Information Sciences 610 (2022) 215–234

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Leveraging multidimensional features for policy opinion sentiment prediction

Wenju Hou [a], Ying Li [a,*], Yijun Liu [b,c], Qianqian Li [b,c,*]

[a] College of Computer Science and Technology, Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, China
[b] Institutes of Science and Development, Chinese Academy of Sciences, Beijing, China
[c] School of Public Policy and Management, University of Chinese Academy of Sciences, Beijing, China

ARTICLE INFO

Article history:
Received 10 February 2022
Received in revised form 30 July 2022
Accepted 1 August 2022
Available online 5 August 2022

Keywords:
Policy opinion
Sentiment prediction
Deep learning
Feature engineering

ABSTRACT

Previous online policy opinion analyses based on social media data have focused on topic detection and sentiment classification of policy opinion after a given period following policy implementation. These approaches are limited and inefficient because they provide no opportunity to change citizens' opinions once they have been formed. Furthermore, incorporating auxiliary information to enrich semantic representations is vital and challenging due to limited texts, and a lack of both semantic information and strict syntactic structure. Therefore, we propose a novel framework to extract and integrate multidimensional features from user-related and policy-related social media information and predict policy comment polarity in the policy release phase. First, we construct four machine learning models for model-induced features to capture topic-related and opinion-related features and identify the policy-opinion nexus. In addition, we integrate basic and behavioral user features. Then, we leverage multidimensional features to construct a stacked learning model for predicting the policy opinion. Finally, we conduct experiments on 20 policy comment datasets to demonstrate that our prediction framework can effectively predict public opinion about a policy once it is released. Our model provides key insights into policy opinions in advance and can enable policymakers to engage in better policy communication before opinion formation.

1. Introduction

Public opinion refers to people's attitudes about an issue when they are members of the same social group. Public opinion on policy issues represents citizens' preferences for certain policies. Such opinions are an important driver of public policy change. Thus, keeping different from other opinions, public preferences will be reflected in policy. Most researchers concur that public opinion may influence public policy and government decision-making and that policy opinion can also be a predictor of policy. A strong positive relationship and congruence between opinion and policy have been found [1]. In the information age, the internet has become an important outlet for members of the public to voice their opinions on public affairs. Specifically, social media platforms are an important channel for communication between governments and citizens [2]. Social media has shifted the ways citizens engage in the policy-making process [3], as citizens can use social networks to

* Corresponding authors.
E-mail addresses: liying@jlu.edu.cn (Y. Li), lqqcindy@casisd.cn (Q. Li).

https://doi.org/10.1016/j.ins.2022.08.004
0020-0255/

express their opinions on policy issues. Public opinion can spread rapidly and broadly, allowing policymakers to quickly capture public reactions to a policy. To improve their policy-making and problem-solving abilities, policymakers attach more importance to opinions expressed on social networks.

Citizens' responses to policies in social networks have attracted considerable attention with regard to sentiment distribution and policy topic detection [3]. Relevant studies have focused mainly on extracting, classifying, and analyzing the majority opinion from a large volume of text data. These policy-opinion analyses have been conducted *after a given period of time following policy implementation*. Once a public opinion has been formed, it is difficult to change it. Therefore, these works are limited and inefficient in terms of practical application.

To address the aforementioned problems, we aim to predict citizens' opinions of a policy *once it is released*. McGregor [4] revealed that when a policy is released, public opinions emerging on social media platforms are time-sensitive and allow the possibility of perceiving policy opinions at a very early stage. Rasmussen [1] verified the predicted coefficients of public opinion on the policy for specific issues. For example, on the issues "military in Afghanistan" and "adoption by same-sex couples", the policies are clearly strongly related to public opinion. Thus, the more accurately and earlier decision-makers understand public responses to a policy, the more effective policy response and better communication strategies they can devise.

However, there are two major challenges in predicting the sentiment of policy opinion on social media platforms. The first challenge is the semantic sparsity and lack of syntactic structure due to the limited length of a short text in social media data. When a policy is released, government microblog accounts publish policy microblogs to inform the public. Despite the cancellation of the 140-character limit in 2016, the long-standing restriction has had a significant impact on the writing style and usage habits of users on Weibo, a social media platform in China that is similar to Twitter. The limit is still in effect in some situations, such as commenting and retweeting. Therefore, policy microblogs cannot embody policy topic information sufficiently. In addition, not all policies receive much attention [5]; thus, acquiring voluminous domain-rich text is problematic. To address policy microblogs' semantic sparsity, we introduce government microblogs [2] to increase semantic information. Government microblogs are social text data of political affairs over a period that contain a large amount of political information.

The second challenge is how to make the connection between citizens' opinions and policies. That is, when a policy is released, how can we determine who will be interested in it and what kinds of emotional comments will be published? To address this challenge, we are inspired by the research of Kosinski [6], who revealed associations between user attributes and users' digital records of behavior on social media. As a result, we aim to identify clues regarding the policy-opinion nexus from user digital record information (such as publishing information and profiles) on social media. A user's historical posting information on social media could disclose his or her interests in a policy, and a user's profile information is related to his or her level of engagement and social status (e.g., activity and authority).

It is worth mentioning that our research does not focus on whether a user has publishing behavior. We focus on predicting the user's opinion sentiment (positive, neutral, negative) toward the policy. Even though a user may not publish comments on a policy, he or she still holds a stance on the policy. Comments on policy microblogs are the explicit thoughts of citizens. Thus, we can utilize these comments to verify our research framework.

The contribution of our research is the proposed deep learning framework for extracting and integrating multidimensional data from user-related and policy-related social media information. To the best of our knowledge, this is the first work to integrate multidimensional features from social media text data for public opinion prediction regarding policy in the policy release phase. We introduce multidimensional features, including model-induced features and basic and behavioral user features, to incorporate various types of auxiliary information and enrich the semantic representations. The model-induced features, including policy microblog topic distribution, user interest topic distribution, user-policy opinion distribution and user emotional status distribution, are extracted from our four constructed machine learning models based on the latent Dirichlet allocation (LDA) topic model and convolutional neural network (CNN). The final prediction model of public opinion on policy is constructed by a hierarchical ensemble learning model that integrates multidimensional features, including model-induced features, user basic features, and user posting behavioral features.

To validate our proposed model, we conduct comprehensive experiments on 20 popular policy datasets on Weibo. The experimental results show that integrating multidimensional features improves the highest accuracy of 68.16 % using single-class user basic attribute features to 73.93 %. The accuracy rate of a stacked learning model of multidimensional features is 82.12 %. The results illustrate that our proposed framework can rapidly and accurately detect policy opinions in advance and identify the policy-opinion nexus from social media data.

The rest of the paper is organized as follows: Section 2 introduces related works. Section 3 describes the process of data collection. Section 4 formalizes the proposed method. Section 5 shows the experimental results. Finally, section 6 presents our discussion and conclusions.

2. Related works

2.1. Behavior prediction based on social network information

As social interactions increasingly happen on social media platforms, the scale, richness, and temporality of such data have made microblogs a popular resource for behavior prediction, ranging from economic activities to political phenomena

[7,8]. On Facebook, user interaction data such as likes, comments, and shares of posts have been used to construct models for user behavior prediction, personality prediction, and even suicide risk detection [9,10]. There is also extensive literature on inferring user characteristics and preferences, including gender, sexual orientation, and political opinions, which are often regarded as sensitive attributes. Text, images, personal profiles, social network structures, and even account names, as well as combinations of these elements, can be useful features for attribute inference tasks [11,12]. Most of this type of research has focused on mining business behavior to predict user interests and purchasing behaviors, as well as personalized product recommendations based on user behavior records on e-commerce platforms or multidimensional data on social platforms, such as clicking, purchasing, reviewing, and structure information [13,14]. Matz [15] characterized user extroversion and introversion to successfully predict user purchase behavior by pushing customized beauty products to female Facebook users.

In the political participation area, opinion mining can lead to a better understanding of a user's political behavior [8]. According to Chauhan et al.'s review of studies that aimed to infer the political stance of online users, researchers have used sentiment-oriented content mostly to predict election results and assess the stances of political campaigns [16]. Brito and Adeodato [17] focused on the repercussions of the posts of official candidates' profiles in major social media and combined traditional polls to train machine learning models individually for each candidate to predict their vote share.

2.2. Sentiment prediction techniques

The goal of sentiment prediction is to automatically identify the polarity (positive, negative, or neutral) of human opinions, attitudes, or sentiments toward a topic in a given text [18]. Owing to the enormous amount of user-generated content, a significant amount of research has focused on sentiment prediction, and these studies can be divided into lexicon-based, machine learning-based, and hybrid methods. Lexicon-based methods rely on sentiment lexicons, which contain words with a fixed sentiment polarity. In this type of method, an annotated corpus is not necessary. However, this approach suffers from sentiment fluctuation in different semantic contexts [19] and word coverage of sentiment words. Machine learning-based methods overcome the limitations of extracting effective features from the text and usually obtain better performance. Different machine learning algorithms are employed for sentiment classification, with the support vector machine (SVM) considered to be the most effective and popular [20]. Phan et al. [21] achieved good results by applying LDA and fuzzy decision trees for sentiment analysis and decision-making support. Ali et al. [22] proposed an ontology-based and LDA-based word embedding system called topic2vec, which improved the effectiveness of machine learning classifiers for sentiment classification. With the development of deep learning, CNN and LSTM, as well as their variants, have been frequently used in sentiment prediction recently due to their feature learning capabilities [23,24]. Abdi et al. [25] utilized word embedding, sentiment, and linguistics knowledge features as input to RNN-LSTM, which uses multifeature fusion to classify users' opinions stated in reviews. Gan et al. [26] presented a multichannel CNN-BiLSTM model with the attention mechanism that can extract both multiscale and original high-level context features.

