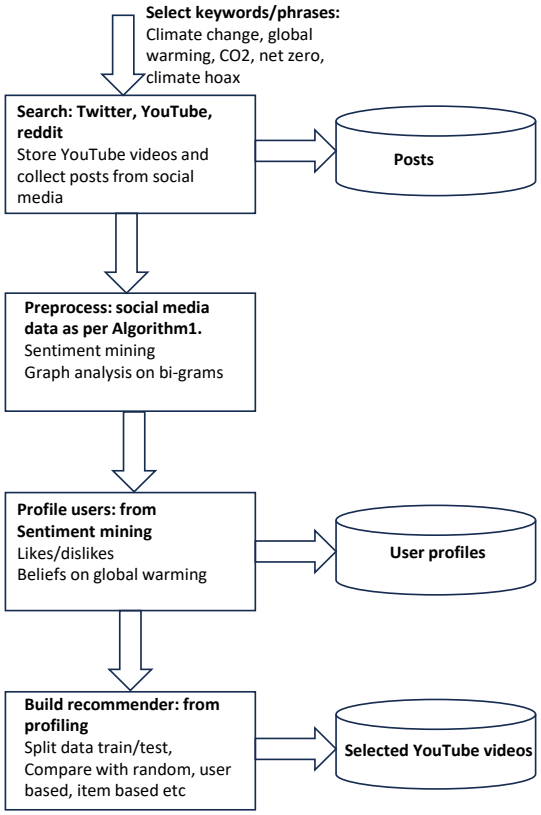


Figure 5. Data collection, storage and processing

The three data sets from Twitter, YouTube and reddit now must be preprocessed prior to sentiment analysis. In algorithm 1 we show the stages of processing the three data sources (Twitter T_{txt}, Reddit R_{txt} and Y_{txt} YouTube). In lines 1 to 4 each text is converted into Corpus and in line 5 they under go removal of stop words, stemming, and removal of punctuation and non-ascii text. Lines 6 to 11 creates the topic maps for each Corpus using a for..loop to build a series of topic maps from 10 to 100 maps. Lines 12 and 13 uses the harmonic mean metric to judge the optimum number of maps for each Corpus. Finally line 14 returns the optimum topic maps and related data structures.

4.1. Sentiment mining

We use the sentimentR package written by Rinker [65], it incorporates the lexicon developed by Ding *et al* [23]. The lexicon consists of words which have been rated as neutral,

Algorithm 1 Data transformation for text mining

Input: Raw text for twitter $\mathbf{T_{txt}}$, reddit $\mathbf{R_{txt}}$, youtube: $\mathbf{Y_{txt}}$;
Output: Corpus for twitter $\mathbf{C}_t$, reddit $\mathbf{C}_r$, youtube $\mathbf{C}_y$;
Topic Maps for each Corpus$\mathbf{TM_t}$; $\mathbf{TM_r}$; $\mathbf{TM_y}$;
optimum number of topic maps $\mathbf{OT_t}$; $\mathbf{OT_r}$; $\mathbf{OT_y}$

Initialize MinWordFreq $\leftarrow 5$
Create corpus $\mathbf{C_t} \leftarrow \mathbf{T_{txt}}$
Create corpus $\mathbf{C_r} \leftarrow \mathbf{R_{txt}}$
Create corpus $\mathbf{C_y} \leftarrow \mathbf{Y_{txt}}$
Preprocess $\mathbf{C_t}$, $\mathbf{C_r}$, $\mathbf{C_y} \leftarrow$ **removal**[stopwords, stemming, punctuation, non-ascii]
repeat
  if (words >= MinWordFreq) **then**
    Build TM in $\mathbf{C} = \forall \mathbf{C}$.
    TM $\forall$ C
  end if
until $\mathbf{TM_t}$; $\mathbf{TM_r}$; $\mathbf{TM_y}$; populated from [10...100]
Calculate harmonic mean in $\mathbf{C} = \forall \mathbf{C}$. for all Corpora
$\mathbf{OT_t}$; $\mathbf{OT_r}$; $\mathbf{OT_y} \leftarrow P(w|z)$
Return [$\mathbf{C_t}$, $\mathbf{C_r}$, $\mathbf{C_y}$, $\mathbf{TM_t}$, $\mathbf{TM_r}$, $\mathbf{TM_y}$, $\mathbf{OT_t}$, $\mathbf{OT_r}$, $\mathbf{OT_y}$]

positive or negative sentiment and have a strength value allocated. The implementation takes into account valence shifters whereby word polarity can be negated, amplified or de-amplified, if ignored sentiment analysis can be less effective and miss the true intent of the author of the posts. The package can easily update a dictionary by adding new words or changing the value of existing words. However, pre-processing of the text is achieved by the tm (text mining) package developed by Feinerer [66,67], this package enables the removal of stop words, stemming and non-ascii character removal.

A paragraph is composed of sentences $p_i = \{s_1, s_2..., s_n\}$), and each sentence can be decomposed into words $s_j = \{w_1, w_2, ..., w_n\}$). The words in each sentence (w_i, j, k) are searched and compared against the dictionary or lexicon [45,68,69]. Sentiment is assessed by various calculations, where N and P are counts of the negative and positive words, O is the count of all words including neutral words [70,71].

P = $(w_i, j, k+)$ and N = $(w_i, j, k-)$, the words are tagged with either +1 or -1, neural words are zero.

$$Sentiment = (P - N)/(P + N + O) \tag{1}$$

4.2. Graph modelling

The igraph package developed by Csardi and Nepusz provides a comprehensive package for conducting analysis into graph theory, it is available across several languages and is regularly updated and maintained [72]. It allows statistics to be computed from the graph network based on the nodes and connectivity patterns. Useful statistics include closeness, betweenness, and hubness amongst others. Furthermore, it is possible to detect community structure where certain nodes strongly interact and form cohesive clusters which may relate to some real-world characteristics about the network. Graph theoretic methods can be applied to any discipline where the entities of interest are linked together through various associations or relationships. Other graph approaches, different to ours involve graph neural networks (GNNS) which are a powerful way of expressing graph data [73].

Hub nodes have many connections to other nodes and therefore of some importance or influence, the deletion of a hub node is more likely to be catastrophic than deletion of a non-hub node. This is a characteristic confirmed in many real-world networks which are typically small world networks with power law degree (number of edges per vertices) distributions [38].

The concept of the *shortest path* is important to centrality measures and can be defined as when two vertices i and j are connected if there exists a sequence of edges that connect i and j. The length of a path is its number of edges. The distance $l(i, q)$ between i and j is the length of the shortest path connecting i and j [39]. The closeness centrality of a given node i in a network is given by the following expression:

$$CC(v_i) = \frac{N-1}{\sum_j d(v_i, v_j)} \tag{2}$$

Betweenness centrality is a measure of the degree of influence a given node has in facilitating communication between other node pairs and is defined as the fraction of shortest paths going through a given node. If $p(v_i, v_j)$ is the number of shortest paths from node i to node j, and $p(v_i, v_k, v_j)$ is the number of these shortest paths that pass through node k in the network, then the BC of node k is given by:

$$BC(v_k) = \sum_i \sum_j \frac{p(v_i, v_j, v_k)}{p(v_i, v_j)}, i \neq j \neq k \tag{3}$$

4.3. Generating the topic models

Latent Dirichlet Allocation (LDA) is commonly used to generate topic models [74]. We use the R Topic model package developed by Grun and Hornik [8,75]. Equation 4 defines the stages, there are three product sums K, M, N that describe the documents, topics and terms.

$$P(W, Z, \theta, \varphi, \alpha, \beta) = \prod_{i=1}^{K} P(\varphi_i; \beta) \prod_{j=1}^{M} P(\theta_j; \alpha) \prod_{t=1}^{N} P(Z_{j,t}\theta_j P(W_{j,t}|\varphi Z_{j,t}) \tag{4}$$

Where: $P(W, Z, \theta, \varphi, \alpha, \beta)$ is the overall probability of the LDA model; $\prod_{i=1}^{K} P(\varphi_i; \beta)$ generates the Dirichlet distribution of the topics over the *terms*; while $\prod_{j=1}^{M} P(\theta_j; \alpha)$ calculates the Dirichlet distribution of the documents over the *topics*; the probability of a topic appearing in a given *document* is given by $\prod_{t=1}^{N} P(Z_{j,t}\theta_j)$; while the probability of a word appearing in a given topic is calculated by $P(W_{j,t}|\varphi Z_{j,t})$. The parameters W, Z, θ, φ where θ and φ hold the document-term matrices; while α, β are the Dirichlet distribution parameters; the indices i, j, t keep track of the number of topics, terms and documents. The term W is the probability that a given word appears in a topic and Z is the probability that a given topic appears in the document [74].

