International Journal of Computational Intelligence Systems (2024) 17:8
https://doi.org/10.1007/s44196-023-00375-7

RESEARCH ARTICLE

A LDA-Based Social Media Data Mining Framework for Plastic Circular Economy

Yangyimin Xue[1] · Chandrasekhar Kambhampati[1] · Yongqiang Cheng[1] · Nishikant Mishra[2] · Nur Wulandhari[2] · Pauline Deutz[3]

Received: 24 February 2023 / Accepted: 21 November 2023

Abstract

The mass production of plastic waste has caused an urgent worldwide public health crisis. Although government policies and industrial innovation are the driving forces to meet this challenge, trying to understand public attitudes may improve the efficiency of this process. Social media has become the main ways for the public to obtain information and express opinions and feelings. This motivated us to mine the perceptions and behavioral responses towards plastic usage using social media data. In this paper, we proposed a framework for data collection and analysis based on mainstream media in the UK to obtain public opinions on plastics. An unsupervised machine learning model based on Latent Dirichlet Allocation (LDA) has been employed to analyse and cluster the topics to deal with the lack of annotation of the data contents. An additional dictionary method was then proposed to evaluate the sentiment of the comments. The framework also provides tools to visualise the model and results to stimulate insightful understandings. We validated the framework's effectiveness by applying it to analyse three mainstream social media, where 6 first-level topic categories and 13 second-level topic categories from the comment texts related to plastics have been identified. The results show that public sentiment towards plastic products is generally stable. The spatiotemporal distribution of each topic's sentiment is highly correlated with the number of occurrences.

Keywords LDA · Model visualisation · Sentiment analysis · Comments' classification

Abbreviations

LDA	Latent Dirichlet Allocation
NLP	Natural Language Processing
VOC	Vocabulary
RNN	Recursive Neural Network
LSTM	Long Short-Term Memory
GRU	Gated recurrent unit
BBC	British Broadcasting Corporation
LSI	Latent Semantic Indexing
HDP	Hierarchical Dirichlet Process
SVD	Singular value decomposition

✉ Yangyimin Xue
y.xue-2019@hull.ac.uk

Chandrasekhar Kambhampati
c.kambhampati@hull.ac.uk

Yongqiang Cheng
y.cheng@hull.ac.uk

[1] Department of Computer Science and Technology, University of Hull, Hull HU6 7RX, UK

[2] Business School, University of Hull, Hull HU6 7RX, UK

[3] Department of Geography, Geology and Environment, University of Hull, Hull HU6 7RX, UK

1 Introduction

Plastic products began to be widely used in the 1950s. Because of the low manufacturing cost, moderate intensity, ease of being processed, and convenience of use, plastic has become almost ubiquitous in use and is recognised with ever-increasing importance in waste management. The Ellen MacArthur Foundation report shows about 500 million tons of plastic are produced every year in the world at present, but only 10% are recycled. Furthermore, much plastic waste leaks from the disposal system to become an environmental threat. As a result, the amount of plastic in municipal waste increased from 1% to more than 10% [1]. The problem reflects that the design, production and use of plastics are linear, not a circular pattern. Taking plastic packaging as an example, about 32% produced globally was discarded into the natural environment such as the ocean and air, and around 40% of it was landfilled. Only

Published online: 10 January 2024

10% of plastic packaging were recycled, and only 2% of them have achieved the same level of regeneration [2]. In contrast, in a circular system, products (including packaging) are designed to last as long as possible, to be reusable and to be recoverable at end of life [1]. Policymakers in the UK, EU and elsewhere are devising policies to ban single use plastics and improve the recovery of plastics. Literature [3, 5] have abstracted the public views of plastic products as a reference for policymakers to formulate relevant policies and also for companies in devising business plans [4].

In psychology, cognitive science, and sociology, studies have shown that people will use comments to guide individual decisions and behaviours [9]. In Maayah B's study, it is illustrated how social media is addictive to the public [28]. But to achieve an effective plastic circular economy, one must be able to understand the range of public opinions. What, for example, is the pubic tolerance for increased costs for environmental products? To what extent are public aware of sympathetic to the causes and consequences of plastics mismanagement? As social media gradually becomes an important medium for people to obtain and publish information, massive amounts of social media and comment data are generated on the internet every day. One approach to gaining these insights is analysis comments on social media, which also has also influenced public behaviour from information acquisition to post-purchase behaviour [6–8]. Furthermore, hundreds of millions of internet users access different types of plastic-related information through the internet and social media. Compared with other traditional communication channels, social media is much more efficient. Generally, with the number of comments exploding, topics and emotions about have rapidly spread and evolved, causing psychological fluctuations and emotional changes throughout society [10]. How to remove spam comments effectively, find valuable comments, and present them to readers, or provide them for further public opinion analysis and text mining tasks, have important application values.

Nowadays, few people do public opinion analysis on specific topics similar to views on plastic products. Studies often conclude that the private sector needs to take greater action to reduce plastic waste. Due to differences in the expertise and motivation of online commentators, comments online are often unpredictable in content and feelings [11]. These text information comments can examine the theme and emotion of the comment in order to analyse the comment point of view. The purpose of this study is to obtain public topic views from plastic-related reviews to reveal the public's response to the plastic economy. We analyse comments based on social media data to receive its temporal and spatial characteristics and its spatial and temporal distribution characteristics of various topics by using an undetermined topic extraction and classification model [12]. At the same time, the number of unknown topics is classified, and the quality of the classification results is evaluated.

In this paper, we propose a framework to streamline opinion mining different aspect on the plastics based on Latent Dirichlet Allocation (LDA) [13], an unsupervised topic modelling method that can identify potential topics in documents has been adopted to explore the hidden topics or frequently appearing words in the corpus. These words are calculated based on the probabilities of the topic document and the word-topic proximity. Unlike the traditional method of using experience to judge the number and result of text classification, this paper introduces two estimation methods and visualises the results to select the appropriate text classification results. After determining each comment's various aspects, we visualise the classification results to generate a more intuitive and clear understanding of the process and the scope of the topics. This helps to summarize the subject content. In order to perform sentiment classification, we first obtain the sentiment score of each word from the established word score list, and then calculate the sentiment score result of the complete comments for normalization [16]. Finally, through topic classification and sentiment scores, insightful findings of the public's perceptions of recycled plastics and behavioural changes can be forwarded to aid policymaking.

The rest of the paper is organised as follows: Sect. 2 is the related work, Sect. 3 gives details of the framework, in Sect. 4, we give the various results and compare our results with other states of the art methods. We conclude the paper in Sect. 5.

2 Related Work

Here, we give the related work used in this research, specifically, the background of NLP, LDA, topic model evaluation and sentiment analysis.

2.1 Natural Language Processing

Natural Language Processing (NLP) is a bridge between machine language and human language. It studies how to make computers understand human language [17]. This technology often embodies the highest task and realm of artificial intelligence. It is a branch of artificial intelligence and is the intersection of computer science and linguistics. NLP is also known as computational linguistics in the linguistic discipline. The primary research of natural language processing includes grammatical analysis, semantic analysis, and text understanding. Each analysis needs a preprocessing stage that includes some significant technologies, such as Tokenisation. In this process, the input text from the document is divided into the smallest units (words, phrases, emoticons, etc.). In most cases, this unit is in the

form of words or sentences or paragraphs. The normalization process can change all letters into lowercase. Remove punctuation and stemming, which converts a complete text into basic words [18]. The next step is to generate a model, a dictionary and a corpus to filter low-frequency words and construct vectors.