2.3. Online policy opinion mining analysis

There are mounting concerns over the possible linkage between public opinion and public policy. From the perspective of the public value paradigm, policymakers should focus not only on the economic objectives but also on fulfilling "the needs and wants of collective citizenry" [27]. Currently, online policy opinion mining models can provide significant useful insights into the benefits expected from public policies. In comparison to traditional public opinion surveys, the advantages of online policy opinions are that they are more spontaneous and can reveal the realistic concerns of citizens about policy in real time [28]. Opinion mining techniques can be used in different areas of policy cycles, including agenda setting (tracking citizens' concerns about policy proposals), policy formation (extracting valuable opinions about policy proposals and performing sentiment analysis of them), and policy implementation and evaluation (recognizing the sentiments of opponents and proponents in evaluating the effectiveness of a policy) [29]. The most common opinion techniques used in policy opinion analysis are topic models, word frequencies, and sentiment analysis. Belkahla et al. [3] proposed a generic framework based on latent semantic analysis (LSA) to identify discussion keywords and topics on diverse policy subjects in social networks. However, this framework needs human resources to ensure the accuracy of subject identification, which is impossible when dealing with hundreds of thousands of texts. Hagen et al. [30] applied topic modeling to identify and visualize emerging topics for the improvement of policy problem definition and agenda setting. Additionally, they found that computational and visual analytics have potentially positive impacts on policy-making. Dahal et al. [31] applied LDA to infer topics of discussion and sentiment analysis to assess the overall attitude and feelings in comparing the nature of discussions on climate change and related policy concerns among various nations on social media.

3. Dataset description

Considering the differences among the reviewers of Weibo accounts, we selected policy microblogs from one Weibo account. The more important an account is, the higher its user participation level. For this reason, we chose the Weibo account of人民日报 (Renmin Daily), which is the largest authoritative newspaper in China. Then, we randomly selected 20

policies that raised discussions on social media from 03/2018 to 05/2019. The policy information is shown in Table 1. Sometimes more than one microblog was published for an individual policy; as a result, we collected 47 policy microblogs from the Weibo account of Renmin Daily for the 20 policies.

For each policy microblog, we collected the comments and user basic information. However, because crawling all historical posting information is very costly, we selected key users by means of user profile information, the detailed process of which will be presented later. In addition, historical microblog posts of key users were crawled from each user's homepage. Finally, an auxiliary corpus, government microblogs on Weibo, was also crawled. Through the above process, we obtained policy-related information and a user-related database. The policy-related information included policy microblogs and government microblogs. The user-related information included comments on public policy, user information (key users), and historical microblog posts (key users). Fig. 1 briefly introduces the data collection process for constructing the policy-related and user-related datasets. The details of our collected dataset are listed in Table 2.

3.1. Policy microblogs

Social networks have been the main platform for citizen engagement. According to the research of Boudjelida [32], 41 % of citizen participation activities are implemented through social media platforms. For policymakers, a microblog is a convenient tool for informing the public and facilitating policy communication [33]. When a policy is released, the official corporate Weibo account and the media Weibo account initially publish policy-related microblogs. An example of a policy microblog is shown in Fig. 2. For the 20 policies presented in Table 1, 47 policy microblogs were collected.

3.2. Comments on public policy

Once a policy microblog is released, users' activities concerning the policy are mostly posting new microblogs, retweeting microblogs, commenting on microblogs, liking microblogs, and so forth. We principally investigated replies to the policy microblog since they are an explicit expression of the policy. Hence, we collected comments from each policy microblog. Finally, we built a dataset with 83,085 policy comments.

3.3. Key user information extraction

Key users are located in this paper based on user activeness and social position, which are significant predictors of key users. According to Zengin [34], when considering user activeness on a specific topic, influential interaction and daily social media usage frequency are both positively connected with key social media users' activeness. Two measurements are employed to measure user activeness. One is the number of user comments, indicating the participation level in the discussion of a policy. The other is the total number of microblogs posted on a user's personal homepage, indicating the user's daily activeness. Social position refers to a user's opinion leadership [35]. We used the number of followers and comment popularity to measure a user's social position. The number of followers represents those who follow a key user, which is a metric that has been used in previous research to measure social position [35]. Comment popularity is the measurement of a user's influence on the policy opinion. By default, the Weibo platform displays comments from high to low popularity according to the number of comments liked, which means that comments with more likes will have more chances of being seen by other users. Then, we define the corresponding thresholds to locate key users: {The number of user comments is greater than 2}∩

Table 1
Policy information.

	Policy Information	Release Date		Policy Information	Release Date
1	Preferential admission to college for students from poor families over other students with the same conditions.	2018.03	11	Institution of paid leave.	2018.10
2	Charges for domestic garbage disposal.	2018.07	12	Raise the supply of inclusive preschool resources.	2018.11
3	Housing incentives for talented students.	2018.07	13	An ID card photo can be chosen from multiple photos.	2018.12
4	The average salary of compulsory education teachers cannot be lower than that of local civil servants.	2018.07	14	Cancellation of infusions for adult outpatients.	2019.01
5	Primary and secondary school teachers cannot be employed for off-campus tutoring.	2018.07	15	Adult female applicants cannot be asked about marital and childbirth status during recruitment.	2019.02
6	Lowering prices of anticancer drugs and ensuring supply.	2018.07	16	Hospitals should build a fantasy castle for use to allay the fears of children seeking health care.	2019.02
7	New college entrance examination program in Beijing.	2018.08	17	Heads of primary and secondary schools and kindergartens should take meals with students.	2019.03
8	New personal income tax law.	2018.08	18	Lowering ticket prices of scenic areas.	2019.03
9	Increasing broadband speed and lowering fees.	2018.09	19	Programs for minors are not allowed to spread puppy love.	2019.04
10	Rectification of ride-sharing services.	2018.09	20	No deposit for shared cars and bicycles.	2019.05

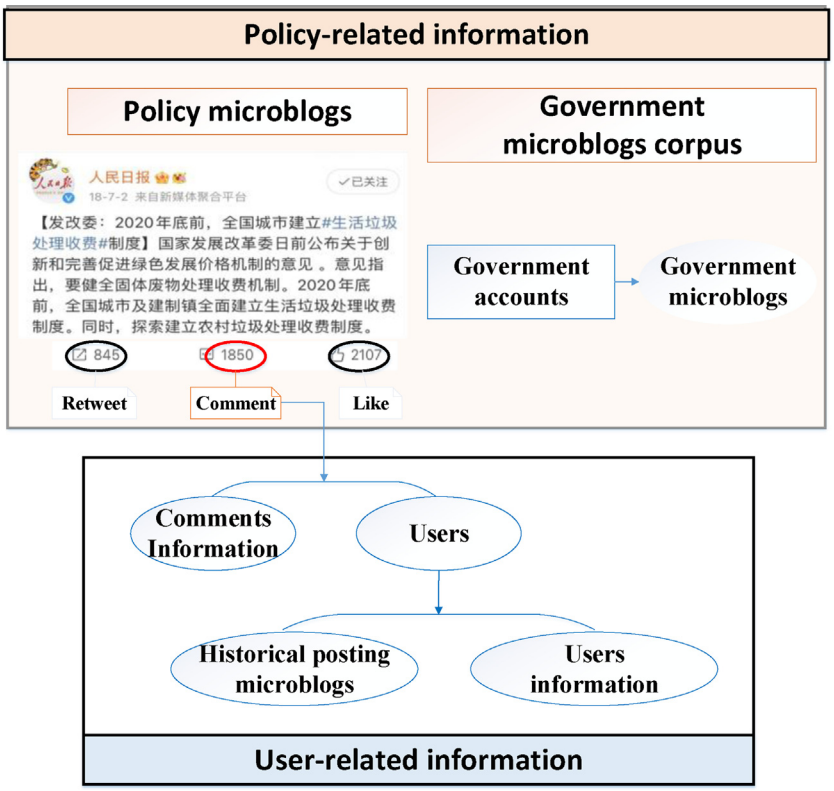

Fig. 1. Policy-related and user-related data collection. All selected policy microblogs were collected from the Weibo account of Renmin Daily. Comments on policy microblogs were crawled and used to filter key users and then to obtain users' historical postings and profile information. Government microblogs were used as an auxiliary corpus.

Table 2
Statistical description of the dataset.

Data Type	Dataset Name	Volume
Policy-related information	Policy microblogs	47
	Government microblogs	87,432
User-related information	Comments on public policy	83,085
	User profile information (key users)	13,390
	Historical microblog posts (key users)	1,904,777

{The total number of microblogs posted on the user's personal homepage is greater than 150}∩{The number of comments liked is greater than 1}∩{The number of followers is greater than 100}. Consequently, 13,390 key users are selected. Then, the basic information of these users (gender, verified status, location, number of followers, etc.) was crawled from their homepages.

3.4. Historical microblog posts (key users)

Understanding users' interests and intentions by tracking their digital behavior has been shown to be effective [7]. Due to the high crawling cost of acquiring users' historical microblog posts, we exploited the strategy of focusing on key users. For the key users selected as described in 3.3, 150 microblogs published on their Weibo homepages were crawled. The crawled information includes microblog content and the number of retweets, comments, and likes. Finally, we obtain 1,904,777 historical microblog posts.

3.5. Government microblogs

Government microblogs are microblogs published by official agencies to convey information about public affairs. Government microblogs improve the efficiency of interaction between the government and citizens [2]. They are published mostly by mainstream media accounts and official agency accounts, which play an important role in policy information dissemina-

Fig. 2. A policy release microblog example on Weibo.

tion and have a great influence on citizens. We collected 87,432 microblogs from media accounts (e.g., CCTV News) and official agency accounts that were used by Xiong [2]. However, not all government microblogs are policy-related. Therefore, we first construct a policy judgment model to discriminate between policy-related and non-policy-related microblogs. The model details are described in the methodology section. Based on the policy judgment model, we extracted 4,186 policy-related information.

4. Proposed method

In this section, we provide a detailed description of our proposed model. The research framework of our model is shown in Fig. 3. The main work includes four steps: data crawling, data preprocessing, multidimensional feature extraction and policy opinion sentiment prediction.

- Data crawling and data preprocessing

First, we construct policy-related and user-related datasets by crawling policy microblogs, government microblog corpora, comments on policy microblogs, historical user microblogs, and user information from social media. Then, we apply natural language processing techniques, including word segmentation, word frequency matrices, and word embedding to preprocess the collected texts.