We generate individual topic models for Twitter data, Reddit data and Youtube data. The optimum number of topics k is determined using a harmonic mean method determined by Griffiths and Steyvers [76,77]. This is shown in equation 5.

$$P(w|z) = \frac{\Gamma(V\beta)}{\Gamma(\beta)^v}^K, \prod_{i=1}^{K} \frac{\prod_V \Gamma(n_k^{(w)})}{\Gamma(n_k^{(\cdot)} + V\beta} \tag{5}$$

Where: w represents the words in the corpus w, and the model is specified by the number of topics K. Gibbs sampling provides the value of $p(w|z, K)$. $p(w|K)$ by taking the harmonic mean of a set of values of $p(w|z, K)$ when z is sampled from the posterior $p(z|w, K)$. Where n_k^w is the frequency of word w has been assigned to topic k in the vector z and Γ is the standard Gamma function.

4.4. Recommendation System

The last component in our system is the RS engine, this contains the information from the sentiment analysis, the statistics from user ratings and user connectivity patterns from graph analysis. We use nonnegative matrix factorization (NMF) to generate the process of collaborative filtering (CF) [78,79]. Strictly speaking NMF is related to Principal Components Analysis (PCA) which is typically used for dimensionality reduction but still keeps a meaningful representation of the solution [80]. Both methods use similar matrix transforms that are linear combinations of the other variables but NMF has a stricter constraint that the values should not be negative. This is an advantage because it enables a clearer *interpretation* of the factors involved since in many applications negative values would be counterintuitive such as negative website visits or negative human height.

There are also improvements in sparsity for feature detection and imputation of missing information. We integrate our recommender system without the framework of the R package by Hahsler, this allows easier testing and comparison [81].

The objective is to determine a matrix of ratings of **V**, where the columns represent the users and the rows represent the video ratings. The use of NMF will approximate this matrix by taking the matrix of users **W** and the matrix of videos **H**. The majority of **V** entries are unknown, these can be predicted by NMF using **W** x **H** $\approx$ **V**. See fig 6, the matrix dimensions of n and m are determined by **V**.

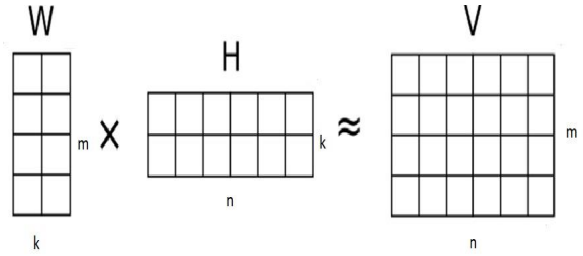


Figure 6. NMF matrix transformations where the dimensions: n and m are determined by the shape of **V** and k is determined by the number of components set by the user.

We use nonnegative matrices **W** and **H**, of rank k from which **V** is approximated by the dot product operator. Where k is a parameter set usually smaller than the number of rows and columns of **V**. The trade-off with k is a fine balance able to capture the key features of the data but to avoid overfitting. In previous work we modified NMF as a data integration method [82] other variations are typically used for data integration with heterogeneous data, especially in chemistry and community detection [83,84].

5. Results

The flow of data and processing the posts begins with the conversion of raw text from Twitter, Reddit and Youtube into Corpora, basically term-document-matrices (as described in algorithm 1). Once they are processed we can extract topic models from the Corpora to aid our understanding of the posts by this logical grouping of keywords.

As an example of sentiment analysis using the R package *sentimentr* on twitter posts is shown in fig 7. The posts are identified by number (1-10), they are ranked as either positive(green), neutral (grey) or negative(red), each with a number denoting the strength of the sentiment. There are three word sentiment lookups available for the Bing, NRC, and Afinn dictionary's, each with differing number of words rated and with differing sentiment values attached to each word. This can be at the word level, sentence level or the entire post (paragraph). As can be seen, the twitter data shown here represents a number of opinions on the climate change debate.

We can see that comment 1 is rated at zero sentiment since the sentence is fairly neutral in its wording. In comment 2 we find the first sentence is neutral but the second sentence has a positive sentiment word (optimistic) and is rated $+.082$. Comment 3 is more negative because of the words *scam*, *scammer* and *hoax*, rated at -147.

The next stage is to develop topic models holding keywords that are coherently related to key concepts and will be data mined for bigrams. The optimum number of topic models for each Corpora is determined using the Harmonic mean described in equation 5. In fig 8 the optimum numbers are presented with 24 for Twitter, 44 for Reddit and 31 for YouTube concepts and issues. We used LDA to generate the topicmaps with a value starting at 10 up to 100 possible topic maps, so at the first iteration 10 topic maps would be selected to describe the Corpora, then 11, 12, 13 until 100 topic maps are generated. Beyond a certain point adding more topic maps simply degrades performance, and when the harmonic

1: .000

so, why are we at 424 ppm of co2 not to mention the other greenhouse gases

2: +.082

Life, uh, finds a way. First optimistic news about climate I've heard in quite a while

3: -.147

Ahh Climate Change the biggest money making scam on the planet where Al Gore and the other Climate Scammers make billions a year with their Climate hoax. The Earth is a lot cooler than it has for a long time and it's going to keep getting cooler.

4: +.357

We may have a tough time in most industrialized countries - with our economies & tech over the 20th century well-established "bad habits" - but, developing nations & growing economies the world over don't have to do it the way we did then in order to become "wealthy" & modern in the 21st century. They have the opportunity to grow & advance in a new way with sustainable resources, economies, etc. while the rest of the industrialized world gets their sh*% together. They could be building permaculture food & water management resources, solar power grids, etc. from the start.

5: +.035

Anyone there know what Scenario 1 of Limits To Growth looks? It had many names, ie, business as usual, the Do Nothing Scenario, among others. At the time it was published Scenario I's Tipping Point would be caused by the creators of NOVEL and Forever chemical compounds failure to invent and have under development by 1975 an effective detoxification response to those chemical compounds already loosed into our planet's atmosphere. Until recently nothing had been invented and developed with that capability - remove all God-made (CO, Methane, + the other GHG) and Man-made chemical compounds, ie, PFAS, fertilizers, leaking Shale Oil and Coal cess pools, most river deltas, and tide pools. COOKED {by exceeding 350 ppm of atmospheric CO2 or CONTAMINATED (by PFAS and the other 300,000 man-made NOVEL and probably Forever chemical compounds accumulating in our spaceship's biosphere Life Support Systems, and us apex consumers. "Work the problem; Failure Is not an option," or is it? DG Sooner than later; no escape pods capable of getting us to the nearest probable planet to invade. So, the Tipping Point for Global Population Overshoot and Collapse about 2025 was confirmed ten or fifteen years ago.

6: .000

who is we?

7: .000

what about micro plastic that we swallow from fish

8: -.188

If we don`t stop breeding like rabbits and decrease the worlds population we are screwed, no matter what the data tells us.

9: .000

We have made tremendous progresses on making the climate change.

10: .000

RCP 1.9!