2.2 LDA

The LDA [13] is a generative probabilistic model that can be used for corpus classification. The LDA is a three-level hierarchical Bayesian probability model, including word, topic and document. The principle of this method is to assume that each word is extracted from a potentially hidden topic behind it. During the generative process, topic selection and word selection are two random processes. In the generation process, topic selection and word selection are two random processes. First, for each document, extract a topic from the topic distribution. Then, a word is extracted from the word distribution corresponding to the selected topic. Repeat the above process until each word in the document is traversed. Both a document belonging to a topic and a topic that can be represented by a word are assumed to follow a multinomial distribution. In short, the purpose of LDA is to identify topics from documents, i.e., turn the document-word matrix into a document-topic matrix (distribution) and a topic-word matrix (distribution).

The advantage of LDA is that it is an unsupervised machine learning without relying on the manually annotated training dataset, i.e., the only inputs are the set of documents and the number of topics. In addition, the LDA can always find representative words to describe each topic.

For the latent text model of LDA, the main disadvantage is that it has not considered the position of a series of words in the text. Hence, it cannot distinguish the different meanings that the same words in different orders can express. Moreover, because long texts contain more words, it is difficult to match their topics. Another problem that needs to be solved is that words composed of different topics in LDA are reused. This leads to an overlapping rather than independence of topics.

First, we define D as the document set. t is a collection of topics. In document set D, document d can be regarded as a word sequence $< W_1, W_2, \ldots, W_n >$, d contains n words. All the different words in this D combined into a large set vocabulary (VOC). In D, the probability of each document d corresponding to different topics $\theta_d < pt_1, pt_2 \ldots, pt_n >$, where pt_i is the probability that d corresponds to the ith topic in t. Intuitive look at calculation, $pt_i = nt_i/n$, where nt_i is the number of words corresponding to the ith topic in d, and n in d is the total number of all words. For each t in T, the probability of generating different words $\varphi_t < pw_1, pw_2 \ldots, pw_n >$. Where pw_i is the probability that t generates the ith word in VOC. The calculation method is also very intuitive, $pw_i = nw_t/n$, where nw_i is the number of the ith word in the VOC corresponding to the topic, and total number of all words corresponding to the topic is n.

$$p(\mathrm{w}|\mathrm{d}) = P(\mathrm{w}|\mathrm{t}) \times P(\mathrm{t}|\mathrm{d}). \tag{1}$$

The formula is the core process which to take the topic layer as the middle layer connecting the words and document, then give the probability of W in document d through θ_d and φ_t. Using the current θ_d and φ_t, we can calculate the $p(w|d)$ of a word in a document when it corresponds to any topic, and then update the topic corresponding to the word according to these results. Then, if the update changes the topic corresponding to the word, it will in turn affect θ_d and φ_t.

Below is the learning process of LDA algorithm:

At the beginning, θ_d and φ_t are randomly assigned (for all d and t). Then, repeated the above core process continuously, until the final convergence result, this is the output of LDA.

For the ith word w_i in one of the document d_s, when the topic corresponding to the word can be t_j, then the above formula can be modified as:

$$P_j\left(w_i|d_s\right) = P\left(w_i|t_j\right) * P\left(t_j|d_s\right). \tag{2}$$

When enumerating the topics in T to get all $P_j\ (w_i|d_s)$, where the value of j is $1 \sim k$. Then, according to these probability values, it is the ith word w_i in d_s,select a topic. The simplest idea is to take the w_i that maximises $P_j\ (w_i|d_s)$ (note that only j is a variable in this formula), which is argmax $[j]P_j\ (w_i|d_s)$.

Next, if the ith word w_i in d_s chooses a topic from the original one, it will affect θ_d and φ_t (It can be deduced from the calculation formula of the above two vectors). This influence will also be retransmitted to the calculation of $P(w|d)$ above. Calculate $P(w|d)$ for all w in all d in D and reselect the topic as an. Execute this method until after n loop iterations, it will cover and get the final result required by LDA.

Here, we acquired two probability $P(word|topic)$ and $P(topic|document)$.

2.3 Topic Model Evaluation

Most of the traditional methods use a visual inspection or prior knowledge to evaluate topic selection models' performance. The most intuitive way is to judge the extracted topic manually, but obviously, this is time consuming. Manual visual judgment mainly includes evaluating clustering results using visualization technology and the introduction of topics meaning by keywords generated from the topic model [15].

On the other hand, automatic evaluation methods include evaluating the clustering effect by the Silhouette Coefficient and coherence to measure whether the words in the topic are

coherent. Silhouette Coefficient is an evaluation method of clustering performance, first proposed by Peter J. Rousseau in 1986. It is a combination of cohesion and resolution. It is one of the evaluation methods of the impact of different algorithms or different operation modes on the clustering results based on the same original data. For each sample i in the data, the average distance from i to other samples in this cluster is *a(i)*. The minimum average distance from I to all samples of other clusters is b*(i)*. The formula can be rewritten as:

$$S(i) = b(i) - a(i) / \max\{a(i), b(i)\} \tag{3}$$

$$S(i) = \begin{cases} 1 - a(i)/b(i), & a(i) < b(i) \\ 0, & a(i) = b(i) \\ b(i)/a(i) - 1, & a(i) > b(i) \end{cases} \tag{4}$$

(1) Calculate the average distance *a(i)* between sample i and other samples in the same cluster. When *a(i)* is smaller, the grouping effect of cluster i is better. $a(i)$ is regarded as the intra cluster dissimilarity of sample i.
(2) The average distance between sample I and other samples in another cluster b is $b(i)$. It is defined as the inter cluster dissimilarity of sample i: $b_i = \min\{b_{i1}, b_{i2}, \ldots b_{ik}\}$
(3) When S_i is close to 1, the clustering of sample i is reasonable. If S_i is close to – 1, the opposite is true.

Topic coherence is to score the topic by comparing the semantic similarity between the words with higher scores in the topic.

$$score(v_i, v_j, \varepsilon) = \log(p(v_i, v_j) + \varepsilon) - \log(p(v_i)p(v_j)) \tag{5}$$

V is a group of words used to describe the subject, ε Means to ensure that the score returns a real number. Here, the smaller the epsilon is, the smaller the result will be. Word probability p is calculated by counting the frequency of words on the corpus. The algorithm is based on the original corpus of the training topic model and does not rely on the external corpus.

2.4 Sentiment Analysis

Sentiment analysis is the process of analysis, processing, induction, and subjective text reasoning with emotional colours [19, 20]. Sentiment analysis can use traditional methods based on sentiment dictionary or deep learning-based methods [25]. The dictionary-based method mainly consists of formulating a series of sentiment dictionaries and rules, disassembling sentences, analysing and matching dictionaries on the text (generally part-of-speech analysis, syntactic dependency analysis), calculating sentiment values, and finally using sentiment values as the sentiment tendency of the text Basis of judgment [16, 21, 22]. Based on deep learning sentiment classification, the sentence is first based on deep learning sentiment classification. The sentence is first preprocessed, such as word segmentation, stop words and simplified and traditional conversion. Then, it is word vector coding, as well as the feature extraction using RNN (Recursive Neural Network) such as LSTM (Long Short-Term Memory) or GRU (gated recurrent unit) [23]. The operating procedures are: the document needs to be composed of sentences and tags. Tokenize the sentence, so a list of words represents it. Then, through simple unigram word features, subjective and objective instances are used, respectively, to maintain a balanced and uniform class distribution in the training set and test set. The features are used to obtain a feature-value representation of our dataset. Then we need to train the classifier on the training set, and finally output the results.

3 The Proposed Framework

The framework consists of four parts, data acquisition and pre-processing, topics classification, model selection and visualisation, and sentiment analysis, as illustrated in Fig. 1.