- Multidimensional feature extraction

Based on the policy-related and opinion-related datasets, we extract multidimensional features, including model-induced features and basic and behavioral user features, to explore the comprehensive and deeper user-policy features. Model-induced features refer to topic-related and opinion-related features extracted from tailored models. Topic-related features are extracted from the user interest topic model and policy microblog topic model based on the LDA algorithm. Opinion-related features are extracted from the opinion sentiment classification model and policy judgment model based on the CNN algorithm. In addition to the model-induced features, some descriptive features are incorporated. These are user basic features crawled from user homepages and user posting behavioral features from historical user microblog statistics.

- Policy opinion sentiment prediction

Finally, policy opinion sentiment prediction is converted into a classification problem (negative, neutral, positive). The prediction model constructs a hierarchical stacked ensemble learning model by integrating single-feature learners, multifeature learners, and a stacked learning model.

4.1. Prediction task statement

Predicting opinion sentiment on public policy can be defined as follows: for a given released policy microblog, the task is to determine what kind of opinion sentiment a user would publish regarding the policy. This problem can be formulated as a

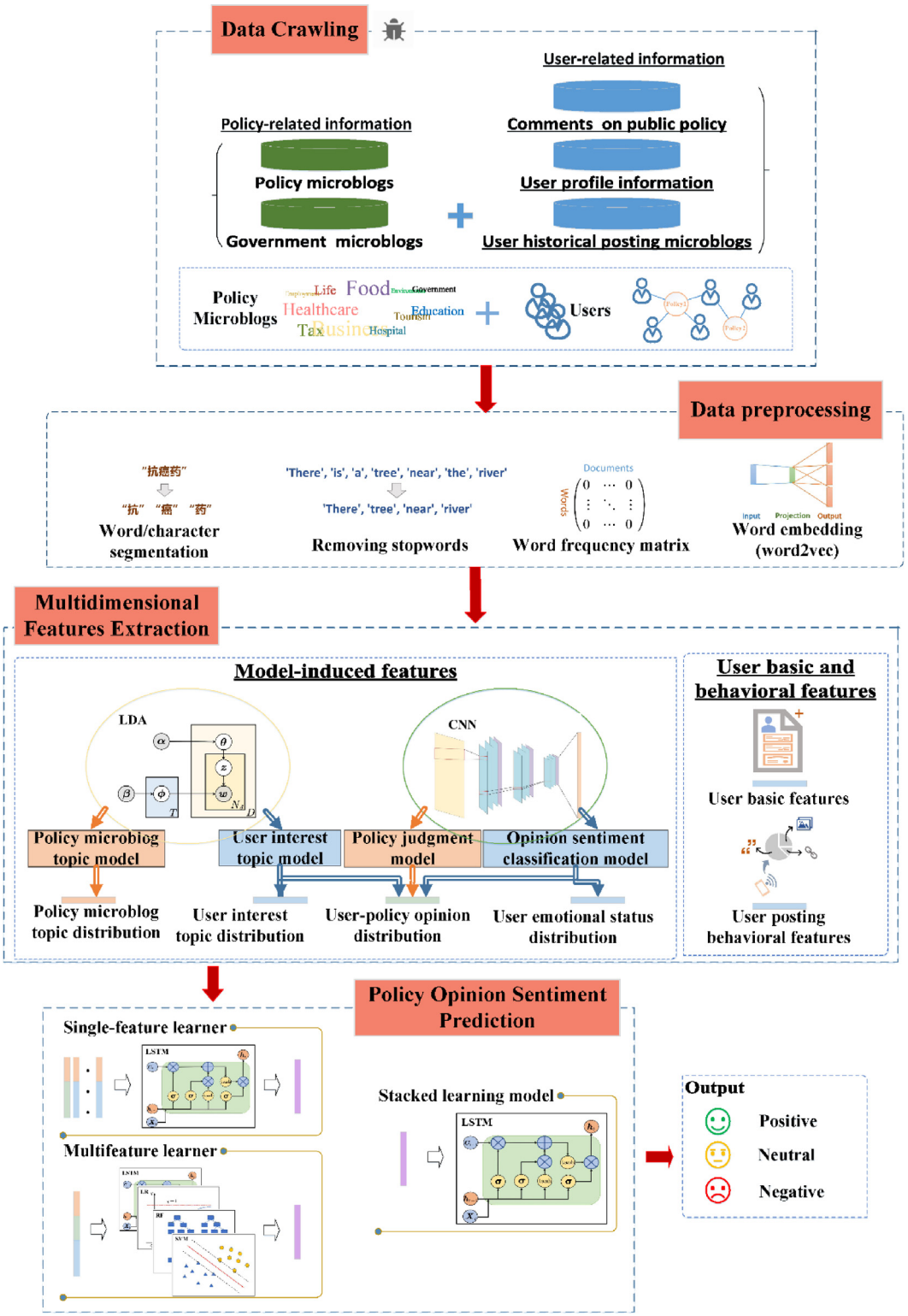

Fig. 3. Research framework. Policy-related and user-related information is collected from Weibo. Extracted multidimensional features include model-induced features, user basic and posting behavioral features. Then, the policy opinion prediction task is conducted based on a hierarchical stacked learning model.

classification problem as follows: let $pw = \langle pw_1, \cdots, pw_K \rangle$ be a sequence of words in a policy microblog pw with length K. The user u's available information on social media, including *historical user microblog posts, user basic information* and *user behavioral information*, are collected. Assume that user u will publish a comment on the policy microblog. The goal is to predict user u's sentiment toward policy pw. The point is the user's attitudes toward the policy rather than the sentiment strength. Hence, this prediction task is transformed into a three-class classification problem with labels u_p={negative, neutral, positive}.

4.2. Data preprocessing

For our collected corpora, we conduct the basic data preprocessing steps, including word/character segmentation, removing stopwords, and tokenizing. Two different segmentation strategies are applied. Word segmentation is used in topic models because it can maintain the accuracy and richness of the semantics. For classification models, character segmentation is used to achieve higher accuracy [36]. Stopwords and unnecessary information, such as emoticons, links, and other nonsense items, are removed. For the topic model, punctuation marks are also removed as useless information. The remaining text is converted into bags of words in the form of a word frequency matrix. When the frequencies of words are counted, words that occur infrequently are also removed. For CNN-based text classification tasks, tokenizing comprises assigning a serial number to each word/character, transforming the text into a sequence of integers, and cropping the text to a uniform length. Punctuations are retained in the classification tasks due to some punctuation marks expressing emotion, such as "?" for doubt and "!" for exclamation. Word embeddings and character embeddings are trained by word2vec. In the word2vec algorithm, the dimension parameter is set to 300, the context window size is set to 8, and the minimum word count is set to 10.

4.3. Multidimensional feature extraction

Multidimensional features include the user interest topic distribution, policy microblog topic distribution, user emotional status distribution, user-policy opinion distribution features, user basic features, and user posting behavioral features. The first four types of features are categorized as model-induced features. The user interest topic distribution and the policy microblog topic distribution are topic-related features derived from the topic model, which can disclose the interpretable topics of user interest and policy. Opinion-related features extracted based on the classification models include user emotional status and user-policy opinion distribution features, which can reveal the user's psychological assessment of and relevance to the policy. Other user features, such as user basic information and user historical post statistics, are added to indicate user involvement attributes on social media. The overall schematic diagram of feature extraction and sentiment prediction is shown in Fig. 4.

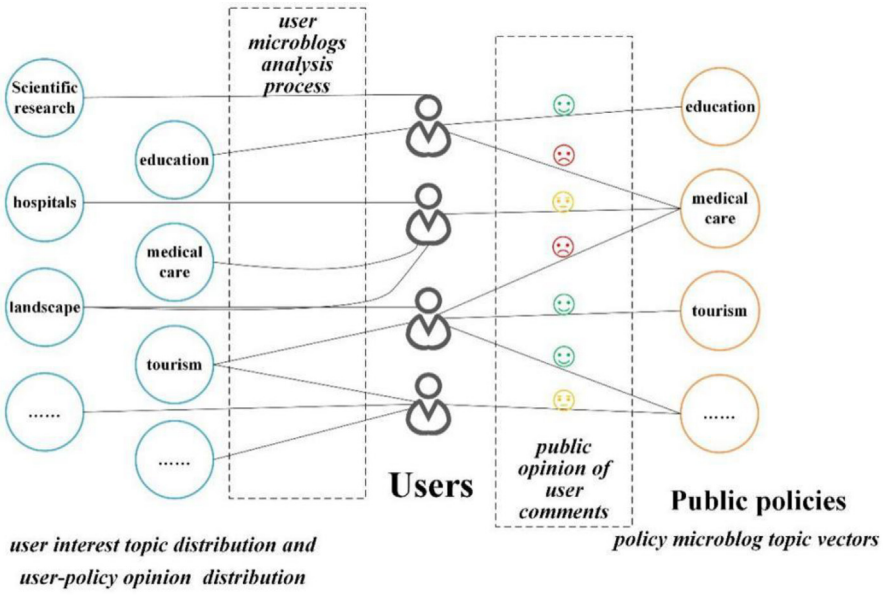

Fig. 4. Schematic of the feature extraction and prediction task. Distributions of users on interest topics and policy topics can be constructed using all of a user's microblogs. In turn, the policy microblogs commented by users can be interpreted as distributions on policy topics.

4.3.1. User interest topic model and user interest topic distribution

LDA, a typical topic model, is widely used in information retrieval and natural language processing. It represents a document as a bag of words, describes a topic as a concept, and expresses a topic as a series of related words and their probability distribution. The topic distribution for each document and word distribution for each topic are both assumed to follow Dirichlet distributions.

A user's historical microblog posts are used as training data. As a single microblog cannot sufficiently reflect the overall distribution of a user's interests, all the microblogs of a user need to be stitched into a long text as a training sample. We randomly select 6,000 users from among key users and concatenate the long texts from their microblogs to compose the training set. After preprocessing, a document is transformed into a collection of words. In the model training process, we remove words with frequencies less than 2 as well as words that appear in more than 95 % of all documents. Then, we use the LDA algorithm to construct the user interest topic model. After comparing topic numbers of 50, 100, 150, 200, 300, and 500, we designate the number of topics in this topic model as 200, aiming at a greater cohesion of words within topics and differences between topics.