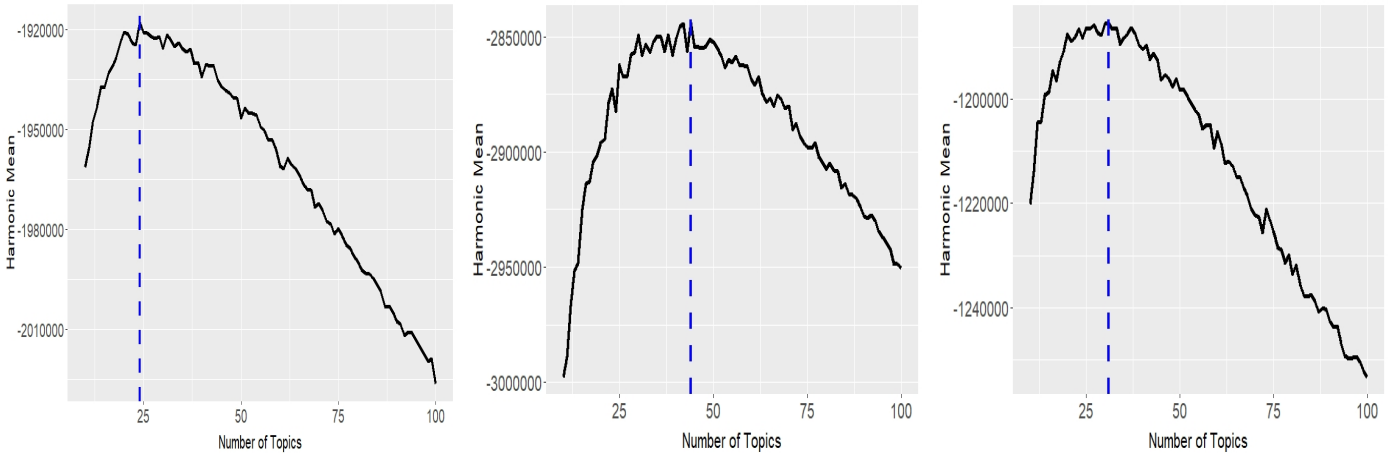
Figure 7. Basic sentiment mining using twitter data on 10 raw text posts.

mean decreases that is the number of maps to use. However, the Harmonic mean method has known instabilities but is generally robust enough.

In fig 9 five out of 24 twitter topicmaps are shown, generally the terms *climate* and *change* are present throughout some of the 24 topicmaps. Topicmap 2 is generally related to energy consumption of fossil fuels such as oil and gas. Topicmap 3 is concerned with public health and net zero. Topicmap 4 has gathered words on environmental impact and statements issued by the Intergovernmental Panel on Climate Change (IPCC). Topicmap 5 seems to have grouped human rights and social justice as key themes.

To augment the statistics and text mining, we also generated Wordclouds which perhaps give a better visualization and understanding of the main themes that dominate user posts. Individual word frequencies are used to highlight the important themes. The more frequent a word then its size increases. In fig 10 wordclouds for twitter, reddit and youtube are presented. Clearly *climate* and *change* totally dominate user posts for twitter, while reddit and youtube have a wider range of concepts with more or less equal frequency of occurrence. Only words that appear with at least five occurrences are displayed.

The next stage is to build graph theoretic models of bi-grams of co-occurring words building of up a picture of sentiment relating to each Youtube video. Graph models of Twitter and Reddit are also constructed to support the ratings/rankings of the videos in terms of the esteem/trust in which the videos producers are held. In table 3 the graph statistics for YouTube are shown for five users, the key variables are *Betweenness* and



(a) Twitter posts generate 24 topic maps.

(b) Reddit posts generate 44 topic maps

(c) Youtube posts generate 31 topic maps

Figure 8. Optimum topic map configuration selected from a range between 10-100

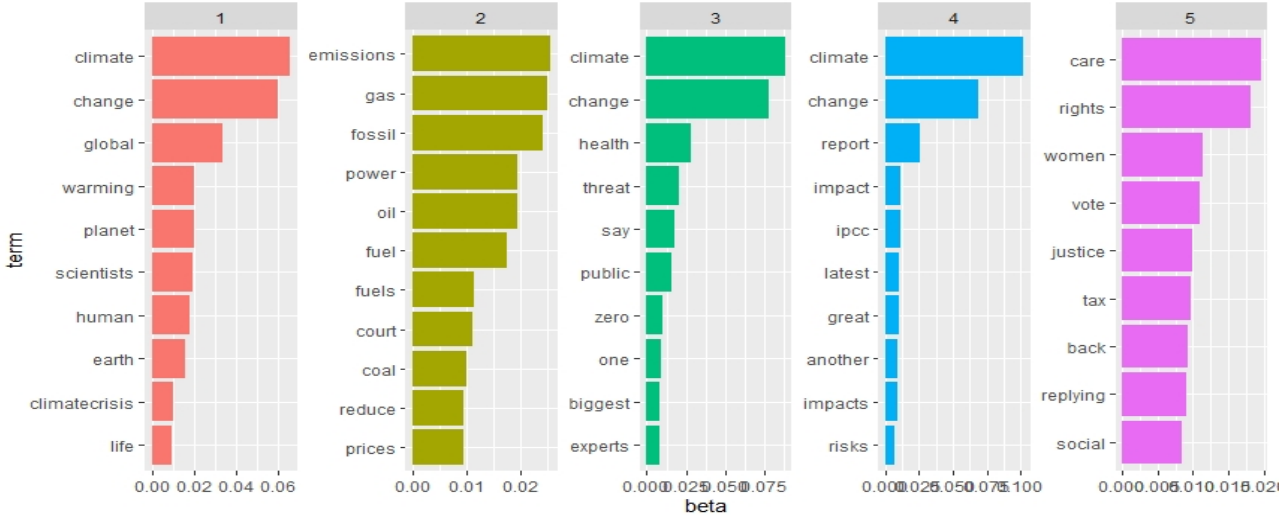


Figure 9. The first five topicmaps for twitter.

Hubness which indicates for each word the relative connectivity importance. The other columns have identical values - mod (modularity) column refers to the structure of the graph and can take a range of 0.0 to 1.0 indicating there is structure and not a random collection of connections between the nodes. *Nedges* indicates the number of connections in this small network, *nverts* is the number of nodes in the network. The *transit* column refers to the transitivity or community strength, it is a probability for the network to have adjacent nodes interconnected.

As the graph is highly disconnected (bigrams linking to other bigrams) it has zero for all entries. *Degree* refers to the average number of connection per node and of course is around 2.0, *diam* the length of the shortest path between the most distanced nodes. *Connect* refers to fully connectedness of the graph and in this case it is not. *Closeness* of a node measures its average distance to all other nodes, high closeness scores suggest a short distances to all other nodes. *Betweenness* detects the influence a given node has over the flow of information in a graph. The *Density* represents the ratio between the edges present in a graph and the maximum number of edges that the graph can contain. The *Hubness* is a value to indicate those nodes with larger number of connections than an average node.

In table 4 we have shown the basic statistics of several YouTube videos. We collect data such as the ID of the video e.g. in the first row, *oJAbATJCugs* would normally be used to se-

(a) Twitter wordcloud. **(b)** Reddit wordcloud **(c)** Youtube wordcloud

Figure 10. Word clouds for the posts

user	mod	path	nedges	nverts	transit	degree	diam	connect	close	between	density	hubness
UC9Di-3Y41sreUEtKD9MuZEQ	0.69	1.06	128.00	109	0.00	2.35	2.00	FALSE	0.30	19.00	0.01	0.00
UCjjVjhAEzLNAvpr-7pDpd8g	0.69	1.06	128.00	109	0.00	2.35	2.00	FALSE	0.20	15.00	0.01	0.00
UCrsD7Oq3yjZu0GYLhquHpVQ	0.69	1.06	128.00	109	0.00	2.35	2.00	FALSE	0.20	14.00	0.01	0.00
UCUB6baFW4kvLsLzlZ-kp5Ug	0.69	1.06	128.00	109	0.00	2.35	2.00	FALSE	0.50	5.00	0.01	0.00
UCjrCf7x7Dgo4VKplUgWKIdg	0.69	1.06	128.00	109	0.00	2.35	2.00	FALSE	0.50	5.00	0.01	0.00

Table 3. Graph theoretic statistics on YouTube bi-graph/bigrams on five users

lect the video in a web browser using the string *"https://www.youtube.com/watch?v=oJAbATJCugs"*. The number of comments received for each video is collected, along with the average number of *likes*, we also collect the number comments that had *zero likes* and we should note that this does not imply the video was disliked only that the person sending the comment neglected to select *like* irrespective of their feelings for the video. The number of unique posters making comments is also recorded. Next we perform sentiment analysis, examining overall sentiment for the video and then breaking down comments into neutral, negative and positive sentiments. The total number of sentiments (positive, negative and neutral) are based on the sentence level and therefore we have more than the overall number of comments i.e. number of posts.