Firstly, the comments data can be explored from social media online. Next, the comments are preprocessed to filter out the noise of repeated content or comments with non-text structure for processing in the comments. Following preprocessed, unsupervised machine learning LDA was trained on the training data set. Afterwards, an evaluation process and a visualisation tool to help us select the number of topics and summarise the topics. At the same time, the sentiment evaluation model gives each comment a sentiment score. The score can be used to classify emotions or study the changes in emotions. Finally, by combining subject classification and emotional scores, we can get a time series of emotional changes in different aspects of plastic products.

As the data explored are from the comments left by readers of major British media outlets including BBC, the Guardian and Mail Online under the news about recyclable plastic, the data has the characteristics of unpredictability and without available classification labels. Therefore, we need to use some techniques that do not require a training set to classify comments. Here, the researcher introduces a word based unsupervised learning method LDA.

3.1 Data Acquisition and Preprocessing

A few media such as the British Broadcasting Corporation (BBC), the Guardian and the Mail Online have largely

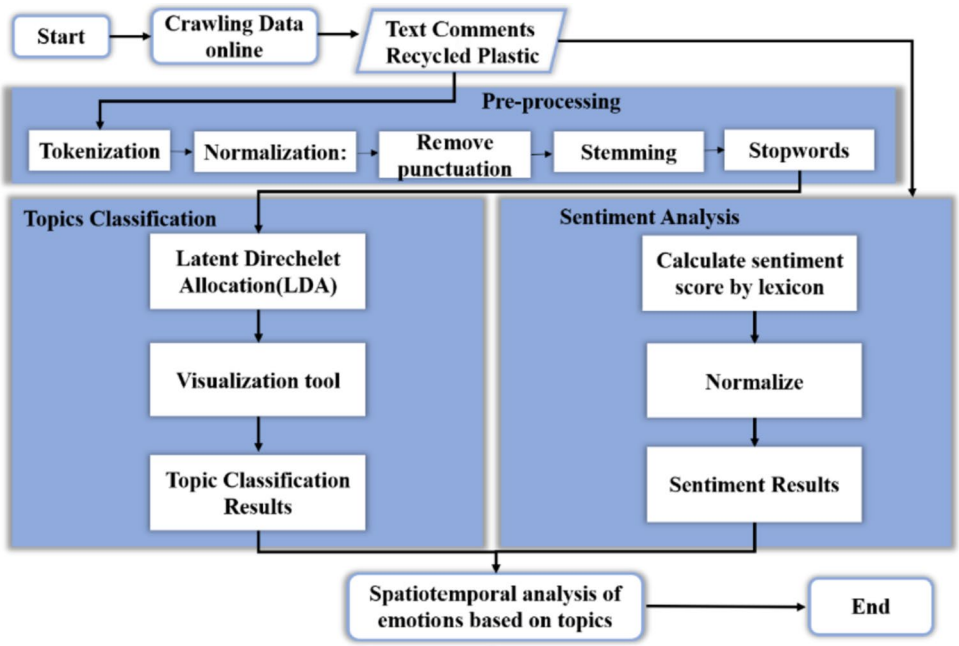

Fig. 1 Research method

dominated online news consumption in the UK [24], Readers' trust of these outlets means no matter what the news is, they will search and read directly and frequently. The comments expressed after reading the news are often highly related to the news topic.

This step is the process taken by algorithm 1, including:

The original comments text contains interference information such as spaces, http links, and punctuation marks. To eliminate noise and improve the efficiency of word segmentation, the original data must be text filtered. We use Python regular expressions to filter the original social media text and remove interference information (such as http links, punctuation), stop words, low quality text, and repeated text. Figure 2 shows the ranking of 30 most frequently occurred words in our collected data.

3.2 Topics Classification

The topic extraction and classification framework are constructed based on the LDA topic model, and public topic emotions are obtained in layers from relevant social media texts. LDA topic model is used for topic extraction to generate a topic probability distribution of each text and word probability distribution of each topic. After that, the sample data of the annotated subject is classified into the entire data set.

Algorithm 1

Start
1. Take the comments that have been crawled in the previous stage.
   For all comment in dataset:
   remove_unwanted_word
   remove_short_word
   *remove_stop_word_*
   make_lowercase
   filtet_noun_adjective
   return tokenized_reviews
2. Calculate frequency words in the data.

End

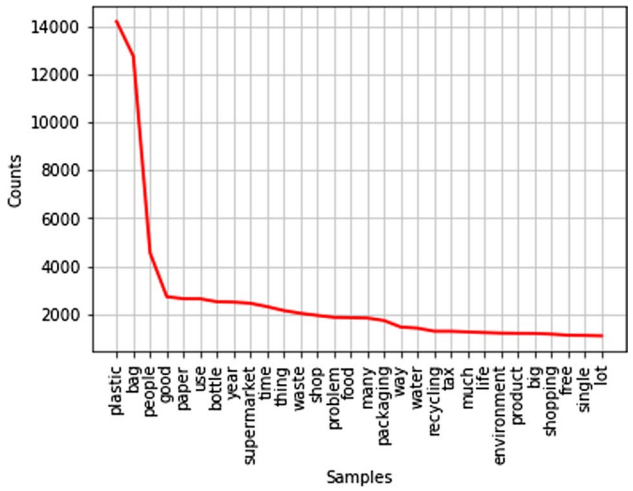

Fig. 2 The most frequent 30 words in the data

LDA models are capable of recognising topics in documents and mining the corpus for hidden information, and have a wide range of uses in scenarios such as topic aggregation, information extraction from unstructured text, and feature selection. It can (1) uncover topic patterns hidden in the corpus; (2) annotate documents according to the topics; and (3) use the annotation to organise, collate, summarise and retrieve the documents. LDA is a probabilistic model for solving the problem of text topic modelling. It is an unsupervised learning method that automatically discovers hidden topic structures from a large number of documents and assigns each document to one or more topics (Fig. 3) [25].

When building a topic model, some main parameters need to be set in advance. It includes the number of topics K, a priori alpha of topic distribution, a priori beta of word distribution, and the number of documents to be used in each training block chunk size. The total number of training evaluation passes. Maintained in this experiment. ($\vec{\alpha} = 5$ / K, $\vec{\beta} = 0.1$).

How to select the appropriate number of topics? The more topics we choose, the more specific the topics will be. However, choosing too many topics may make it impossible to distinguish comments in a meaningful way. At the same time, too few topics will lead to the combination of comments towards the same aspects, which should belong to different categories.

By comparing the quality of the generated topics, the optimal number of topics is selected. The next step is to calculate and compare the probabilities of the topics that a single review belongs, and determine the topics of comment. Finally, the theme is used to classify all documents.

Different topics are constructed from the LDA model [26], where each topic is a combination of keywords, and each keyword contributes a certain weight to the topic. From the keyword list, the words conducive to our understanding and summary of topics. These keywords and weights will be used to summarise the content of the topic.