Through the user interest topic model, a long text concatenated by all historical microblog posts of a user will be represented as a 200-dimensional numeric vector, which is taken as the user interest topic distribution. This user interest topic model can grasp the user's interests through the user's historical microblog information. These features intuitively inform us about what individuals are usually saying and following on social media. This topic vector serves as a contextual backdrop for each of the user's posts, and some of the topics are undoubtedly relevant to policy issues that will be discussed later. Table 3 shows some of the topics and keywords, which can describe users' different interests in daily life.

4.3.2. Policy microblog topic model and policy microblog topic distribution

In contrast to the user interest topic model, the number of released policy microblogs cannot satisfy the requirements of LDA model training, which requires a large amount of text data. For this reason, government microblogs are introduced to alleviate semantic sparsity and develop the ability of policy topics to describe the real situation in various fields of social life. In our policy microblog topic model (**LDA-pol**), we set the topic number to 5–50 in turn. After repeated comparison, we identify 30 as the best topic number. Then, the policy microblog topic model generates a 30-dimensional vector representing the released policy microblog topic distribution. Policy microblog topic distribution reflects the overall semantic information about a policy. The policy topic, likewise, informs us about what this policy states and what the concerns are. The content and sentiment encoded in the policy text, as the target of user comments, has a significant effect on the sentiment opinions that people may develop [37,38]. Simultaneously, policy topic distribution is employed to distinguish among various policies.

4.3.3. Opinion sentiment classification model and user emotional status distribution

The opinion sentiment classification model is built using a three-layer 1D CNN. The length of each microblog is sliced into 50 characters (i.e., the first 50 characters are retained, and a 0 is added to texts with lengths below 50). Each one-dimensional convolutional layer is followed by a ReLU and a 1D maxpooling layer, and the feature maps are flattened and concatenated before the fully connected layer. We train the whole network to minimize the cross-entropy with the Adam optimizer. The dropout rate is set as 0.4, and each incremental training is performed for 10 epochs. We formulate the sentiment distribution of one microblog with a 5-point Likert scale, where 1 indicates strongly negative and 5 represents strongly positive. We manually construct an emotion annotation as the ground truth. Examples for the sentiment classification of each microblog are shown in Table 4.

Then, we train the opinion sentiment classification model (**CNN-opi**). The training is synchronized with the correction and enhancement of opinion annotations of the microblogs. In addition, we extract an expanded dataset including only strong sentiment (strongly negative and strongly positive) to manage the class imbalance due to the lack of strong sentiment samples. In the **CNN-opi** model training process, we first minimize the inconsistency between the prediction and the ground

Table 3
Keywords from user Weibo topics (excerpts).

Label	Keywords
Politics (international)	International, China, U.S., news, report, strategy, country, Iran, Russia, arms control, agreement, Syria, on-site, Ministry of Foreign Affairs, Chinese government
Fashion (women's)	Fashion, star, matching, style, elegance, black, in style, temperament, fashionable, classic, design, feminine, retro, style, name brand
Fashion (men's)	Fashion week, autumn and winter, men, hot, fashion, men's clothing, matching, fad, exciting content, series, teenager, spring and summer, suit
Affection (family)	Like, children, female students, men, parents, marriage, friends, love, girl, woman, life, in love, male students, mother, broke up
Engineering (energy)	Three Gorges, China, Yangtze River, hydropower station, project, engineering, white crane, power station, wind power, offshore, development, protection, construction, energy
Research	Study, science, teacher, thesis, professor, scientist, China, webpage, student, research, publication, academic, popular science, laboratory
Basketball	NBA, Kobe, James, basketball, tease, assists, rebounds, Lakers, players, three-pointers, CBA, games, zero distance

Table 4
Examples of microblog sentiment annotation.

Label	Example comments	Opinion classification
1	Oppose! Strongly oppose!	Strongly negative
2	[Sad], [Disappointed]	Negative
3	This policy has nothing to do with you!	Neutral
4	Support! [Good], [Congratulatory]	Positive
5	This policy is very good! We must support this!	Strongly positive

truth. Second, the model is used to select some microblogs with strong opinions (strongly negative and strongly positive) from the unlabeled microblogs. Adding strong opinion samples to the expanded dataset is intended to alleviate the class imbalance.

Self-expression on social media can reveal a user's emotional status. In our model, we construct the opinion sentiment classification model to detect a user's daily emotional status from his or her historical microblog posts. What kinds of emotions do users post on social media in everyday life, and are they more positive or negative? We assume that a user's comment on a released policy is related to his or her daily emotional status [15]. A user maintaining a negative attitude in everyday life is more likely to express negative opinions on public affairs. A user's emotional status distribution is the mean of sentiment distribution for that user's historical microblog posts. For example, a user's emotional status distribution of [0.6, 0.1, 0.1, 0.1, 0.1] indicates that this user tends to voice negative opinions. Thus, the probability of this user opposing the policy is relatively high.

4.3.4. Policy judgment model for government microblogs and user-policy opinion distribution

The policy judgment model is devised to identify policy-related microblogs from government microblogs. The architecture of the policy judgment model is a three-layer 1D CNN. We choose government microblogs as an auxiliary corpus enhancing semantics. Based on the average text length, each microblog is sliced into texts of 100 characters. While the policy judgment model (**CNN-pol**) is trained, the selection and labeling of government microblogs progress concurrently. This process is called reverse labeling combined with manual labeling (RLCML), which has the following three steps.

Step 1 **Initialization**. First, we take the released policy microblogs as positive samples and then randomly select 50 microblogs from historical user microblogs as negative samples. These samples compose the initial dataset. We train a simple model for initialization of these microblogs.

Step 2 **Iterative expansion**. In the government microblog dataset, not all microblogs are policy-related. We set an initial threshold $\theta = 0.5$ and use the judgment model to screen policy-related samples from government microblogs. After collecting a certain number of samples using this process, we obtain an expanded dataset in which the newly included microblogs are all labeled positive samples (policy-related microblogs). We correct the labels of samples that are misclassified by the model and retain the expanded dataset. Then, we retrain the model using the extended dataset. In addition, we slightly increase the threshold to reduce manual adjustment. These processes are repeated several times until the number of false positives (samples misclassified by the model as policy-relevant) in the newly collected samples is reduced to zero.

Step 3 **Performance improvement**. After collecting a sufficient number of samples, we focus on improving the performance of the policy judgment model for government microblogs. First, we use the judgment model to assign pseudo labels to all samples in the expanded dataset and screen the samples with inconsistent labels, which yields the inconsistent set $\boldsymbol{I}$. Then, we downsample the negative samples by deleting them randomly due to class imbalance. Finally, the judgment model is retrained. This process is repeated until the inconsistent set is empty. We set the threshold θ to 0.6 based on the final inconsistent set, which is used in subsequent feature extraction. In other words, if a microblog is classified as policy-related information by the policy judgment model, the probability that this microblog is actually classified as policy-related information is greater than or equal to 0.6.

In the following, we address the challenge of building the connection between citizens' opinions and policy. First, we make the following assumptions: every user's historical microblog posts are related to policies, but each microblog has a different degree of relevance. We design an algorithm to calculate the user-policy opinion distribution, which fits a user's support tendency for various policy topics and is divided into two parts. Second, we use the distribution of historical user microblogs on policy topics with relevance and opinion weights to fit users' opinions. The set of a user's historical microblogs is denoted as $\boldsymbol{W}$:

$$W = \{w_1, w_2, \ldots, w_j, \ldots, w_k\} \tag{1}$$

where w_j is a user's historical microblog posts and k is the total number of microblogs. For each microblog w_j, we calculate the *distribution on policy topics*, *emotional labels*, and *political relevance* as follows.

Distribution on policy topics. This is the user's microblog distribution on policy topics, which embodies the user's historical microblog post relevance to the released policy. We decompose each user's microblog w_j into a policy topic vector to represent the content of the policy embedded in this microblog, which is generated by the policy microblog topic model.

The policy topic distribution is denoted as $\boldsymbol{D}$ for microblog w_j, which is formalized as.

$$\boldsymbol{D} = \{p(t_i|w_j)|i = 1, \cdots, n\} \tag{2}$$

where t_i is the i^{th} topic and n is the number of topics.

Emotional label. This is the result ranging from 1 to 5 generated by the opinion sentiment classification model taking a user's microblogs as input and indicating that the opinion changes from *strongly negative* to *strongly positive*. The emotional label is denoted as $\boldsymbol{L}$.

Political relevance. This reflects the degree to which historical microblog posts are relevant to policy. The probability of microblogs being classified as policy related is used as political relevance, which is generated by the policy judgment model for government microblogs. This probability is denoted as $\boldsymbol{P}$.

Then, the user-policy opinion distribution of w_j can be calculated by the product of the microblog semantic distribution on policy topic $\boldsymbol{D}$, emotional label $\boldsymbol{L}$, and political relevance $\boldsymbol{P}$. The user-policy opinion distribution of w_j is denoted as $\boldsymbol{r} = (r_1, \cdots, r_n)$:

$$r_i = P * L * p(t_i|w_j), i = 1, \cdots, n, P \in [0, 1], L \in \{1, 2, 3, 4, 5\} \tag{3}$$

A user's daily microblog distribution on a released policy microblog is the base, and the political relevance is used to assign a weight to each microblog reflecting the policy-related relevance, whereas the emotional label reflects a user's tendency to support a policy.