	vid_id	title	num_comments	likes	zero_likes	num_posters	overall_sentiment	neg_sent_count	pos_sent_count	neut_sent_count
1	oJAbATJCugs	VID1	500	3311	412	57	7.81	501	581	437
2	n-Z0eG1pKhA	VID2	602	1947	364	525	-5.64	528	514	427
3	vFDnknU0h0s	VID3	628	219	536	477	16.88	912	999	669
4	2CQvBGSiDvw	VID4	1114	3399	789	587	0.7	1038	1103	1004
5	ga-RBuhcJ7w	VID5	689	158	627	468	-15.71	911	873	756
6	eDWq7-eP5sE	VID6	680	1185	428	481	30.13	943	1085	714
7	DticpNH3a2Q	VID7	587	72	536	458	-90.32	642	457	371
8	rwdxffEzQ9I	VID8	708	877	561	502	63.59	549	701	565
9	uynhvHZUOOo	VID9	769	300	616	520	-17.37	636	674	651
10	dcBXmj1nMTQ	VID10	625	155	529	488	45.01	594	732	502
11	tMwFNMfjFuU	VID11	98	1289	3	95	14	40	76	43
12	48zAWYkrBIw	VID12	305	1238	162	201	-9.6	288	259	252
13	eDWq7-eP5sE	VID13	679	1185	427	481	29.54	942	1082	713
14	DYWrehjaMFQ	VID14	737	303	585	424	33.96	860	964	565
15	I2OHAuvoUkQ	VID15	374	575	269	196	10.42	367	411	314
16	rweblFwt-BM	VID16	731	1103	512	522	35.25	750	881	659
17	pl1Rnz4zNkg	VID17	628	772	446	534	-35.53	515	454	362
18	qXLqoFHGmv0	VID18	763	1283	503	335	-17.99	811	740	708
19	m3hHi4sylxE	VID19	653	155	536	481	-48.9	773	710	760

Table 4. Main statistics on selected YouTube videos

In fig 11, the YouTube bigrams are displayed, we only show those words that have at least 100 co-occurrences based on key topic map groupings. The bigrams for YouTube are more strongly linked to coherent topics and follow a logical pattern of subjects with

more linkages between bigrams. Generally, the comments on YouTube are more calm and balanced with some thought given to the subject of global warming.

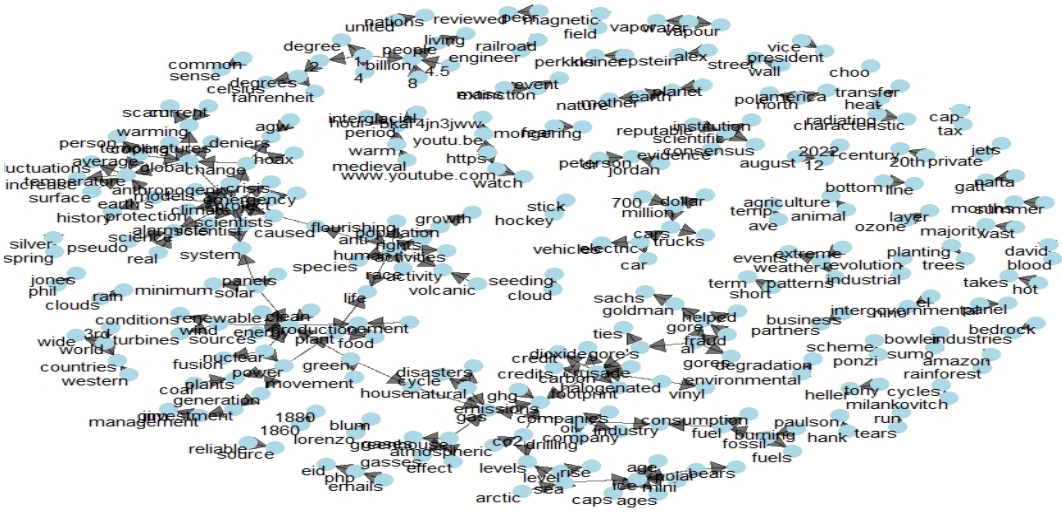


Figure 11. Bi-gram chart of YouTube linked pairs of words.

In fig 12, the situation for Twitter bigrams is a more complicated, furthermore because of the large number of posts we filter the number of word co-occurrences to 200 before they can appear on the plot. However, it provides a richer source of data illuminating the issues and concerns once the most frequently occurring words are revealed based on key topic map groupings. The general trend for twitter posts seems to contain a lot of off-topic issues such as legal aspects and gun violence. Another issue is the text limits on tweets (280 characters) which may cause posters some constraints in their dialog. The text limit has been raised for fee paying subscribers to 4,000 characters.

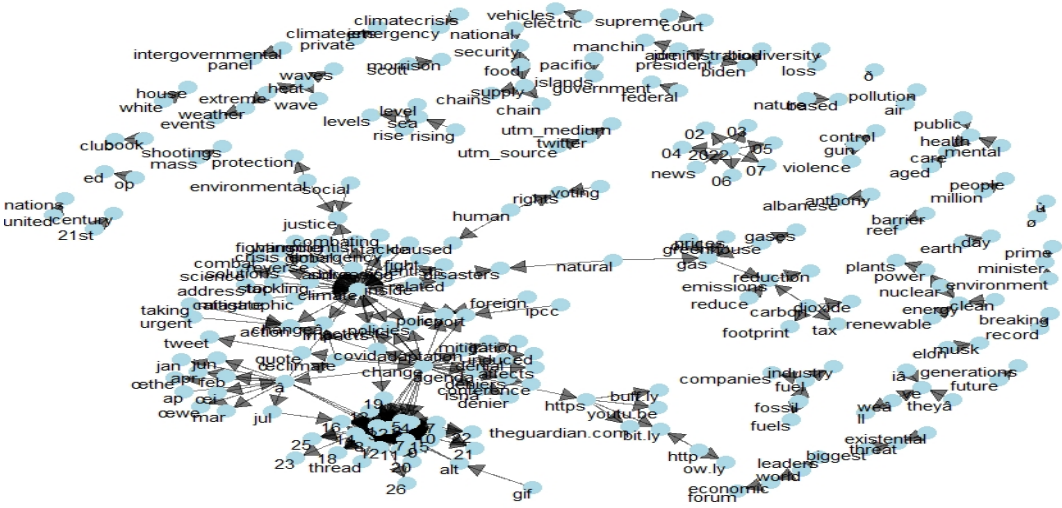


Figure 12. Bi-gram chart of Twitter linked pairs of words.

In fig 13 the Reddit bigrams are displayed, again for clarity we only show those words that have at least 100 co-occurrences based on key topic map groupings. Similar in tone and style to YouTube posts the majority of Reddit posts are more objective and less inclined to be sensationalist. Reddit has very strict rules on posting and any user breaching these may be banned. Other inducements for good behavior are *karma awards* and *coins* given to a poster by other users.

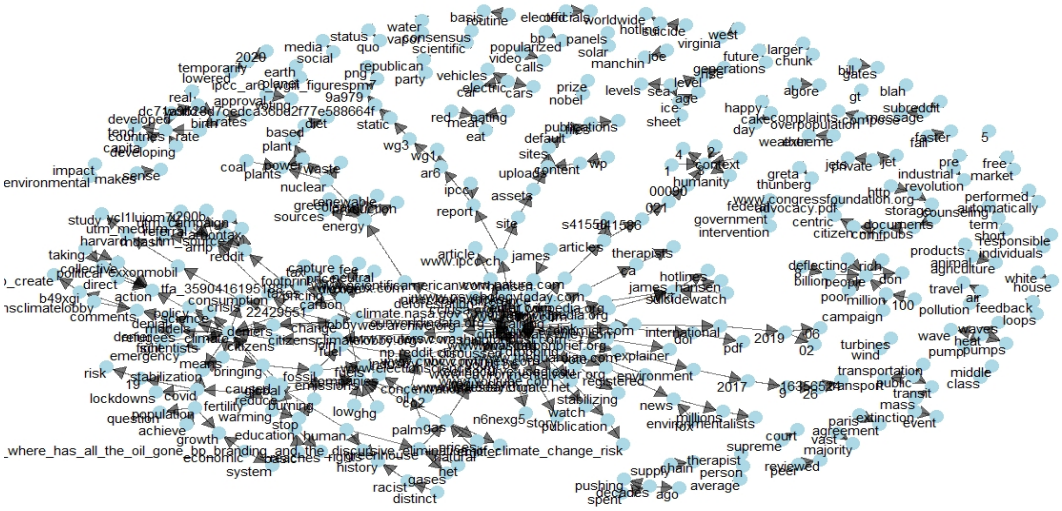


Figure 13. Bi-gram chart of Reddit linked pairs of words.