For each LDA process, the following steps are followed as Algorithm 2:

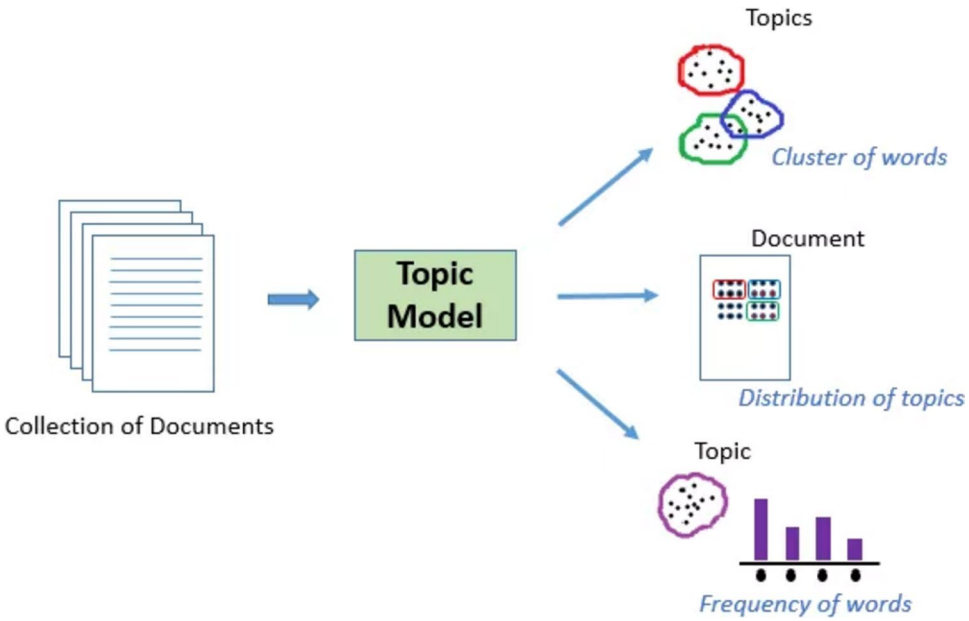

Fig. 3 A workflow of the typical theme model [25]

Algorithm 2

Start
1. From the previous pre-processed step, get the term list.
2. Selected LDA model, calculate the probability of each word appearing in the document and corpus, also the probability of proximity of document-topic and word-topic.
   for all topics k ∈ [1, K]:
   sample mixture components $\vec{\varphi}k \sim Dir(\vec{\beta})$
   end for
   for all comments m ∈ [1, M]:
   sample mixture proportion $\vec{v}m \sim Dir(\vec{\alpha})$
   sample document length $N_m \sim Poiss(\xi)$
   for all words n ∈ [1, N_m*]:*
   sample topic index $z_{m,n} \sim Mult(\vec{v}m)$
   sample term for word $w_{m,n} \sim Mult(\vec{\varphi}z_{m,n})$
   end for
   end for
   Parameters and variables:
   K: the number of topics
   $\vec{\varphi}k$*: word distribution for topic k*
   $\vec{\alpha}$, $\vec{\beta}$*: Dirichlet parameters*
   M: the total number of comments
   V: vocabulary size
   $\vec{v}m$*: topic distribution for comments m*
   N_m*: the length of comments m*
   $z_{m,n}$*: topic index of nth word in comments m*
   $w_{m,n}$*: a particular word for word placeholder [m, n]*
3. Display the results of topic classification (key words and word frequency under Hidden topic content)
4. Summary topics.
End

3.3 Model Selection and Visualisation

Next, we uses a web-based interactive tool to show the results of the LDA model [14], including the meaning of each topic, the prevalence of each topic and the relation of each topic.

LDAvis is a web-based interactive visualisation system. The system not only provides a universal ordering of words within a topic, but also displays the thematic distribution of words of particularity, exclusionary words. In turn, it proposes a new criterion relevance. The paper also utilises the results of a user study to demonstrate that the selection of words in decreasing order of probability is not optimal for topic interpretation. For topic summarisation, the most optimal explanation is not simply words by top frequency, nor is it absolutely special words. Rather, words that tend to be more relevant should be utilised to explain topic categorisation.

The operation interface is illustrated in Fig. 4, it can be selected to view the particular topic by hovering over the circle on the left. After selection, the right hand side will show the top 30 words that relevant to this topic. It can be used by these words to summarise the mean of this topic. The size of the circle shows the frequency of this topic. Here, we use the multi-dimensional analysis, extract the principal components as dimensions, and put the topic distribution on these two dimensions. The distance between the topic represents their relevance. The order of the words on the right hand side illustrates the weights of the words contributing to the topic. The relevance can be adjusted by the parameter $\lambda \subseteq [0,1]$. Changing the value of λ will change the weight order of the words in the topic, which helps to visualise the core content of the topic. The larger the λ, the more high frequency of the words, whilst the smaller the λ, the more special words.

$$\text{relevance} = \lambda * \text{p(w|t)} + (1 - \lambda) * \frac{\text{p(w|t)}}{\text{p(w)}}. \tag{6}$$

The goal of topic classification is to have as few topic categories as possible and relatively independent. Because too many topic classifications are meaningless. Overlapping topics will confuse the research results. The framework has flexible functions to set the number of categories and topics.

3.4 Sentiment Analysis

We uses NLTK's Vader analysis tool [16] in this framework as Algorithm 3. VADER (Valence Aware Dictionary and sEntiment Reasoner), is a dictionary and rule-based sentiment analysis tool that does not need to be trained or

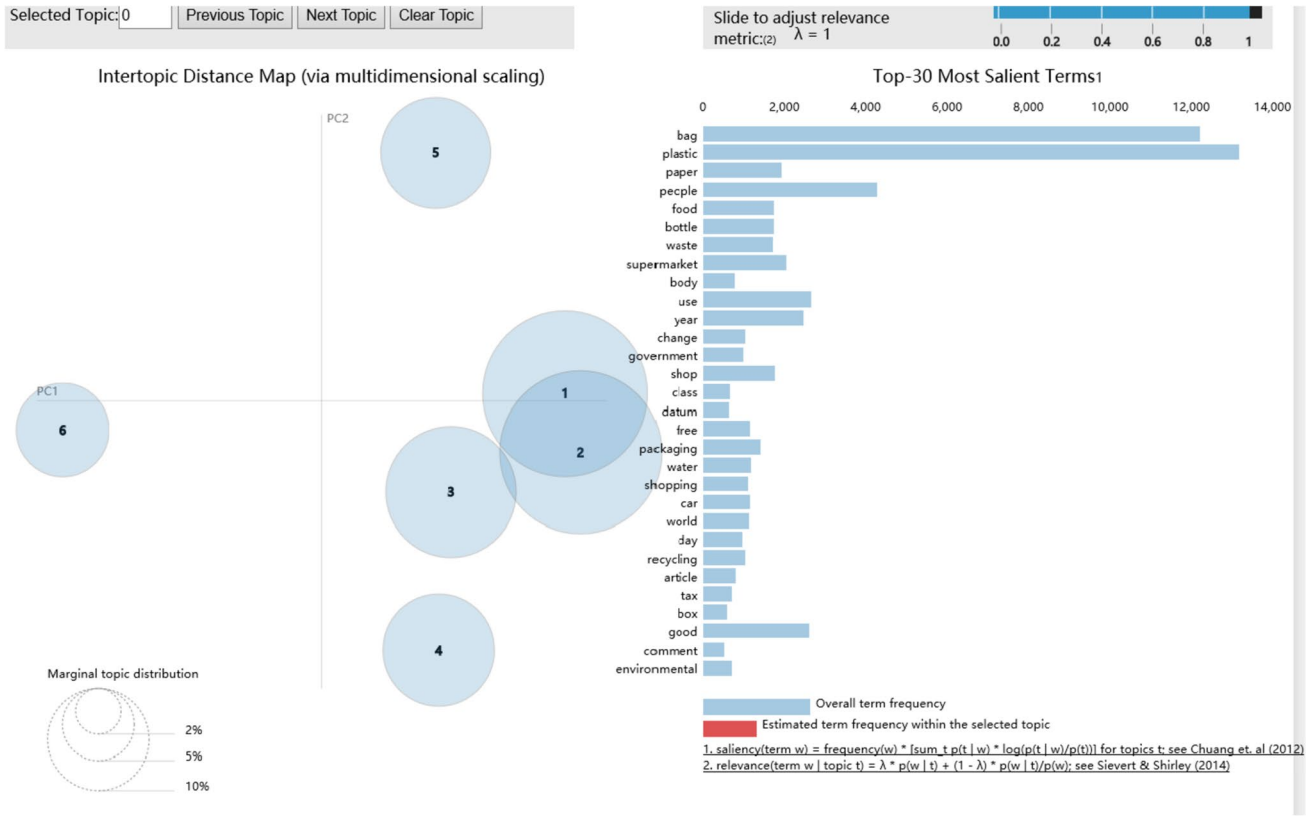

Fig. 4 Topic model visualisation

Algorithm 3

Start
1. Take the original comments.
2. By iterating the SentiText lexicon, then calculate whole sentiment score for comment.
   for w in lexicon:
   w_score = w_dict
   end for
   tol_score =sum(w_score)
3. Normalize the score.
   *fin_score = total_score / sqrt (total_score * total_score + aplha)*
4. Save the result.