The above description shows a single historical microblog post's relevance to the released policy. When considering all of a user's historical microblog posts, the situation becomes complicated. When a microblog is retweeted, the content of the retweet and users' replies to the retweet are both collected. A natural assumption is that if a user retweets, the original tweet matches his or her interest, and the user not only approves of the content but also shares a similar opinion to that expressed by the tweet. However, not only the topic itself but also opinions related to the topic might have an impact on a user's retweet decision. A user's opinions are evaluated mainly on the basis of microblogs published by him or her [39]. In this case, we calculate the emotional label of a user's reply, while we use the content of a retweet to calculate the distribution of policy topics and political relevance. Therefore, we divide $\boldsymbol{W}$ into a retweet set $\mathbf{W_r}$ and an original microblog set $\boldsymbol{W_p}$. We differentiate the topic distribution $\boldsymbol{D}$ and probability $\boldsymbol{P}$ in the same way using consistent subscripts. The user-level opinion distribution is the sum of the distributions of all microblogs published by the user, reflecting the policy topics that people pay attention to and talk about and how people feel about discussing these issues. The overall process is described in Algorithm 1, and the calculation formula is shown in Eq. (4) and Eq. (5), where $\boldsymbol{R}$ is a 30-dimensional vector. For each dimension, the larger the value is, the more supportive the opinion tends to be

$$R_i = \sum_{w \in W_p} P_p * L * p(t_i|w) + \sum_{w \in W_r} P_p * L * p(t_i|w), i = 1, \cdots, n \tag{4}$$

$$\boldsymbol{R} = \sum_{w \in W_p} P_p * L * D_p + \sum_{w \in W_r} P_r * L * D_r, i = 1, \cdots, n \tag{5}$$

Algorithm 1

Input: User's historical microblog posts **W**
Require: Policy judgment model for government microblogs **CNN-pol**, opinion sentiment classification model **CNN-opi**, and policy microblog topic model **LDA-pol**
Output: User-policy opinion distribution **R**
1: for w in **W** do
2: if w in $\mathbf{W_p}$ then
3: **r** ←**CNN-pol**(w) * **LDA-pol**(w) * **CNN-opi** (w)
4: else if w in $\mathbf{W_r}$ then
5: *reply*, *retweet* ← User's reply to the retweet, content of the retweet
6: P ← **CNN-pol**(*retweet*)
7: **D** ← **LDA-pol**(*retweet*)
8: L ← **CNN-opi**(*reply*)
9: **r** ←P * **D** * L
10: end if
11: Add **r** to **R**
12: end for
13: **return R**

Algorithm 1. User-policy opinion distribution extraction.

Finally, we calculate the overall attention of the user to the policy, which is a 3-dimensional vector. These vectors are the proportions of microblogs that are classified as policy-related information in $\boldsymbol{W}$, $\mathbf{W_r}$ and $\boldsymbol{W_p}$ respectively. The classification probability is greater than or equal to the threshold of 0.6. These ratios imply how much attention users pay to policy information at a macrolevel. The more attention people pay to policy information, the higher the probability that they will participate in the policy discussion.

4.3.5. User basic features

User basic features, including numbers of followers, followings, published microblogs, account verification status and gender, are obtained directly from each user's profile information. Other information, such as age, job, region, and interest tags, is difficult to utilize effectively due to a large number of missing values caused by registration requirements. The statistical information on the user basic features is listed in Table 5.

4.3.6. User posting behavioral features

User posting behavioral features reflect a user's posting habits, which are the statistics of his or her daily posting behavior. These features consist of the proportion and frequency of retweets, videos, URLs, pictures, etc. In addition, user discussion participation behaviors are also included, such as the use of the mention symbol (@) and the hashtag symbol (#). The mean values of these features for users are shown in Table 6.

Basic and posting behavioral features can both be utilized to determine the user's level of participation in social media discussions. Furthermore, user basic features represent the user's social position, whereas posting behavioral features are the user's habits when using Weibo.

4.4. Policy opinion sentiment prediction

4.4.1. Benchmark models

The final policy opinion sentiment prediction is a two-layer classification model using a stacked ensemble learning strategy. This mechanism can fuse the base learning result generated by different individual learners called the base learner. The base learner is employed to extract the primary feature from the training dataset and output the secondary training dataset input into the secondary learner.

To evaluate the performance of single-type features, in this paper, four machine learning models, including random forest (RF), k-nearest neighbor (KNN), long short-term memory (LSTM), and multilayer perceptron (MLP), are used. These four machine learning models are also used as secondary learners to prevent an excessive combination of features. To leverage the multidimensional features in the prediction task, we apply the 10 machine learning models that are most often used for text analysis as base learners to enhance heterogeneity in order to test the performance of the multidimensional features, which are listed below.

RF: The RF model is an ensemble method employing decision trees as the base learners and combining their predictions by bagging. The RF model has been popular as a result of its excellent practical performance, particularly in high-dimensional situations where it can also be used to filter key information.

XGBoost: This is also a tree-based method and a type of boosting algorithm.

GDBT: The gradient boosting decision tree (GDBT) is an iterative decision tree algorithm consisting of multiple decision trees and is considered to be an algorithm with high generalization capability.

Logistic regression: LR is a type of classification for analyzing the relationship between numerous influencing factors and a binary or categorical outcome.

SVM-RBF: This model is a support vector machine with a Gaussian kernel (SVM-RBF). The SVM model is one of the most commonly used machine learning classifiers for text analysis. As a benchmark, a Gaussian kernel function is taken as the default option.

SVM-POLY: This model is a support vector machine with a polynomial kernel (SVM-POLY). Polynomial kernels are another option for us.

KNN: This is one of the simplest machine learning algorithms. Based on the class of the nearest sample or samples, the approach identifies the class to which the samples to be classified belong.

LSTM: A 3-layer LSTM is constructed due to the limited number of samples and features.

Table 5
Mean values of the user basic features.

Features	#Followers	#Followings	#Microblogs
Gender			
Male	13,557	381	1936
Female	3438	362	1919
Account status			
Verified	51,929	599	5579
Unverified	892	335	1378

Table 6
Mean values of user posting behavioral features.

Posting behavioral feature	Mean value
Proportion of retweets	0.54
Proportion of microblogs containing videos	0.17
Proportion of microblogs containing @	0.17
Proportion of microblogs containing #	0.11
Average times each microblog has been reposted	6
Average times each microblog has been commented on	1.32
Average times each microblog has been liked	6
Average number of pictures per microblog	0.18
Average number of videos per microblog	0.02
Average number of URLs per microblog	0.08
Average number of @ per microblog	0.33
Average number of # per microblog	0.14

MLP: An MLP with a 3-layer structure is capable of handling nonlinearities.
CNN: This CNN includes two layers of fully connected structures and three convolution-maxpooling pairs.
These are fairly traditional models that are used in many similar works [16,20,24,40]. There is much variety in the models in terms of machine learning methods. For stacked ensembles, the heterogeneity of the base learners is crucial to improve the prediction performance of the final ensemble model; namely, the more diverse the individual learners are, the better the prediction result will be [41]. Furthermore, some learners' specific roles are valuable in subsequent analyses, as when RF is used to determine feature importance and filter out significant features.

4.4.2. Model evaluation
The evaluation indicators include the accuracy, precision, recall, and F1 score based on the confusion matrix. These evaluation indicators are defined in Eqs. (6)-(9).

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{6}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{7}$$

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

$$F1 = \frac{2 \times TP}{T2 \times TP + FP + FN} \tag{9}$$

5. Results

In this section, we present the performance of the policy opinion prediction task and compare the feature performances for the base learner and stacked ensemble learning.

5.1. Experimental results

The results of the single-type feature prediction by base learners are shown in Table 7. Fig. 5 displays the average performance of different base learners for single-type features. Single-type features refer to user basic attributes, behavioral features, user interest distribution, emotional status, and association distribution. The association distribution represents the user-policy opinion distribution. Each type of prediction task involves policy microblog topic features; thus, the policy microblog topic features are not listed as a single-type feature. All accuracies of base learners for single-type features are lower than 70 %. The average performance reveals that user basic attributes achieve better performance than other single-type features, while posting behavioral features extracted from text-level metadata perform much worse. This finding may be related to the correlation between the sentiment of the tweets of the user and text-level metadata [42]. For example, personal attributes, such as social status and activity level, have a clear and stable relationship with emotional expression on social media. Furthermore, our experiments reveal that the relationship between the implicit sentiment toward policies in user microblogs and user attributes still exists, which may be a reason for the good performance of the user basic features. However, the specific impact of each type of user attribute on opinions about policies on different topics needs further exploration. By using user basic features, LSTM achieves the best performance, with an accuracy of approximately 68.16 %. Association distribution features achieve better performance than user interest distribution and emotional status.

Table 7
Experimental results of single-type features for the base learner.

Single-type features	Model	Accuracy	Precision	Recall	F1
Basic attributes	RF	0.6452	0.6147	0.6411	0.6311
	KNN	0.6710	0.6027	**0.6689**	**0.6331**
	LSTM	**0.6816**	**0.6337**	0.6574	0.6259
	MLP	0.6672	0.6137	0.6637	0.6326
	Average	0.6663	0.6162	0.6578	0.6307
Behavioral features	RF	0.6356	0.6028	0.6374	0.6123
	KNN	**0.6444**	**0.6123**	**0.6466**	**0.6301**
	LSTM	0.6173	0.5940	0.6322	0.6195
	MLP	0.6112	0.5856	0.6218	0.6085
	Average	0.6271	0.5987	0.6345	0.6176
Interest distribution	RF	0.6442	0.5982	0.6414	0.6240
	KNN	**0.6688**	0.6037	**0.6731**	0.6283
	LSTM	0.6452	**0.6177**	0.6584	**0.6323**
	MLP	0.6607	0.6140	0.6504	0.6256
	Average	0.6547	0.6084	0.6558	0.6275
Emotional status	RF	0.6421	0.6025	0.6433	0.6228
	KNN	**0.6546**	0.5986	**0.6545**	**0.6245**
	LSTM	0.6464	0.6056	0.6318	0.6185
	MLP	0.6473	**0.6118**	0.6433	0.6117
	Average	0.6476	0.6046	0.6432	0.6194
Association distribution	RF	0.6384	0.6017	0.6433	0.6242
	KNN	0.6575	0.6050	0.6589	**0.6293**
	LSTM	0.6557	**0.6135**	0.6608	0.6180
	MLP	**0.6701**	0.6036	**0.7018**	0.6285
	Average	0.6554	0.6059	0.6662	0.6250
Average		0.6502	0.6068	0.6515	0.6240

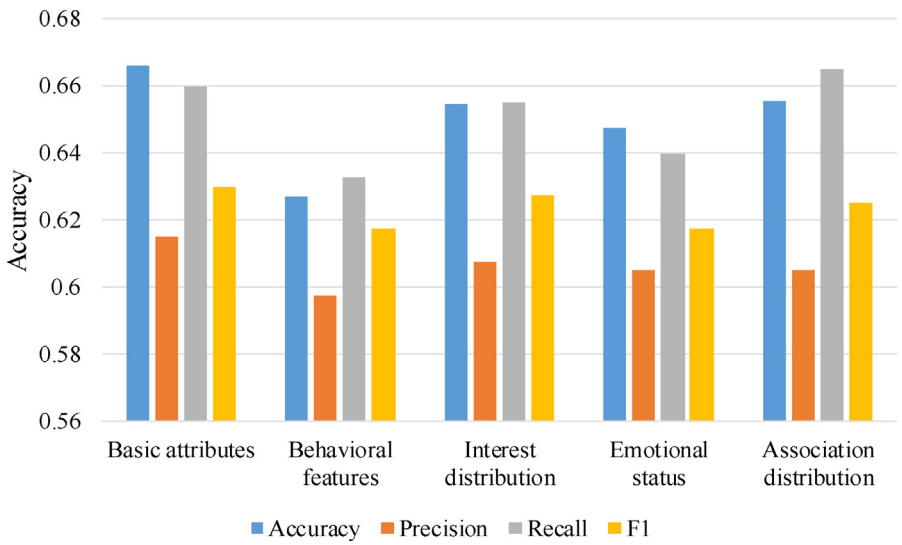

Fig. 5. Average performance of different base learners for single-type features.