Having gathered statistics from sentiment analysis of the topic maps, comments and bigrams of paired common words we now structure the data to build the recommendation engine. The difficulty we face is that the matrix of items (videos) and users is very sparse, this is alleviated to a certain extent by generic profiling of users from Twitter and Reddit data.

The rating matrix is formed from the YouTube user rankings of videos, these are normalized by centering to remove possible rating bias by subtracting the row mean from all ratings in the row. In table 5 we present normalized ratings data for a small fraction of the overall ratings matrix. Empty spaces represent missing values where no rating data exists. The missing values are usually identified in our software as *Not Available* (NA), we did not attempt to impute missing values.

	vid1	vid2	vid3	vid4	vid5	vid6	vid7	vid8	vid9	vid10	vid11	vid12	vid13	vid14	vid15	vid16	vid17	vid18
usr1	-0.24	1.18	0.88	-1.17	0.88	0.06	-0.89	0.25	-0.88	-1.34	-0.52	1.67	0.55	1.36	-1.75	-0.21	0.79	0.72
usr2		0.49	-1.38	-0.22	-0.63	-1.34	1.12	0.55	0.00	-1.34	-0.52	0.06	0.18	0.49	1.17	-0.84	-1.06	
usr3	0.30	-0.56	0.59	-1.17	-0.25		-1.18	-0.34	0.44	0.94	-1.15	0.06	-1.65	-0.97	-0.08	-1.16		
usr4	-0.50	0.49	0.59	-0.22	-1.01	1.47	0.83	-2.11	-0.88	0.62	1.67	-0.90	0.55	-0.10	0.33	0.42	-0.79	-0.92
usr5	1.63	-1.25	0.88	-0.54		0.34	-0.03		-1.77	-0.04	1.35	1.35	-1.28		-0.92	1.37	1.06	-0.92

Table 5. Normalized ratings of YouTube videos for five users with up to 18 videos - with missing values

In table 6 we present the models error for the test data. We evaluate predictions is to compute the deviation of the prediction from the true value. This is the Mean Average Error (MAE). We also use the Root Mean Square Error (RMSE) since it penalizes larger errors than MAE and thus is suitable for situations where smaller errors are likely to be encountered. Where UBCF is user based collaborative filtering and IBCF refers to item based collaborative filtering. We also use mean squared error (MSE) as a measure.

	RMSE	MSE	MAE
UBCF	5.860	34.344	5.279
IBCF	6.216	38.637	5.573

Table 6. Recommender model error on test data

In table 7 we display the confusion matrix where n is the number of recommendations per list, TP, FP, FN and TN are the entries for true positives, false positives, false negatives and true negatives. The remaining columns contain precomputed performance measures. We calculate the average for all runs from four-fold crossvalidation.

	TP	FP	FN	TN	N	precision	recall	TPR	FPR	n
1	1.00	0.00	10.80	5.20	17.00	1.00	0.09	0.09	0.00	1.00
2	2.20	0.80	9.60	4.40	17.00	0.73	0.18	0.18	0.15	3.00
3	3.40	1.60	8.40	3.60	17.00	0.68	0.28	0.28	0.29	5.00
4	7.20	2.80	4.60	2.40	17.00	0.72	0.61	0.61	0.52	10.00
5	10.80	4.20	1.00	1.00	17.00	0.72	0.91	0.91	0.80	15.00
6	11.80	5.20	0.00	0.00	17.00	0.69	1.00	1.00	1.00	20.00

Table 7. Confusion matrix for Recommender model indicating averaged error rates - (four fold crossvalidation)

Examining the evaluation of popular items and the user-based CF methods appear to have a better accuracy and performance than the other methods. In fig14a and fig 14b we see that they provide better recommendations than the other method since for each length of top predictions list they have superior values of TPR and FPR. Thus we have validated our model and are reasonably certain of its robustness.

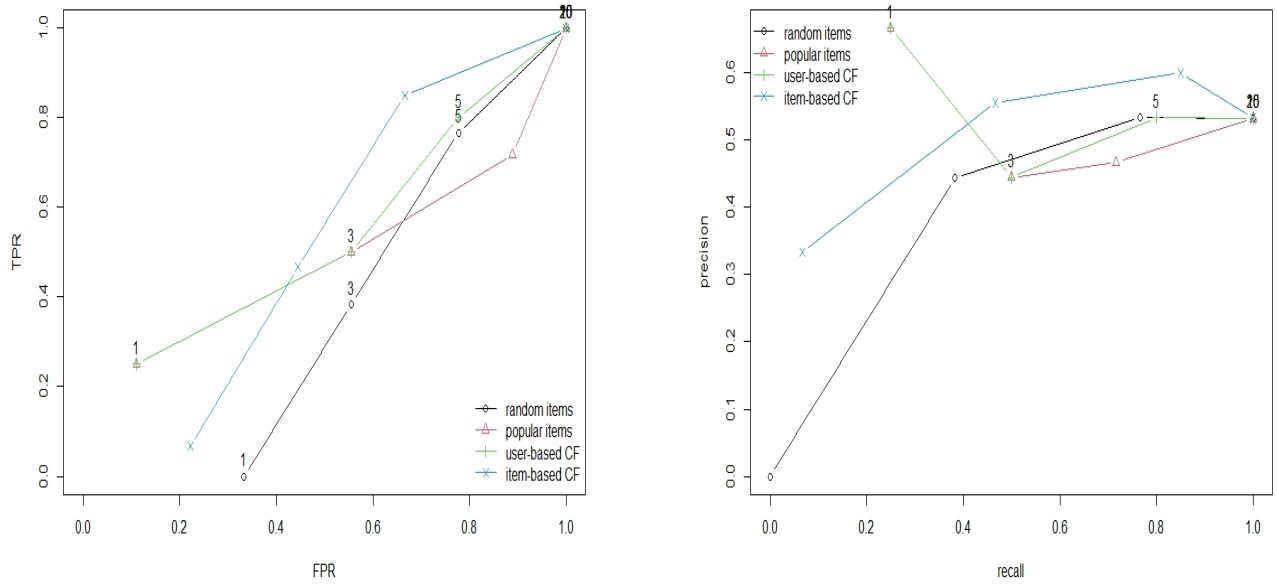
ID	User	Youtube ID	Video Title	Views	Score
1	1	$cl4Uv9_7KJE$	Fleeing climate change — the real environmental disaster	2M	1.0
2	1	$K9MaGf-Su9I$	Climate change: Europe's melting glaciers | DW Documentary	5.7M	1.0
3	1	$3CM_kkDuzGQ$	Friendly Guide to Climate Change - and what you can do to help	319K	1.0
4	1	$2bXn2F58OsM$	This tool will help us get to zero emissions (Bill Gates)	4.5M	1.0
5	2	$uynhvHZUOOo$	See what three degrees of global warming looks like	3M	1.0
6	2	$zrM1mcKmX_c$	Why NITIN GADKARI is pushing GREEN HYDROGEN	2.4M	1.0
7	2	$06m5d3mczvE$	Bill Gates Talks About How To Avoid A Climate Disaster	1.4M	1.0
8	2	$SI9GxjJwGqo$	How long before all the ice melts? - BBC World Service	89K	1.0
9	3	$rwdxffEzQ9I$	El Niño 2023 could be a monster!	1.2M	1.0
10	3	$GystZIxWQ3o$	The melting ice of the Arctic (1/2) | DW Documentary	2.5M	1.0
11	4	$Zklo4Z1SqkE$	Hydrogen Will Not Save Us. Here's Why.	1.6M	1.0
12	4	$N-yALPEpV4w$	Why renewables can't save the planet | Michael Shellenberger | TED	5.2M	1.0
13	4	$dPfIU27RGow$	SCIENTISTS JUST MADE HYDROGEN OUT OF NOTHING BUT AIR!!!	104K	1.0
14	4	$yqgMECkW3Ak$	Donald Trump Believes Climate Change Is A Hoax | MSNBC	307K	1.0
15	5	$pwvVephT1HU$	Global warming: why you should not worry	773K	1.0
16	5	$m3hHi4sylxE$	The Truth About Climate Change	2.1M	1.0
17	5	$Qdg4uQW8Dlg$	There is no climate crisis: Tom Harris	1M	1.0
18	5	$YBdmppcfixM$	"There's no emergency" – dissident climatologist Dr Judith Curry	657K	0.9
19	5	$9Q2YHGIlUDk$	The Models Are OK, the Predictions Are Wrong	876K	0.9
20	5	$1zrejG-WI3U$	Global Warming: Fact or Fiction? Featuring Physicists Soon and Bloom	1M	0.9