End

customised in order to obtain a sentiment assessment score, and it is specially tuned to the sentiments expressed in social media. It is designed for social media content, so it performs best on content you can find on social media.

The tool will provide a compound score for each comment. The compound score is calculated by adding each word's corresponding scores in the dictionary, then adjusting them according to the rules, and finally normalising them to be between − 1 and + 1. The dictionary contains more than 9000 words manually marked with sentiment scores. The normalise formula as:

$$score_{\text{norm}} = \frac{\text{score}}{\sqrt{\text{score}^2+\alpha}}. \tag{7}$$

α in the formula is a parameter used to adjust the interval distribution of normalized results. The effect of the Eq. (7) is shown in Fig. 5. To classify sentiment more conveniently, alpha = 15 has been selected in our paper. This makes the normalised result more uniform within the range.

Finally, set a standardized threshold to classify sentences:

positive sentiment: compound score ≥ 0.05.

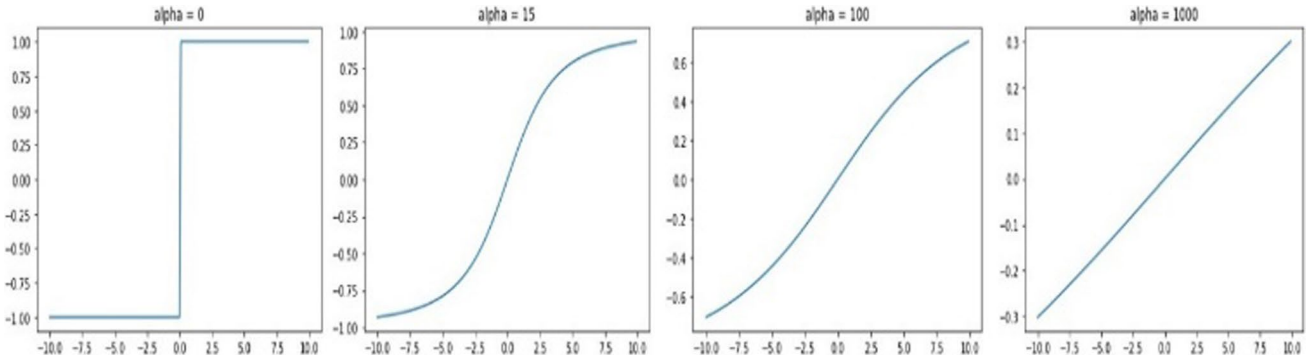

Fig. 5 How alpha affect normalized result

neutral sentiment: (compound score > – 0.05) and (compound score < 0.05).
negative sentiment: compound score ≤ – 0.05.

4 Results and Analysis

4.1 Datasets

The data are captured from the reader feeds news articles from BBC, the Guardian and Daily Mail Online. Specially, we have 7924 comments from 26 news from BBC, date ranges from August 2017 to October 2019; 22,857 comments from 97 news from Guardian, date ranges from July 2016 to October 2019; and 3221 comments from 22 news from Daily Mail Online, date ranges from January 2018 to October 2019.

4.2 Topics Classifications

Table 1 shows the results of categorising comments according to the topic. We divided the comments into six categories: *plastic product*, *shopping*, *policy*, *family*, *food* and *others*. People pay more attention to plastic products, shopping, and policy than the content of family, food, etc., 10,168, 9281 and 5700, respectively. The order of the keywords is in order of magnitude of the coefficients. The coefficient is the frequency of occurrence of the word. Words marked in red are key words that are summarised to facilitate the reader's understanding of the topic.

To understand the fine details, we conducted a secondary classification for the top three topics with the largest number of messages, i.e. plastic products, shopping and policy as Table 2.

The topic of plastic products includes seven sub-themes: reuse, resource, recycling, water package, pollution and others. There are three sub-topics under the shopping topic: reused package, paid package, and delivery package. In addition, business, media, and environment are sub-themes of the policy topic.

4.3 Model Selection

We have compared the coherence scores of different topic numbers in Fig. 6. We can see that when the number of topics is 6, the coherence score is the highest. Table 3

Table 1 First level topics description

	Keywords	Numbers
Topic1: product	0.090*"plastic" + 0.020*"bottle" + 0.020*"waste" + 0.013*"water" + 0.012*"oil" + 0.012*"product" + 0.012*"recycling" + 0.011*"use" + 0.010*"packaging" + 0.010*"glass"	10,168
Topic 2: shopping	'0.145*"bag" + 0.064*"plastic" + 0.023*"paper" + 0.021*"supermarket" + 0.021*"use" + 0.018*"shop" + 0.014*"people" + 0.014*"free" + 0.013*"shopping" + 0.011*"year"	9281
Topic 3: policy	0.034*"people" + 0.020*"good" + 0.019*"change" + 0.018*"government" + 0.017*"thing" + 0.015*"world" + 0.013*"problem" + 0.013*"tax" + 0.013*"environmental" + 0.011*"country"	5700
Topic 4: family	0.022*"people" + 0.022*"year" + 0.018*"day" + 0.016*"time" + 0.015*"car" + 0.014*"man" + 0.013*"kid" + 0.012*"child" + 0.011*"coffee" + 0.010*"many"	3550
Topic 5: food	0.027*"food" + 0.016*"box" + 0.014*"packaging" + 0.014*"fruit" + 0.012*"meat" + 0.011*"stuff" + 0.011*"good" + 0.010*"local" + 0.009*"loose" + 0.009*"small"	2763
Topic: other	0.028*"body" + 0.024*"class" + 0.023*"datum" + 0.019*"comment" + 0.016*"article" + 0.010*"demand" + 0.009*"underline" + 0.009*"thank" + 0.009*"true" + 0.008*"question"	2119