Table 8 shows the base learners' prediction performance for multidimensional features when all the single-type features are incorporated. Compared to the performance with single-type features, multidimensional features bring up to a 12.07 % improvement in accuracy on average. Prediction performance based on multidimensional features is commonly better, demonstrating the powerful predictive ability of multidimensional features. SVM-RBF achieves the best performance with an accuracy of approximately 73.93 %.

The experimental results of the stacked classifier are presented in Table 9 and Fig. 6. "Single-type" means that the base learners accept only one of five types of features as input at a time, and all five outputs are taken as input for the secondary learners, which are obtained from the same base classifier (RF, KNN, LSTM, or MLP) by inputting each of five types of features each time: user basic attributes, behavioral features, user interest distribution, emotional status, and association distribution. In Table 9, "multidimensional" refers to the secondary learner using the outputs of the 10 base learners listed in Table 8

Table 8
Experimental results of multidimensional features for the base learner.

Model	Accuracy	Precision	Recall	F1
RF	0.7345	0.6181	0.7335	0.6345
XGBoost	0.7190	0.6118	0.7228	0.6409
GBDT	0.7343	0.5881	0.7335	0.6301
LR	0.7219	0.6011	0.7384	0.6360
SVM-RBF	**0.7393**	0.6161	0.7264	0.6283
SVM-POLY	0.7386	0.5961	0.7380	0.6296
KNN	0.7256	**0.6475**	0.7256	0.6456
LSTM	0.7379	0.6128	**0.7405**	0.6422
MLP	0.7077	0.6119	0.7098	**0.6463**
CNN	0.7288	0.6193	0.7308	0.6341
Average	0.7287	0.6123	0.7299	0.6367

(Experimental results of multidimensional features for the base learner) as input. For the stacked classifiers, we input the results of different kinds of base learners.

Due to the rich information contained in the multidimensional joint data, the probability generally outperforms the label. However, this pattern is not evident in the task where a single-type feature classifier is used as the base learner. Using single-type feature classifiers as the base learners, LSTM achieves outstanding performance compared to other machine learning models as the base learner, with an average accuracy of 78.54 % on probability data, 77.51 % on label data, and 78.60 % on joint data. An accuracy of 82.12 % is obtained using the joint data of the multiple-feature learner. The results sufficiently prove the advantages of integrating multidimensional features. A stacked learning model can integrate multimodal features and can significantly improve performance.

5.2. Feature importance

To explore the relationship among various features and user policy opinions, we use RF to determine the importance of features and filter out the most important features in each feature type (Fig. 7). Among the user basic features, the three most important are the number of followings, the number of microblogs, and the number of followers, which can be used to measure a user's social activity and discussion participation. Gender and social status are the least important features. In user posting behavioral features, the average times each microblog has been liked, the proportion of retweets, and the average number of @ per microblog are the most important. These features also indicate frequent communication between a target user and other users, and the more that user's opinions are recognized, the more likely it is that his or her opinions will be detected. Among user interest topics, personal preferences (No. 195), feelings for family and lovers (No. 63), and perceptions of life (No. 118) are the most prominent, and these topics are closely related to opinions and emotions, as shown in Table 10.

Moreover, for user emotional status features, strong opinions (*strongly negative* and *strongly positive*) are more helpful in revealing users' opinions on policies. Since the association distribution also considers both the relevance to the policy distribution and user opinions, the accuracy of the association distribution is higher than that of the features that contain only user information. The association distribution values reflect the degree of concern about policy topics, with higher values indicating more frequent discussion or stronger support of a particular policy topic. Here, we use the topics to reflect the level of relevance. These topics are related to the credit system (No. 21: protection, information, dishonest conduct, internet, disclosure, etc.), tourism (No. 18: scenic spots, flights, travel, international, transportation, etc.), drugs and health care (No. 5: drugs, focus, catalog, anticancer drugs, negotiation, etc.), employment (No. 17: employment, reform, abolition, recruitment, economy, etc.), and taxation (No. 26: special, deduction, taxpayer, income, personal income tax, etc.) and received high importance scores in the association distribution. Credit systems, employment and taxation are all closely related to individuals' income and ability to make a living. This result suggests that a user's concern about his or her economic circumstances may influence his or her perception and evaluation of policies in other areas.

Our experimental results show that it is feasible to utilize a user's digital footprint to infer his or her opinions on policies. This allows decision-makers to repeatedly collect opinions on different policies for a fixed user community and devise effective persuasive communication strategies [15]. The effective features extracted in our method have universal applicability to various classification algorithms. Not only do the opinion clues exist between users and policies, but users' personal attributes also improve the forecast. This has implications for a better understanding of the relationship among user profiles, comment texts, and opinions on other entities.

6. Discussion and conclusions

With an increasing number of people taking to social media to debate public policy, quickly capturing the public response to public policies and further revising and promoting public policies are of great significance to policymakers. Many researchers have focused on opinion mining after policy implementation, and a few studies have empirically predicted policy opinions in the early stage of policy-making, such as the release stage. However, there has been a lack of investigation of

Table 9
Experimental results of stacked classifiers.

Input	Model	Accuracy	Precision	Recall	F1
Single-type probability (RF)	RF	**0.7066**	0.6172	**0.7065**	0.6418
	KNN	0.7005	**0.6230**	0.7007	**0.6475**
	LSTM	0.6511	0.6211	0.6492	0.6399
	MLP	0.6751	0.6140	0.6839	0.6447
	Average	0.6833	0.6188	0.6851	0.6435
Single-type probability (KNN)	RF	**0.6719**	0.6098	**0.6731**	0.6328
	KNN	0.6581	**0.6135**	0.6594	0.6330
	LSTM	0.6701	0.6126	0.6720	**0.6354**
	MLP	0.6626	0.6085	0.6616	0.6352
	Average	0.6657	0.6111	0.6665	0.6341
Single-type probability (LSTM)	RF	0.7677	0.7423	0.7614	0.7514
	KNN	0.7864	**0.7693**	0.7870	**0.7707**
	LSTM	0.7936	0.7643	**0.7909**	0.7687
	MLP	**0.7941**	0.7438	0.7898	0.7491
	Average	0.7854	0.7549	0.7823	0.7600
Single-type probability (MLP)	RF	0.6430	0.6054	0.6426	0.6219
	KNN	0.6445	0.6060	0.6450	0.6230
	LSTM	**0.6489**	**0.6101**	**0.6507**	**0.6277**
	MLP	0.6426	0.6046	0.6437	0.6171
	Average	0.6447	0.6065	0.6455	0.6224
Single-type label (RF)	RF	**0.7058**	0.6178	0.7067	0.6434
	KNN	0.7001	**0.6181**	0.7001	0.6412
	LSTM	0.6988	0.6143	0.7007	0.6389
	MLP	0.7034	0.6156	**0.7077**	**0.6447**
	Average	0.7020	0.6164	0.7038	0.6420
Single-type label (KNN)	RF	0.6751	**0.6121**	0.6755	0.6369
	KNN	0.6807	0.6069	0.6805	0.6348
	LSTM	0.6845	0.6073	0.6833	**0.6382**
	MLP	**0.7279**	0.6068	**0.7327**	0.6296
	Average	0.6920	0.6083	0.6930	0.6349
Single-type label (LSTM)	RF	**0.7847**	**0.7694**	**0.7855**	0.7609
	KNN	0.7551	0.7375	0.7537	0.7447
	LSTM	0.7811	0.7567	0.7824	**0.7632**
	MLP	0.7793	0.7293	0.7794	0.7369
	Average	0.7751	0.7482	0.7752	0.7514
Single-type label (MLP)	RF	0.6496	0.6093	0.6496	0.6267
	KNN	0.6494	**0.6108**	0.6475	0.6268
	LSTM	0.6513	0.6101	0.6507	**0.6276**
	MLP	**0.6532**	0.6004	**0.6535**	0.6251
	Average	0.6509	0.6076	0.6503	0.6265
Single-type joint (RF)	RF	**0.7068**	0.6172	**0.7067**	0.6429
	KNN	0.7047	0.6184	0.7048	0.6432
	LSTM	0.7031	**0.6240**	0.6988	0.6465
	MLP	0.6992	0.6193	0.7050	**0.6486**
	Average	0.7034	0.6197	0.7038	0.6453
Single-type joint (KNN)	RF	0.6715	0.6082	0.6718	0.6314
	KNN	0.6571	**0.6133**	0.6530	0.6308
	LSTM	0.6739	0.6109	0.6722	0.6355
	MLP	**0.6781**	0.6055	**0.6769**	**0.6429**
	Average	0.6701	0.6095	0.6685	0.6351
Single-type joint (LSTM)	RF	0.7667	0.7441	0.7583	0.7436
	KNN	0.7902	0.7729	0.7903	0.7736
	LSTM	**0.7951**	**0.7771**	**0.7943**	**0.7763**
	MLP	0.7919	0.7729	0.7885	0.7667
	Average	0.7860	0.7667	0.7828	0.7650
Single-type joint (MLP)	RF	0.6419	0.6043	0.6426	0.6217
	KNN	**0.6496**	**0.6122**	**0.6475**	**0.6278**
	LSTM	0.6479	0.6090	0.6460	0.6253
	MLP	0.6462	0.6106	0.6460	0.6261
	Average	0.6464	0.6090	0.6455	0.6252
Multidimensional probability	RF	0.7415	0.5783	0.7411	0.6362
	KNN	0.7243	0.5921	0.7244	0.6373
	LSTM	**0.7462**	**0.6165**	**0.7460**	0.6375
	MLP	0.7449	0.6072	0.7456	**0.6377**
	Average	0.7392	0.5985	0.7393	0.6372