Table 8. Recommendations for 5 users selected at random

In operation the recommender system makes suggestions for selected users of YouTube based on their ratings of previous videos, their comments (if applicable) and related statistics. In table 8 we highlight 20 suggestions based on 5 users selected at random. Each user may obtain a differing number of recommendations Column one, identifies the video, column two gives the user ID (selected at random), column three gives the YouTube video ID (which can be pasted into a browser), column four gives the title of the video, column five gives the number of views and finally column six gives the recommender score. Where the videos stand in relation to climate change is obvious from the titles, with the exception of video 20 which appears to take a neutral stance. The score or ranking of a video is based on a value between 0.0 and 1.0, formed by the statistics generated and YouTube recommendations. Experimentally we have determined that values below 0.5 are unlikely to be of interest as we detected videos that are off-topic and little related to global warming.

6. Conclusions and Future Work

In this paper, we constructed a recommendation system based on sentiment analysis on topic maps, bigrams and graph analysis. The main source of data and was from the posts, comments and rating statistics attached to each YouTube video. From this data we were able to profile those agreeing with the global warming situation and those who were more skeptical. Although our model is successful in certain conditions it has major limitations, mainly we cannot usually identify posters from one forum to another. Posters typically have different user-names and so we would unlikely to be able extract further information, hence we went for a generic person profiling. We tried to alleviate that drawback by attempting to judge the character, sentiment and beliefs of the users. Future work must deal with improving user profiling based on their sentiment, type of language they use



(a) Comparison of ROC curves for several recommender methods.

(b) Comparison of PR curves for several recommender methods

Figure 14. Evaluation of Recommender methods

and thus gather their opinions and beliefs. Furthermore, it would be interesting to see if some users (tracked overtime) change their beliefs. Another interesting possibility would be to suggest videos that conflict with the users initial beliefs, assuming the user is open to persuasion and debate.

Author Contributions: K. McGarry conducted the experimental work and the write-up.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: MDPI Research Data Policies at https://www.mdpi.com/ethics.

Acknowledgments: The authors would like to thank Michael Hahsler for details of his R package and the three anonymous reviewers for their suggestions to improve the quality of this paper

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Spiliotopoulos, D.; Margaris, D.; Vassilakis, C. On Exploiting Rating Prediction Accuracy Features in Dense Collaborative Filtering Datasets. *Information* **2022**, *13*. https://doi.org/10.3390/info13090428.
2. Bai, Y.; Li, Y.; Wang, L. A Joint Summarization and Pre-Trained Model for Review-Based Recommendation. *Information* **2021**, *12*, 223. https://doi.org/10.3390/info12060223.
3. Kaur, P.; Goel, S. Shilling attack models in recommender system. In Proceedings of the 2016 International Conference on Inventive Computation Technologies (ICICT), 2016, Vol. 2, pp. 1–5. https://doi.org/10.1109/INVENTIVE.2016.7824865.
4. Lam, S.K.; Riedl, J. Shilling Recommender Systems for Fun and Profit. In Proceedings of the Proceedings of the 13th International Conference on World Wide Web; Association for Computing Machinery: New York, NY, USA, 2004; WWW '04, p. 393–402. https://doi.org/10.1145/988672.988726.
5. Sharma, R.; Gopalani, D.; Meena, Y. An anatomization of research paper recommender system: Overview, approaches and challenges. *Engineering Applications of Artificial Intelligence* **2023**, *118*, 105641. https://doi.org/https://doi.org/10.1016/j.engappai.2022.105641.

6. Halim, Z.; Hussain, S.; Hashim Ali, R. Identifying content unaware features influencing popularity of videos on YouTube: A study based on seven regions. *Expert Systems with Applications* **2022**, *206*, 117836. https://doi.org/https://doi.org/10.1016/j.eswa.2022.117836.
7. Zappin, A.; Malik, H.; Shakshuki, E.M.; Dampier, D.A. YouTube Monetization and Censorship by Proxy: A Machine Learning Prospective. *Procedia Computer Science* **2022**, *198*, 23–32. 12th International Conference on Emerging Ubiquitous Systems and Pervasive Networks and the 11th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare, https://doi.org/https://doi.org/10.1016/j.procs.2021.12.207.
8. Grün, B.; Hornik, K. topicmodels: An R Package for Fitting Topic Models. *Journal of Statistical Software* **2011**, *40*, 1–30. https://doi.org/10.18637/jss.v040.i13.
9. Smith, B.; Linden, G. Two Decades of Recommender Systems at Amazon.com. *IEEE Internet Computing* **2017**, *21*, 12–18. https://doi.org/10.1109/MIC.2017.72.
10. Covington, P.; Adams, J.; Sargin, E. Deep Neural Networks for YouTube Recommendations. In Proceedings of the Proceedings of the 10th ACM Conference on Recommender Systems; Association for Computing Machinery: New York, NY, USA, 2016; p. 191–198. https://doi.org/10.1145/2959100.2959190.
11. Abul-Fottouh, D.; Song, M.Y.; Gruzd, A. Examining algorithmic biases in YouTube's recommendations of vaccine videos. *International Journal of Medical Informatics* **2020**, *140*, 104175. https://doi.org/https://doi.org/10.1016/j.ijmedinf.2020.104175.
12. Chung, C.Y.; Hsu, P.Y.; Huang, S.H. βP: A novel approach to filter out malicious rating profiles from recommender systems. *Decision Support Systems* **2013**, *55*, 314–325. https://doi.org/https://doi.org/10.1016/j.dss.2013.01.020.
13. Kim, Y.; Shim, K. TWILITE: A recommendation system for Twitter using a probabilistic model based on latent Dirichlet allocation. *Information Systems* **2014**, *42*, 59–77. https://doi.org/https://doi.org/10.1016/j.is.2013.11.003.
14. Abolghasemi, R.; Engelstad, P.; Herrera-Viedma, E.; Yazidi, A. A personality-aware group recommendation system based on pairwise preferences. *Information Sciences* **2022**, *595*, 1–17. https://doi.org/https://doi.org/10.1016/j.ins.2022.02.033.
15. Leng, Y.; Yu, L.; Niu, X. Dynamically aggregating individuals' social influence and interest evolution for group recommendations. *Information Sciences* **2022**, *614*, 223–239. https://doi.org/https://doi.org/10.1016/j.ins.2022.09.058.
16. Wu, B.; Ye, Y. BSPR: Basket-sensitive personalized ranking for product recommendation. *Information Sciences* **2020**, *541*, 185–206. https://doi.org/https://doi.org/10.1016/j.ins.2020.06.046.
17. Wang, R.; Zhou, D.; Jiang, M.; Si, J.; Yang, Y. A Survey on Opinion Mining: From Stance to Product Aspect. *IEEE Access* **2019**, *7*, 41101–41124. https://doi.org/10.1109/ACCESS.2019.2906754.
18. Singh, N.; Tomar, D.; Sangaiah, A. Sentiment analysis: a review and comparative analysis over social media. *Journal of Ambient Intelligence and Humanized Computing* **2020**, *11*, 97–117. https://doi.org/https://doi.org/10.1007/s12652-018-0862-8.
19. Birjali, M.; Kasri, M.; Beni-Hssane, A. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems* **2021**, *226*, 107134. https://doi.org/https://doi.org/10.1016/j.knosys.2021.107134.
20. Phand, S.A.; Phand, J.A. Twitter sentiment classification using stanford NLP. In Proceedings of the 2017 1st International Conference on Intelligent Systems and Information Management (ICISIM), 2017, pp. 1–5. https://doi.org/10.1109/ICISIM.2017.8122138.
21. Kim, R.Y. Using Online Reviews for Customer Sentiment Analysis. *IEEE Engineering Management Review* **2021**, *49*, 162–168. https://doi.org/10.1109/EMR.2021.3103835.
22. Taboada, M.; Brooke, J.; Tofiloski, M.; Voll, K.; Stede, M. Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics* **2011**, *37*, 267–307. https://doi.org/10.1162/COLI_a_00049.
23. Ding, Y.; Li, B.; Zhao, Y.; Cheng, C. Scoring tourist attractions based on sentiment lexicon. In Proceedings of the 2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), 2017, pp. 1990–1993. https://doi.org/10.1109/IAEAC.2017.8054363.
24. Mishev, K.; Gjorgjevikj, A.; Vodenska, I.; Chitkushev, L.T.; Trajanov, D. Evaluation of Sentiment Analysis in Finance: From Lexicons to Transformers. *IEEE Access* **2020**, *8*, 131662–131682. https://doi.org/10.1109/ACCESS.2020.3009626.
25. Crone, S.F.; Koeppel, C. Predicting exchange rates with sentiment indicators: An empirical evaluation using text mining and multilayer perceptrons. In Proceedings of the 2014 IEEE Conference on Computational Intelligence for Financial Engineering and Economics (CIFEr), 2014, pp. 114–121. https://doi.org/10.1109/CIFEr.2014.6924062.
26. Romero, C.; Ventura, S. Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man and Cybernetics. Part C Appl. Rev.* **2010**, *40*, 601–618.
27. Kumar, A.; Jai, R. Sentiment analysis and feedback evaluation. In Proceedings of the in 2015 IEEE 3rd International Conference on MOOCs, Innovation and Technology in Education (MITE), 2015, pp. 433–436.
28. Missikoff, M.; Velardi, P.; Fabriani, P. Text mining techniques to automatically enrich a domain ontology. *Applied Intelligence* **2003**, *18*, 323–340.
29. McGarry, K.; McDonald, S. Computational methods for text mining user posts on a popular gaming forum for identifying user experience issues. In Proceedings of the British HCI 2017 Conference Digital Make Believe; , 2017. https://doi.org/10.14236/ewic/HCI2017.100.
30. Bose, S., RSentiment: A Tool to Extract Meaningful Insights from Textual Reviews. In *Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications : FICTA 2016, Volume 2*; Springer Singapore, 2017; pp. 259–268. https://doi.org/10.1007/978-981-10-3156-4_26.