Table 2 Topic description in

Fist level topic	Second level topic	Weights and keywords
PLASTIC PRODUCT	Other	0.035*"name" + 0.033*"link" + 0.032*"body" + 0.028*"datum" + 0.025*"com" + 0.024*"class" + 0.010*"petroleum" + 0.009*"underline" + 0.009*"big" + 0.007*"wind"
	Reuse	0.105*"bottle" + 0.052*"glass" + 0.024*"plastic" + 0.020*"milk" + 0.019*"container" + 0.018*"deposit" + 0.015*"shop" + 0.012*"people" + 0.011*"scheme" + 0.009*"supermarket"
	Resource	0.083*"plastic" + 0.024*"oil" + 0.021*"use" + 0.018*"energy" + 0.018*"product" + 0.016*"bag" + 0.011*"alternative" + 0.010*"paper" + 0.008*"single" + 0.008*"material"
	Recycle	0.074*"plastic" + 0.028*"waste" + 0.018*"bag" + 0.017*"recycling" + 0.016*"food" + 0.016*"packaging" + 0.012*"problem" + 0.012*"people" + 0.011*"recycle" + 0.009*"much"
	Water package	0.076*"water" + 0.021*"plastic" + 0.014*"oil" + 0.012*"demand" + 0.012*"tap" + 0.011*"year" + 0.009*"growth" + 0.009*"coal" + 0.009*"stuff" + 0.009*"bottled"
	Pollution	0.113*"plastic" + 0.032*"bag" + 0.018*"ocean" + 0.016*"waste" + 0.012*"food" + 0.012*"country" + 0.011*"sea" + 0.010*"rubbish" + 0.009*"problem" + 0.009*"thing"
POLICY	Business	0.032*"plastic" + 0.022*"government" + 0.022*"tax" + 0.018*"bag" + 0.016*"problem" + 0.015*"money" + 0.014*"people" + 0.010*"environmental" + 0.010*"good" + 0.009*"big"
	Media	0.031*"good" + 0.022*"change" + 0.018*"thing" + 0.017*"climate" + 0.017*"point" + 0.013*"article" + 0.011*"people" + 0.011*"way" + 0.010*"environment" + 0.010*"great"
	Environment	0.040*"people" + 0.023*"world" + 0.014*"human" + 0.014*"country" + 0.013*"many" + 0.009*"global" + 0.008*"poor" + 0.008*"population" + 0.008*"planet" + 0.008*"thing"
SHOPPING	Reused package	0.015*"people" + 0.009*"time" + 0.008*"stuff" + 0.008*"person" + 0.007*"bottle" + 0.007*"car" + 0.006*"year" + 0.006*"coffee" + 0.006*"good" + 0.005*"thing"
	Paid package	0.028*"bag" + 0.025*"shop" + 0.016*"plastic" + 0.016*"paper" + 0.014*"charge" + 0.014*"supermarket" + 0.013*"customer" + 0.013*"people" + 0.011*"money" + 0.011*"free"
	Delivery package	0.150*"bag" + 0.082*"plastic" + 0.025*"use" + 0.020*"paper" + 0.017*"supermarket" + 0.015*"shopping" + 0.014*"people" + 0.013*"year" + 0.013*"single" + 0.011*"carrier"

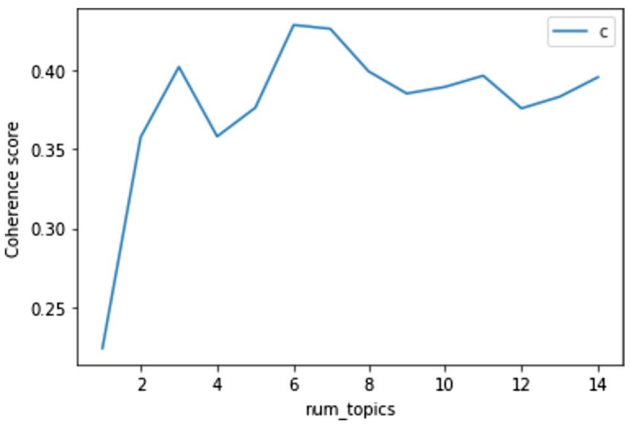

Fig. 6 The coherence score of different number topics

summarises the details of the order of keyword in different topics.

Our basic method of summarizing clusters is using visualization function to find typic words with high frequency ($\lambda = 1$) and special characteristics ($\lambda = 0$) as well as a balance indicator ($\lambda = 0.5$) between high frequency and special characteristics according to six topics including product, food, family, policy, shopping and other. As seen in Table 5, when λ is 1, the words that appear more frequently are more relevant to the topic. When λ is set to 0, the words that are more special and exclusive are more relevant to the topic. For example, when we look inside the topic policy, no matter what the value of λ is, the common words are change, government, tax and environment. These are important words which help us summarize the topic. In addition, climate, poor, global, job and action appear in special words, and people and world appear in high-frequency words. Therefore, we think this topic is about policy.

According to the distribution of words under the topic model generated by different R, this is naturally suitable as a naming scheme: only the number of most likely words (e.g. 5–10) and the most special words in the distribution need to be used as the topic descriptor. This usually works well.

4.4 Sentiment Analysis of the Topics

The average sentiment scores as shown in Table 4. The overall average sentiment tends to be neutral. People's comments on shopping, food and family are relatively positive. The others are neutral.

Figures 7 and 8 illustrate the most frequent words that appear in the positive and negative comments. Among them, people are generally more negative about waste.

4.5 Performance Comparisons

We compared the results of LDA, LSI (Latent Semantic Indexing) and HDP(Hierarchical Dirichlet Process) [27]. The LSI is a simple and practical topic model. LSI is based on singular value decomposition (SVD) to get the topic of

Table 3 The order of keyword in different topic

	$\lambda=0$	$\lambda=0.5$	$\lambda=1$
Product	Bottle, waste, water, recycling, glass, name, link, recycle, energy, com	Plastic, bottle, waste, water, recycling, oil, product, glass, name, link	Plastic, bottle, waste, water, oil, product, recycling, use packaging, glass
Food	Box, fruit, meat, loose, cardboard, fish, vegetable, garbage, legislation, fresh	Food, box, fruit, meat, loose, cardboard, fish, vegetable, garbage, packaging	Food, box, packaging, fruit, meat, stuff, good, local, loose, small
Family	Man, kid, child, coffee, cup, woman, school, self, dog, self	Man, day, kid, child, coffee, cup, car, woman, year, school	People, year, day, time, car, man, kid, child, coffee, many, old
Policy	Change, government, tax, environment, climate, poor, global, job, term, action	Change, government, people, tax, world, good, environment, issue, planet, human	People, good, change, government, thing, world, problem, tax, environmental, issue
Shopping	Bag, paper, free, shopping, charge, carrier, liner, charity, grocery, cole	Bag, plastic, paper, supermarket, shop, free, shopping, use, charge, carrier	Bag, plastic, paper, supermarket, use, shop, people, free, shopping, year
Other	Body, class, datum, comment, underline, thank, question, animal, land, soft	Body, class, datum, comment, underline, thank, article, question, land, animal	Body, class, datum, comment, article, demand, underline, thank, true, question

Table 4 The average sentiment score

	Average sentiment score
Whole comments	0.04524
Policy	0.04664
Shopping	0.08497
Food	0.06074
Other	0.02746
Family	0.00502
Plastic product	0.02172

the text. HDP model uses the property of infinite category division of Dirichlet process in finite space, adaptively adapts the number of topics to get the topic set with optimal allocation of document set structure.