Table 9 (*continued*)

Input	Model	Accuracy	Precision	Recall	F1
Multidimensional label	RF	0.7320	0.6148	**0.7327**	0.6339
	KNN	0.6925	0.6052	0.6926	0.6247
	LSTM	0.7313	**0.6246**	0.7286	0.6310
	MLP	**0.7324**	0.6134	0.7301	**0.6357**
	Average	0.7221	0.6145	0.7210	0.6313
Multidimensional joint	RF	0.8203	0.8322	**0.8208**	**0.7888**
	KNN	0.8084	0.8056	0.8084	0.7761
	LSTM	**0.8212**	**0.8410**	0.8195	0.7859
	MLP	0.8062	0.7551	0.8052	0.7575
	Average	0.8140	0.8085	0.8135	0.7771

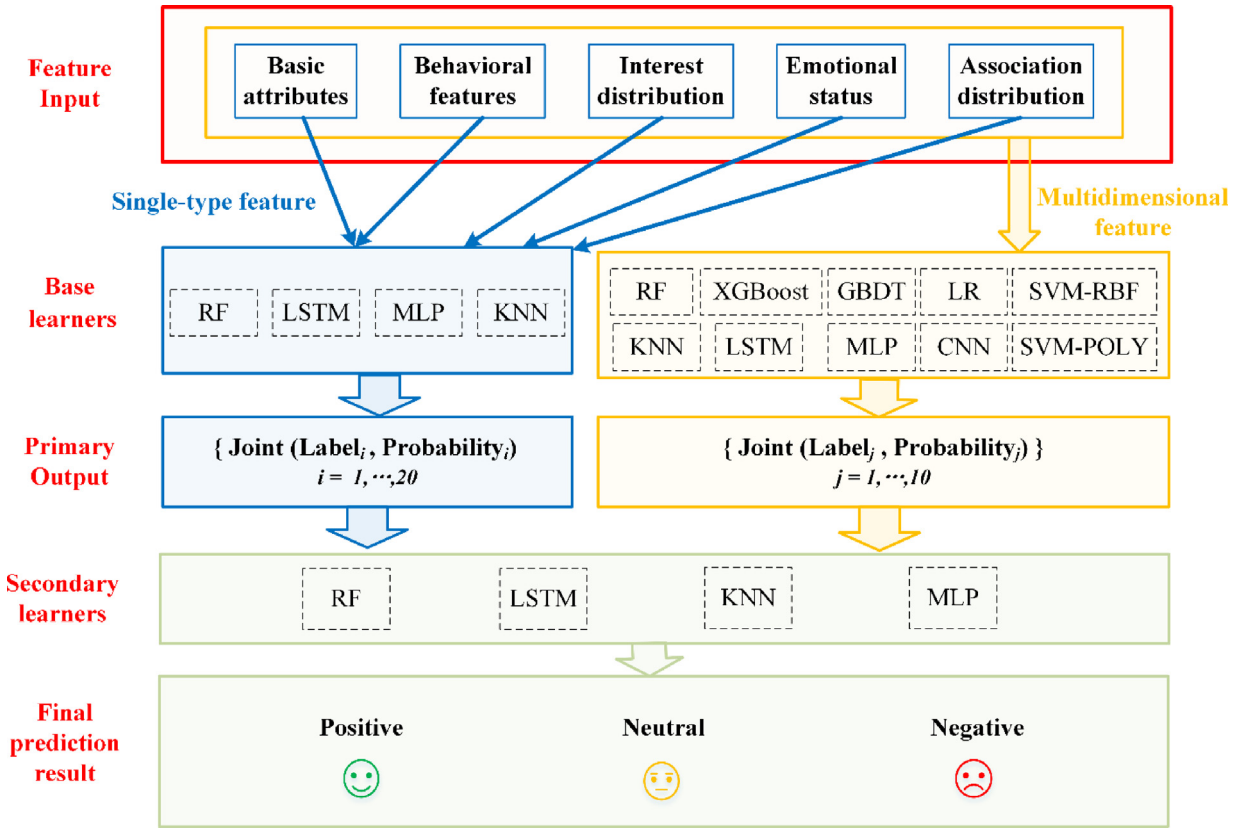

Fig. 6. Proposed stacking ensemble prediction model. Each base learner is trained on a subset of randomly selected 70% samples to increase heterogeneity among them. "Probability" and "label" represent the base learner's probability and label output format, respectively. "Joint" means that both probability and label data are used in the stacked classifiers.

what features impact users' attitudes toward a policy. To bridge this gap, this study develops a novel deep learning framework for policy opinion sentiment prediction, helping policymakers achieve a timely and precise perception of public opinion in the policy release phase.

This research has important theoretical and practical implications for policy opinion analysis and policy-making processes. First, it advances our understanding of policy opinion prediction. This is a complex task that has rarely been studied in the literature. In previous studies, topic detection and sentiment mining have been conducted to determine the main concerns from public policy opinion [3]. However, although policymakers can easily perceive public opinion about a policy [2], it is challenging to determine a previously unknown user's sentiment toward a new release policy if he or she has not publicly commented on the policy. An attention flow network was developed in a prior study to capture people's attention flows and to depict alterations in intention toward entities [7]. The key focus of this approach is how to comprehend interactions between users and entities as well as correlations among entities without focusing on various user-generated information.

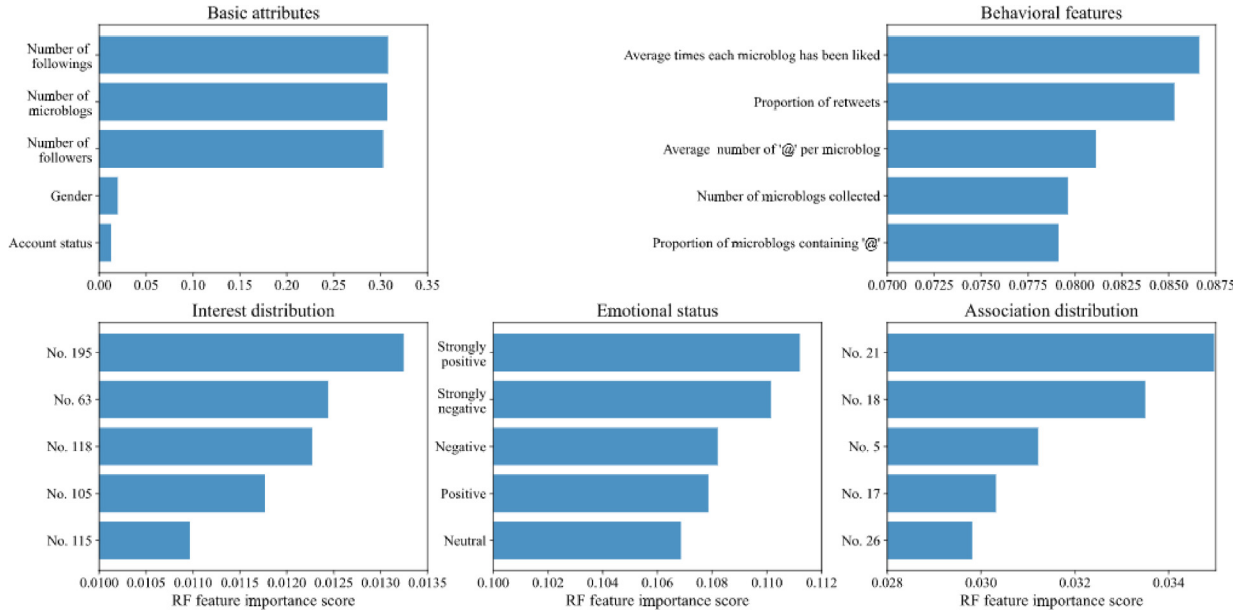

Fig. 7. RF feature importance score. The RF classifier generated importance scores, and after sorting, the five most important features of each category were selected.

Table 10
Keywords from user Weibo topics (the 5 most important).

No.	Label	Keywords
195	Personal preferences	feeling, happy, hope, lovely, thanks, pleasure, share, nice, friend
63	Affection (family)	like, children, parents, marriage, friends, love, girl, woman, life, in love, mother
118	Perceptions of life	daily life, life, love, time, emotions, effort, lives, happiness, forever, joy, good, hope, choice, friends
105	Affection (friend)	friend, reply, feel, special, bad, only, nice, home, sure, less, situation
115	Society	society, China, country, United States, government, economy, development, politics, locality, education

More specifically, our study contributes to the literature on sentiment prediction by focusing on digital footprint information on social media. We utilize user-generated information to predict a user's sentiment toward a policy before the user publishes explicit comments. Second, government administrators are often eager to estimate how the public reacts to the released policy. This automatic opinion prediction approach provides policy opinion detection capability to officials. When policymakers capture policy opinion early, they can adjust the policy or devise more appropriate policy communication strategies, such as recommending policy interpretation to policy opponents.

In this paper, we realize policy opinion sentiment prediction by a hierarchical stacked ensemble model combining multiple features. We devise different features to capture the policy opinion nexus from policy-related and user-related information on social media platforms. The multidimensional features are categorized into model-induced features and user basic and behavioral features. The model-induced features are extracted from four machine learning models: the user interest model, policy microblog topic model, opinion sentiment classification model, and policy judgment model. The model-induced features, which are more advanced, semantic and interpretable, are intended to grasp the embedded relevance between the released policy and public opinion. Then, we use a hierarchical stacked ensemble learning model to integrate multidimensional features as the final prediction model of public opinion on policy in the policy release phase. We conduct comprehensive experiments to examine the performance of multidimensional features for policy opinion sentiment prediction. The experimental results on policy comment datasets demonstrate that when our model considers multiple features, it can significantly improve sentiment prediction performance.