31. Seetharamulu, B.; Reddy, B.N.K.; Naidu, K.B. Deep Learning for Sentiment Analysis Based on Customer Reviews. In Proceedings of the 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2020, pp. 1–5. https://doi.org/10.1109/ICCCNT49239.2020.9225665.
32. Thakur, N. Sentiment Analysis and Text Analysis of the Public Discourse on Twitter about COVID-19 and MPox. *Big Data and Cognitive Computing* **2023**, *7*. https://doi.org/10.3390/bdcc7020116.
33. Fellnhofer, K. Positivity and higher alertness levels facilitate discovery: Longitudinal sentiment analysis of emotions on Twitter. *Technovation* **2023**, *122*, 102666. https://doi.org/https://doi.org/10.1016/j.technovation.2022.102666.
34. Di Cara, N.H.; Maggio, V.; Davis, O.S.P.; Haworth, C.M.A. Methodologies for Monitoring Mental Health on Twitter: Systematic Review. *J Med Internet Res* **2023**, *25*. https://doi.org/10.2196/42734.
35. Kavitha, K.; Shetty, A.; Abreo, B.; D'Souza, A.; Kondana, A. Analysis and Classification of User Comments on YouTube Videos. *Procedia Computer Science* **2020**, *177*, 593–598. The 11th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN 2020) / The 10th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH 2020) / Affiliated Workshops, https://doi.org/https://doi.org/10.1016/j.procs.2020.10.084.
36. Alhabash, S.; hwan Baek, J.; Cunningham, C.; Hagerstrom, A. To comment or not to comment?: How virality, arousal level, and commenting behavior on YouTube videos affect civic behavioral intentions. *Computers in Human Behavior* **2015**, *51*, 520–531. https://doi.org/https://doi.org/10.1016/j.chb.2015.05.036.
37. Shiryaeva, T.A.; Arakelova, A.A.; Tikhonova, E.V.; Mekeko, N.M. Anti-, Non-, and Dis-: the linguistics of negative meanings about youtube. *Heliyon* **2020**, *6*, e05763. https://doi.org/https://doi.org/10.1016/j.heliyon.2020.e05763.
38. Albert, R.; Barabasi, A. Statistical mechanics of complex networks. *Rev Mod Physics* **2002**, *74*, 450–461.
39. Barabasi, A. *Network Science*, 1st ed.; Cambridge University Press, 2016.
40. McGarry, K.; McDonald, S. Complex network theory for the identification and assessment of candidate protein targets. *Computers in Biology and Medicine* **2018**, *97*, 113–123. https://doi.org/10.1016/j.compbiomed.2018.04.015.
41. Ward, K. Social networks, the 2016 US presidential election, and Kantian ethics: applying the categorical imperative to Cambridge Analytica's behavioral microtargeting. *Journal of Media Ethics* **2018**, *33*, 133–148. https://doi.org/10.1080/23736992.2018.1477047.
42. Kolaczyk, E., Statistical Research in Networks - Looking Forward. In *Encyclopedia of Social Network Analysis and Mining*; Springer New York: New York, 2014; pp. 2056–2062. https://doi.org/10.1007/978-1-4614-6170-8_41.
43. Cai, H.; Zhang, F. Detecting shilling attacks in recommender systems based on analysis of user rating behavior. *Knowledge-Based Systems* **2019**, *177*, 22–43. https://doi.org/https://doi.org/10.1016/j.knosys.2019.04.001.
44. Cruickshank, I.; Carley, K. Characterizing communities of hashtag usage on twitter during the 2020 COVID-19 pandemic by multi-view clustering. *Applied Network Science*, *5*. https://doi.org/10.1007/s41109-020-00317-8.
45. Ahmad, U.; Zahid, A.; Shoaib, M.; AlAmri, A. HarVis: An integrated social media content analysis framework for YouTube platform. *Information Systems* **2017**, *69*, 25–39. https://doi.org/https://doi.org/10.1016/j.is.2016.10.004.
46. Wang, S.; Zhang, P.; Wang, H.; Yu, H.; Zhang, F. Detecting shilling groups in online recommender systems based on graph convolutional network. *Information Processing & Management* **2022**, *59*, 103031. https://doi.org/https://doi.org/10.1016/j.ipm.2022.103031.
47. Keramatfar, A.; Amirkhani, H.; Bidgoly, A.J. Multi-thread hierarchical deep model for context-aware sentiment analysis. *Journal of Information Science* **2023**, *49*, 133–144. https://doi.org/10.1177/0165551521990617.
48. Keramatfar, A.; Rafaee, M.; Amirkhani, H. Graph Neural Networks: A bibliometrics overview. *Machine Learning with Applications* **2022**, *10*, 100401. https://doi.org/https://doi.org/10.1016/j.mlwa.2022.100401.
49. Nilashi, M.; Ali Abumalloh, R.; Samad, S.; Minaei-Bidgoli, B.; Hang Thi, H.; Alghamdi, O.; Yousoof Ismail, M.; Ahmadi, H. The impact of multi-criteria ratings in social networking sites on the performance of online recommendation agents. *Telematics and Informatics* **2023**, *76*, 101919. https://doi.org/https://doi.org/10.1016/j.tele.2022.101919.
50. Heidari, M.; Jones, J.H.J.; Uzuner, O. An Empirical Study of Machine learning Algorithms for Social Media Bot Detection. In Proceedings of the 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), 2021, pp. 1–5. https://doi.org/10.1109/IEMTRONICS52119.2021.9422605.
51. Heidari, M.; Jones, J.H.; Uzuner, O. Deep Contextualized Word Embedding for Text-based Online User Profiling to Detect Social Bots on Twitter. In Proceedings of the 2020 International Conference on Data Mining Workshops (ICDMW), 2020, pp. 480–487. https://doi.org/10.1109/ICDMW51313.2020.00071.
52. K, H.; S, S.; M, M. Social media bot detection with deep learning methods: a systematic review. *Neural Computing and Applications* **2023**, *35*, 8903–8918. https://doi.org/10.1007/s00521-023-08352-z.
53. Schneider, L.; Scholten, J.; Sándor, B. Charting closed-loop collective cultural decisions: from book best sellers and music downloads to Twitter hashtags and Reddit comments. *Eur. Phys. J. B* **2021**, *94*. https://doi.org/https://doi.org/10.1140/epjb/s10051-021-00173-0.
54. Madsen, M.A.; Madsen, D.O. Communication between Parents and Teachers of Special Education Students: A Small Exploratory Study of Reddit Posts. *Social Sciences* **2022**, *11*, 518. https://doi.org/10.3390/socsci11110518.
55. Harel, T.L. Archives in the making: documenting the January 6 capitol riot on Reddit. *Internet Histories* **2022**, *6*, 391–411. https://doi.org/10.1080/24701475.2022.2103989.
56. SemRush-Inc. Reddit statistics. https://www.semrush.com/website/reddit.com/overview/, 2023. [Online; accessed 04-February-2023].