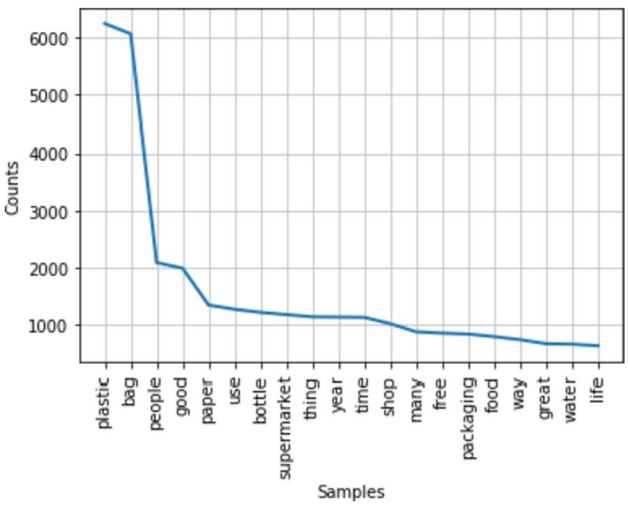

Fig. 7 The most frequent 20 words in the positive comments

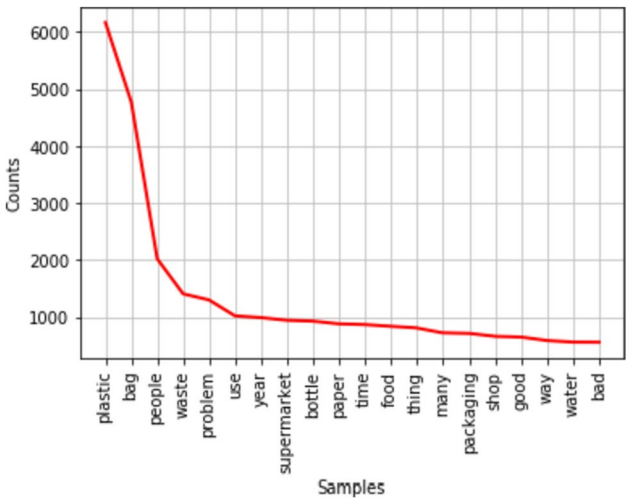

Fig. 8 The most frequent 20 words in the negative comments

Table 5 Topics model's performance

Topic Model Performance			
Performance Metrics	LDA	LSI	HDP
Silhouette Coefficient	0.055307	– 0.065895	– 0.021314
Coherence Score	0.320015	0.274701	0.174872

LDA's silhouette Coefficient score of 0.055 is closer to 1, indicating that LDA's clustering is reasonable. The other two theme models are negative, indicating that their classification results are more than supposed to be less accurate.

In the Coherence Score, LDA achieved the highest score of 0.32, which indicates that words with similar meanings in the LDA model tend to appear in similar contexts. Most of the words under its topic classification are closely related, then the topic is considered more coherent. In Table 5, the results show that LDA has advantages in clustering effect and topic coherence of short text documents.

Table 6 Sentiment results on topics analysis

Sentiment Results on Topics		
Topic	Sentiment	Results (in Percent)
Policy	Negative	7.8467
	Neutral	6.4590
	Positive	2.6682
Shopping	Negative	12.5369
	Neutral	8.8115
	Positive	6.2893
Food	Negative	3.4067
	Neutral	2.7277
	Positive	2.0934
Other	Negative	2.4448
	Neutral	2.0220
	Positive	1.8433
Family	Negative	3.9457
	Neutral	3.7402
	Positive	2.8856
Product	Negative	11.7775
	Neutral	10.7323
	Positive	7.7693
Total Percentage		100

4.6 Topic-Based Sentiment Analysis

It can be concluded from Table 6 that readers often provide comments on plastic news. The result of the analysis is 14,090 positive comments, 11,583 negative comments and 7908 neutral comments. Although there are many positive comments, there are still many negative comments, which means that the public is not satisfied with the current situation of plastic products, or needs some improvement to eliminate negative comments. In terms of policies, there were 2635 positive comments and 896 negative comments. In terms of shopping, positive reviews were 4210 and negative reviews were 2112. In terms of products, positive reviews were 3955 and negative reviews were 2609.

From the average sentiment score, the public is conservative and optimistic about plastics' current situation. Among them, shopping, dining, and family are positive. However, judging from the number of comments classified by sentiment, the number of negative comments accounted for the main component. There were more negative comments on the six topics than positive comments.

4.7 Time-Based Sentiment Analysis

The time series analysis of the six first-level topic categories is shown in Fig. 9. Except for the shopping topic that declined slightly after 2018, the remaining topics are all on the rise. Especially after 2017, the volume of comments has skyrocketed.

In Fig. 10, it shows the time series of people's emotional changes. People's attitudes towards policy gradually changed from negative to positive. People's emotions about shopping became positive. The public's satisfaction with food and plastic products has declined. The sentiment of the family topic fluctuates, but it is neutral. Public sentiment towards plastic products changed from negative in 2013 to positive in 2014 and gradually became neutral.

5 Conclusion

We proposed a topic extraction and classification framework on obtaining and analysing public perceptions based on social media. Based on this framework, we

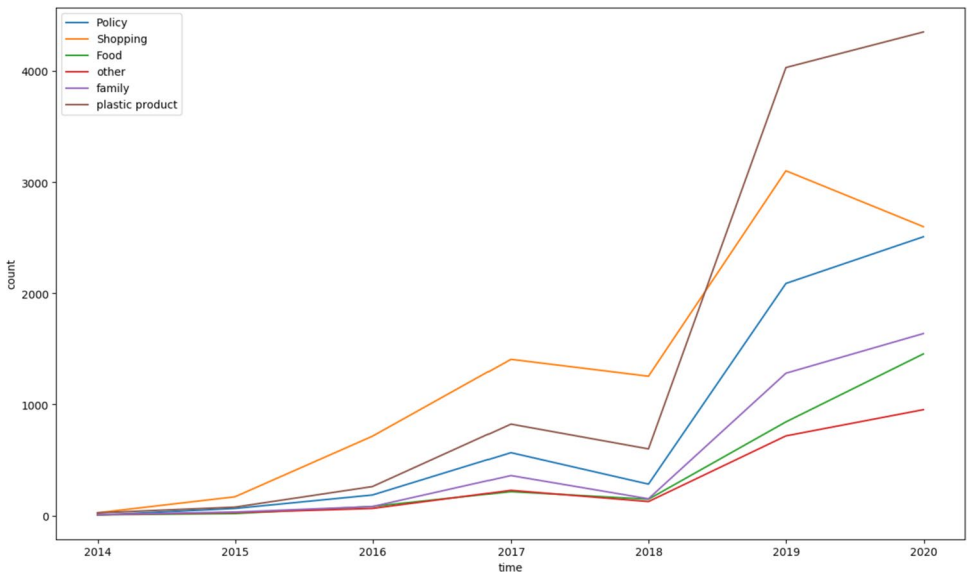

Fig. 9 Related first-level topic time series

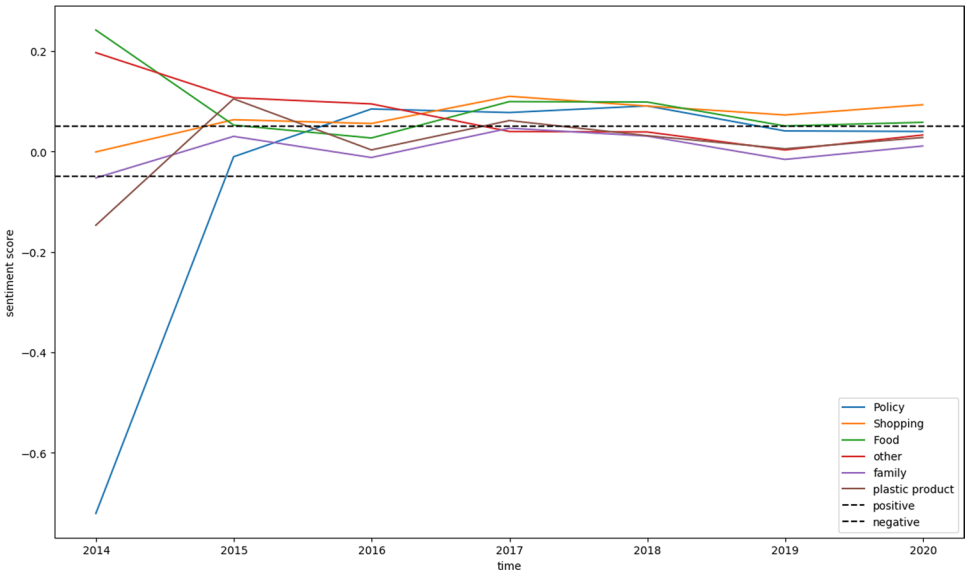

Fig. 10 Related sentiment of first-level topic time serie

comprehensively analysed the social media data from people's opinions about plastic products. The evaluation results show that the topic extraction and classification method proposed in this paper are feasible. At the same time, combined with the emotional score of social media, it is analysed in the time dimension, and the results can find the changes of emotion under the classification.