A limitation of our research is that we did not consider a user's social structural information. If one user follows another, the policy information and opinion of the first user can influence those of the second via communication between the users, and two users who are friends are likely to hold the same or similar attitudes toward a certain policy. A user's social structure is generally also a part of his or her own attributes and can help improve the prediction accuracy of public opinion about public policy. The distributions of policy formulation and policy dissemination in social media are unbalanced in reality, and user groups and policy information are not regularly matched. The use of data to truly reflect the distribution and dissemination of policy and effectively bridge the gap between actual and media communication is important for the practical significance and application of the proposed method and is thus a future research direction. The proposed method can also

be applied to attitude prediction in other fields, such as inferring user attraction to advertising or promotional wording and examining the spread and influence of rumors and fake news in different user groups.

CRediT authorship contribution statement

Wenju Hou: Conceptualization, Methodology, Software, Writing – original draft. **Ying Li:** Conceptualization, Methodology, Writing – original draft. **Yijun Liu:** Resources, Supervision, Funding acquisition. **Qianqian Li:** Conceptualization, Methodology, Software, Investigation, Formal analysis, Writing – original draft, Funding acquisition.

Data availability

Data will be made available on request.

Declaration of Competing Interest

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

Acknowledgments

The authors acknowledge financial support from the Natural Science Foundation of Beijing (No. 9222030), Key Project of the Education Department of Jilin Province (No. JJKH20221012KJ), National Natural Science Foundation of China (No. 71774154, 72074205, 71573247).

References

[1] R. Anne, R. Stefanie, T. Dimiter, The opinion-policy nexus in Europe and the role of political institutions, European Journal of Political Research. 58 (2019) 412–434.
[2] J. Xiong, X. Feng, Z. Tang, Understanding user-to-User interaction on government microblogs: An exponential random graph model with the homophily and emotional effect, Inf. Process. Manage. 57 (2020) 102229.
[3] O. Belkahla Driss, S. Mellouli, Z. Trabelsi, From citizens to government policy-makers: Social media data analysis, Government Information Quarterly. 36 (2019) 560–570.
[4] S.C. McGregor, Social media as public opinion: How journalists use social media to represent public opinion, Journalism. 20 (2019) 1070–1086.
[5] R. Medaglia, D. Zhu, Public deliberation on government-managed social media: A study on Weibo users in China, Government Information Quarterly. 34 (2017) 533–544.
[6] M. Kosinski, D. Stillwell, T. Graepel, Private traits and attributes are predictable from digital records of human behavior, PNAS 110 (2013) 5802–5805.
[7] Y. Chen, Y. Dai, X. Han, Y. Ge, H. Yin, P. Li, Dig users' intentions via attention flow network for personalized recommendation, Inf. Sci. 547 (2021) 1122–1135.
[8] A. Khan, H. Zhang, N. Boudjellal, A. Ahmad, J. Shang, L. Dai, B. Hayat, Election Prediction on Twitter: A Systematic Mapping Study, Complexity. 2021 (2021) 5565434.
[9] N. Annalyn, M.W. Bos, L. Sigal, B. Li, Predicting personality from book preferences with user-generated content labels, IEEE Trans. Affective Comput. 11 (2020) 482–492.
[10] Y. Ophir, R. Tikochinski, C.S.C. Asterhan, I. Sisso, R. Reichart, Deep neural networks detect suicide risk from textual facebook posts, Sci. Rep. 10 (2020) 16685.
[11] B. Mei, Y. Xiao, R. Li, H. Li, X. Cheng, Y. Sun, Image and attribute based convolutional neural network inference attacks in social networks, IEEE Trans. Network Sci. Eng. 7 (2020) 869–879.
[12] Y. Li, L. Yang, B. Xu, J. Wang, H. Lin, Improving user attribute classification with text and social network attention, Cognitive Comp. 11 (2019) 459–468.
[13] C. Gao, X. He, D. Gan, X. Chen, F. Feng, Y. Li, T. Chua, D. Jin, Neural multi-task recommendation from multi-behavior data, in: 2019 IEEE 35th International Conference on Data Engineering (ICDE), 2019, pp. 1554–1557.
[14] L. Shi, W. Wu, W. Guo, W. Hu, J. Chen, W. Zheng, L. He, SENGR: sentiment-enhanced neural graph recommender, Inf. Sci. 589 (2022) 655–669.
[15] S.C. Matz, M. Kosinski, G. Nave, D.J. Stillwell, Psychological targeting as an effective approach to digital mass persuasion, PNAS 114 (2017) 12714–12719.
[16] P. Chauhan, N. Sharma, G. Sikka, The emergence of social media data and sentiment analysis in election prediction, J. Ambient Intellig. Humanized Comp. 12 (2021) 2601–2627.
[17] K.d.S. Brito, P.J.L. Adeodato, Predicting Brazilian and U.S. Elections with Machine Learning and Social Media Data, in: 2020 International Joint Conference on Neural Networks (IJCNN), 2020, pp. 1–8.
[18] X. Yang, Y. Li, Q. Li, D. Liu, T. Li, Temporal-spatial three-way granular computing for dynamic text sentiment classification, Inf. Sci. 596 (2022) 551–566.
[19] M.Z. Asghar, F.M. Kundi, S. Ahmad, A. Khan, F. Khan, T-SAF: Twitter sentiment analysis framework using a hybrid classification scheme, Expert Syst. 35 (2018) e12233.
[20] Y. Liu, J.-W. Bi, Z.-P. Fan, A method for multi-class sentiment classification based on an improved one-vs-one (OVO) strategy and the support vector machine (SVM) algorithm, Inf. Sci. 394–395 (2017) 38–52.
[21] H.T. Phan, N.T. Nguyen, V.C. Tran, D. Hwang, An approach for a decision-making support system based on measuring the user satisfaction level on Twitter, Inf. Sci. 561 (2021) 243–273.
[22] F. Ali, D. Kwak, P. Khan, S. El-Sappagh, A. Ali, S. Ullah, K.H. Kim, K.-S. Kwak, Transportation sentiment analysis using word embedding and ontology-based topic modeling, Knowl.-Based Syst. 174 (2019) 27–42.
[23] A. Yadav, D. Vishwakarma, Sentiment analysis using deep learning architectures: a review, Artif. Intell. Rev. 53 (2019) 4335–4385.
[24] L. Kong, C. Li, J. Ge, F.F. Zhang, Y. Feng, Z. Li, B. Luo, Leveraging multiple features for document sentiment classification, Inf. Sci. 518 (2020) 39–55.
[25] A. Abdi, S.M. Shamsuddin, S. Hasan, J. Piran, Deep learning-based sentiment classification of evaluative text based on Multi-feature fusion, Inf. Process. Manage. 56 (2019) 1245–1259.
[26] C. Gan, Q. Feng, Z. Zhang, Scalable multi-channel dilated CNN–BiLSTM model with attention mechanism for Chinese textual sentiment analysis, Future Gener. Comp. Syst.. 118 (2021) 297–309.

[27] E. Ferro, E.N. Loukis, Y. Charalabidis, M. Osella, Policy making 2.0: From theory to practice, Govern. Inf. Q. 30 (2013) 359–368.
[28] C.G. Reddick, A.T. Chatfield, A. Ojo, A social media text analytics framework for double-loop learning for citizen-centric public services: A case study of a local government Facebook use, Govern. Inf. Q. 34 (2017) 110–125.
[29] P. Sobkowicz, M. Kaschesky, G. Bouchard, Opinion mining in social media: Modeling, simulating, and forecasting political opinions in the web, Govern. Inf. Q. 29 (2012) 470–479.
[30] L. Hagen, T.E. Keller, X. Yerden, L.F. Luna-Reyes, Open data visualizations and analytics as tools for policy-making, Govern. Inf. Q. 36 (2019) 101387.
[31] B. Dahal, S.A.P. Kumar, Z. Li, Topic modeling and sentiment analysis of global climate change tweets, Soc. Network Anal. Min. 9 (2019) 24.
[32] A. Boudjelida, S. Mellouli, J. Lee, Electronic citizens participation: Systematic review, in: Proceedings of the 9th International Conference on Theory and Practice of Electronic Governance, 2016, pp. 31–39.
[33] E. Bonsón, D. Perea, M. Bednárová, Twitter as a tool for citizen engagement: An empirical study of the Andalusian municipalities, Govern. Inf. Q. 36 (2019) 480–489.
[34] Z. Zengin Alp, Ş. Gündüz Öğüdücü, Identifying topical influencers on twitter based on user behavior and network topology, Knowl.-Based Syst. 141 (2018) 211–221.
[35] J. Shi, C.T. Salmon, Identifying opinion leaders to promote organ donation on social media: Network study, J. Med. Internet Res. 20 (2018) e7643.
[36] X. Chen, L. Xu, Z. Liu, M. Sun, H. Luan, Joint learning of character and word embeddings, in: Proceedings of the 24th International Conference on Artificial Intelligence, 2015, pp. 1236–1242.
[37] X.D. Feng, K.X. Hui, X. Deng, G.Y. Jiang, Understanding how the semantic features of contents influence the diffusion of government microblogs: Moderating role of content topics, Inf. Manage. 58 (2021) 103547.
[38] U. Yaqub, S.A. Chun, V. Atluri, J. Vaidya, Analysis of political discourse on twitter in the context of the 2016 US presidential elections, Govern. Inf. Q. 34 (2017) (2016) 613–626.
[39] S.N. Firdaus, C. Ding, A. Sadeghian, Topic specific emotion detection for retweet prediction, Int. J. Mach. Learn. Cybern. 10 (2019) 2071–2083.
[40] K. Cortis, B. Davis, Over a decade of social opinion mining: a systematic review, Artif. Intell. Rev. 4873–4965 (2021).
[41] P.S. Mashhadi, S. Nowaczyk, S. Pashami, Stacked ensemble of recurrent neural networks for predicting turbocharger remaining useful life, Appl. Sci. 10 (2020) 69.
[42] S. Mishra, J. Diesner, Detecting the correlation between sentiment and user-level as well as text-level meta-data from benchmark corpora, in: Proceedings of the 29th on Hypertext and Social Media, 2018, pp. 2–10.