57. Chew, R.F.; Kery, C.; Baum, L.; Bukowski, T.; Kim, A.; Navarro, M.A. Predicting Age Groups of Reddit Users Based on Posting Behavior and Metadata: Classification Model Development and Validation. *JMIR Public Health and Surveillance* **2021**, *7*.
58. Barker, J.; Rohde, J. Topic Clustering of E-Cigarette Submissions Among Reddit Communities: A Network Perspective. *Health Education and Behavior* **2019**, *46*. https://doi.org/doi:10.1177/1090198119863770.
59. Gaffney, D.; Matias, J. Caveat emptor, computational social science: Large-scale missing data in a widely-published Reddit corpus. *PLoS ONE* **2018**, *13*. https://doi.org/https://doi.org/10.1371/journal.pone.0200162.
60. Jhaver, S.; Appling, D.S.; Gilbert, E.; Bruckman, A. "Did You Suspect the Post Would Be Removed?": Understanding User Reactions to Content Removals on Reddit. *Proc. ACM Hum.-Comput. Interact.* **2019**, *3*. https://doi.org/10.1145/3359294.
61. Baumgartner, J.; Zannettou, S.; Keegan, B.; Squire, M.; Blackburn, J. The Pushshift Reddit Dataset. *Proceedings of the International AAAI Conference on Web and Social Media* **2020**, *14*, 830–839. https://doi.org/10.1609/icwsm.v14i1.7347.
62. Rivera, I. Reddit Data Extraction Toolkit, 2023. https://cran.r-project.org/web/packages/RedditExtractoR/index.html, last accessed on 2023-06-29.
63. Gertzel, B.; Ackland, R.; Graham, T.; Borquez, F. VosonSML: Collecting Social Media Data and Generating Networks for Analysis, 2022. https://cran.r-project.org/web/packages/vosonSML/index.html, last accessed on 2023-06-29.
64. Bauchi, C. Twitter Climate Change Sentiment Dataset, 2018. https://www.kaggle.com/datasets/edqian/twitter-climate-change-sentiment-dataset, last accessed on 2023-06-29.
65. Rinker, T.W. *sentimentr: Calculate Text Polarity Sentiment*. Buffalo, New York, 2021. version 2.9.0, available at *github.com/trinker/sentimentr*.
66. Feinerer, I.; Hornik, K.; Meyer, D. Text Mining Infrastructure in R. *Journal of Statistical Software* **2008**, *25*, 1–54. https://doi.org/10.18637/jss.v025.i05.
67. Feinerer, I.; Hornik, K. *tm: Text Mining Package*, 2023. R package version 0.7-11, *Available at https://CRAN.R-project.org/package=tm*.
68. Chen, Y.L.; Chang, C.L.; Yeh, C.S. Emotion classification of YouTube videos. *Decision Support Systems* **2017**, *101*, 40–50. https://doi.org/https://doi.org/10.1016/j.dss.2017.05.014.
69. Chang, W.L.; Chen, L.M.; Verkholantsev, A. Revisiting Online Video Popularity: A Sentimental Analysis. *Cybernetics and Systems* **2019**, *50*, 563–577. https://doi.org/10.1080/01969722.2019.1646012.
70. Gandhi, A.; Adhvaryu, K.; Poria, S.; Cambria, E.; Hussain, A. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion* **2023**, *91*, 424–444. https://doi.org/https://doi.org/10.1016/j.inffus.2022.09.025.
71. Rouhani, S.; Mozaffari, F. Sentiment analysis researches story narrated by topic modeling approach. *Social Sciences & Humanities Open* **2022**, *6*, 100309. https://doi.org/https://doi.org/10.1016/j.ssaho.2022.100309.
72. Csardi, G.; Nepusz, T. The igraph software package for complex network research. *InterJournal* **2006**, *Complex Systems*, 1695.
73. Li, J.; Wang, Y.; Tao, Z. A Rating Prediction Recommendation Model Combined with the Optimizing Allocation for Information Granularity of Attributes. *Information* **2022**, *13*. https://doi.org/10.3390/info13010021.
74. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
75. Grün, B.; Hornik, K. R Package topicmodels. https://cran.r-project.org/web/packages/topicmodels/index.html, 2022. Accessed: 2023-12-06.
76. Chang, J.; Gerrish, S.; Wang, C.; Boyd-graber, J.; Blei, D. Reading Tea Leaves: How Humans Interpret Topic Models. In Proceedings of the Advances in Neural Information Processing Systems; Bengio, Y.; Schuurmans, D.; Lafferty, J.; Williams, C.; Culotta, A., Eds. Curran Associates, Inc., 2009, Vol. 22.
77. Griffiths, T.; Steyvers, M. Finding scientific topics. *Proc Natl Acad Sci USA* **2004**, *101*, 5228–5235. https://doi.org/10.1073/pnas.0307752101.
78. Gaujoux, R.; Seoighe, C. A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics* **2010**, *11*, 367. https://doi.org/10.1186/1471-2105-11-367.
79. Greene, D.; Cunningham, P. A Matrix Factorization Approach for Integrating Multiple Data Views. In Proceedings of the Machine Learning and Knowledge Discovery in Databases; Buntine, W.; Grobelnik, M., Eds. Springer Berlin Heidelberg, 2009, pp. 423–438.
80. Vlachos, M.; Dunner, C.; Heckle, R.; Vassiliadis, A.; Parnell, T.; Atasu, K. Addressing interpretability and cold-start in matrix factorization for recommender systems. *IEEE Transactions on Knowledge and Data Engineering* **2019**, *31*, 1253–1266.
81. Hahsler, M. recommenderlab: An R Framework for Developing and Testing Recommendation Algorithms, 2022, [arXiv:cs.IR/2205.12371].
82. McGarry, K.; Graham, Y.; McDonald, S.; Rashid, A. RESKO: Repositioning drugs by using side effects and knowledge from ontologies. *Knowledge Based Systems* **2018**, *160*, 34–48. https://doi.org/10.1016/j.knosys.2018.06.017.
83. Wang, J.; Fan, Z.; Cheng, Y. Drug disease association and drug repositioning predictions in complex diseases using causal inference probabilistic matrix factorization. *Journal of Chemical Information and Modeling* **2014**, *54*, 2562–2569.
84. Li, W.; Xie, J.; Mo, J. An overlapping network community partition algorithm based on semi-supervised matrix factorization and random walk. *Expert Systems with Applications* **2018**, *91*, 277–285.