After comparing the algorithm results of different topic models, we chose a more suitable LDA model to classify comments by comparing the two topic classification model indicators of silhouette efficiency and coherence score. Combined with the emotion evaluation score, we classify the sentiment under the theme classification. The classification results show that the number of negative comments from the public is more than that of positive comments. Finally, in combination with time series analysis, it is found that the discussion on plastics is increasing year by year. However, the public's attitude towards the classification of plastic topics has gradually become more peaceful and neutral.

From the perspective of sentiment scores, public comments generally believe that plastic products' lowest score means that plastics have more room for improvement in products. According to the proportion of comments classified by emotion, there are more negative comments, and the public is more negative about the policy. At the same time, the research results show that although people have paid more attention to plastic-related information since 2018 (the number of comments has increased sharply), the public's emotional expression has shown varying degrees of decline. This may be because in 2018, the U.K. government issued a new Resources and Waste Strategy dedicated to reducing plastic waste. As a result, people's attention to plastics' current situation has increased. More problems have been discovered, which has led to more negative expressions. The problem of plastics has attracted increasing attention, resulting in a great increase in public attention and sympathy for environmental analysis.

Acknowledgements I would like to thank my supervisor, Chandrasekhar, for his guidance throughout every stage of the process. I would like to thank Professor Nishikant for providing the initial clues to the research data.

Author Contributions YX: conceptualization, methodology, format analysis, software, validation, writing—original draft, visualization, writing—review and editing. CK: data curation, writing—review and editing, supervision, project administration. YC: data curation, writing—review and editing, supervision, project administration. NM: data curation, collection of data sources. NW: result analysis, writing—review. PD: writing—review and editing, project administration. All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by YX. The first draft of the manuscript was written by YX and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding The authors declare that we have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Availability of Data and Materials All the data underlying this article will be shared on reasonable request to the corresponding author.

Declarations

Conflict of Interest No conflict of interest exits in the submission of this manuscript, and manuscript is approved by all authors for publication. I would like to declare on behalf of my co-authors that the work described was original research that has not been published previously, and not under consideration for publication elsewhere, in whole or in part. All the authors listed have approved the manuscript that is enclosed.

References

1. Stahel, W.R.: The circular economy. Nature **531**(7595), 435–438 (2016)
2. Ellen MacArthur Foundation, McKinsey Center for Business and Environment. Growth within: a circular economy vision for a competitive Europe[M]. Ellen MacArthur Foundation (2015)
3. Schnurr, R.E.J., Alboiu, V., Chaudhary, M., et al.: Reducing marine pollution from single-use plastics (SUPs): a review. Mar. Pollut. Bull.Pollut. Bull. **137**, 157–171 (2018)
4. Dilkes-Hoffman, L.S., Pratt, S., Laycock, B., et al.: Public attitudes towards plastics. Resour. Conserv. Recycl.. Conserv. Recycl. **147**, 227–235 (2019)
5. Breznau, N.: Positive returns and equilibrium: simultaneous feedback between public opinion and social policy. Policy Stud. J. Stud. J. **45**(4), 583–612 (2017)
6. Mangold, W.G., Faulds, D.J.: Social media: the new hybrid element of the promotion mix. Bus. Horiz.Horiz. **52**(4), 357–365 (2009)
7. Fischer, E., Reuber, A.R.: Social interaction via new social media:(How) can interactions on Twitter affect effectual thinking and behavior? J. Bus. Ventur.Ventur. **26**(1), 1–18 (2011)
8. Williams, R.L., Cothrel, J.: Four smart ways to run online communities. MIT Sloan Manag. Rev.Manag. Rev. **41**(4), 81 (2000)
9. Margetts, H.: Political behaviour and the acoustics of social media. Nat. Hum. Behav.Behav. **1**(4), 1–3 (2017)
10. Laroche, M., Habibi, M.R., Richard, M.O., et al.: The effects of social media based brand communities on brand community markers, value creation practices, brand trust and brand loyalty. Comput. Hum. Behav.. Hum. Behav. **28**(5), 1755–1767 (2012)
11. Sparks, B.A., Browning, V.: Complaining in cyberspace: The motives and forms of hotel guests' complaints online. J. Hosp. Market. Manag.Manag. **19**(7), 797–818 (2010)
12. Sundaresan, N., Zhang, Y., Baudin, C., et al.: System and method for topic extraction and opinion mining: U.S. Patent 8,533,208, 10 Sept 2013
13. Hidayatullah, A.F., Ma'arif, M.R.: Pre-processing tasks in Indonesian Twitter messages. J. Phys. Conf. Ser. **801**(1), 012072 (2017)
14. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
15. Romaszko K.P.: Automatic detection and visualization of topics in large text data sets. Zakład Projektowania Systemów CAD/CAM i Komputerowego Wspomagania Medycyny (2018)
16. Bakshi, R.K., Kaur, N., Kaur, R., et al.: Opinion mining and sentiment analysis. In: 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom). IEEE, pp. 452–455 (2016)
17. Hutto, C.J., Gilbert, E.: Vader: a parsimonious rule-based model for sentiment analysis of social media text. In: Eighth international AAAI conference on weblogs and social media (2014)
18. Chowdhary, K.R.: Natural language processing. In: Fundamentals of Artificial Intelligence, pp. 603–649. Springer, New Delhi (2020)
19. Jianqiang, Z., Xiaolin, G.: Comparison research on text preprocessing methods on twitter sentiment analysis. IEEE Access **5**, 2870–2879 (2017)
20. Priyantina, R., Nopember, I.T.S., Sarno, R., et al.: Sentiment analysis of hotel reviews using latent Dirichlet allocation, semantic similarity and LSTM. Int. J. Intell. Eng. Syst. **12**(4), 142–155 (2019)
21. Luo, L.: Network text sentiment analysis method combining LDA text representation and GRU-CNN. Pers. Ubiquit. Comput.Ubiquit. Comput. **23**(3–4), 405–412 (2019)
22. Hemmatian, F., Sohrabi, M.K.: A survey on classification techniques for opinion mining and sentiment analysis. Artif. Intell. Rev.. Intell. Rev. **52**(3), 1495–1545 (2019)
23. Jongeling, R., Datta, S., Serebrenik. A.: Choosing your weapons: On sentiment analysis tools for software engineering research. In: 2015 IEEE International Conference on Software Maintenance and Evolution (ICSME). IEEE, pp. 531–535 (2015)
24. Ofcom.org.uk. [online] Available at: https://www.ofcom.org.uk/__data/assets/pdf_file/0027/157914/uk-news-consumption-2019-report.pdf (2021). Accessed 16 Nov 2021
25. Do, H.H., Prasad, P.W.C., Maag, A., et al.: Deep learning for aspect-based sentiment analysis: a comparative review. Expert Syst. Appl. **118**, 272–299 (2019)
26. Hidayatullah, A.F., Aditya, S.K., Karimah, S.T.G., et al.: Topic modeling of weather and climate condition on twitter using latent dirichlet allocation (LDA). IOP Conf. Ser. Mater. Sci. Eng. **482**(1), 012033 (2019)
27. Řehůřek, R., Sojka, P.: Gensim—statistical semantics in python. Retrieved from genism.org (2011)
28. Maayah, B., Arqub, O.A.: Hilbert approximate solutions and fractional geometric behaviors of a dynamical fractional model of social media addiction affirmed by the fractional Caputo differential operator. Chaos, Solitons Fractals: X **10**, 100092 (2023)