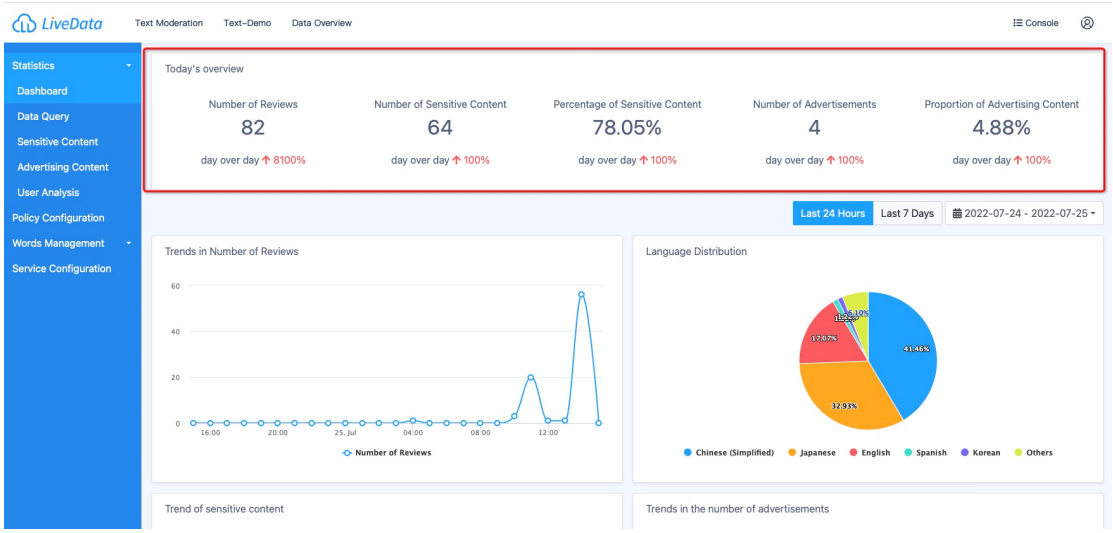
LiveData Text moderation Console User Manual

1、Statistics

1.1 Dashboard

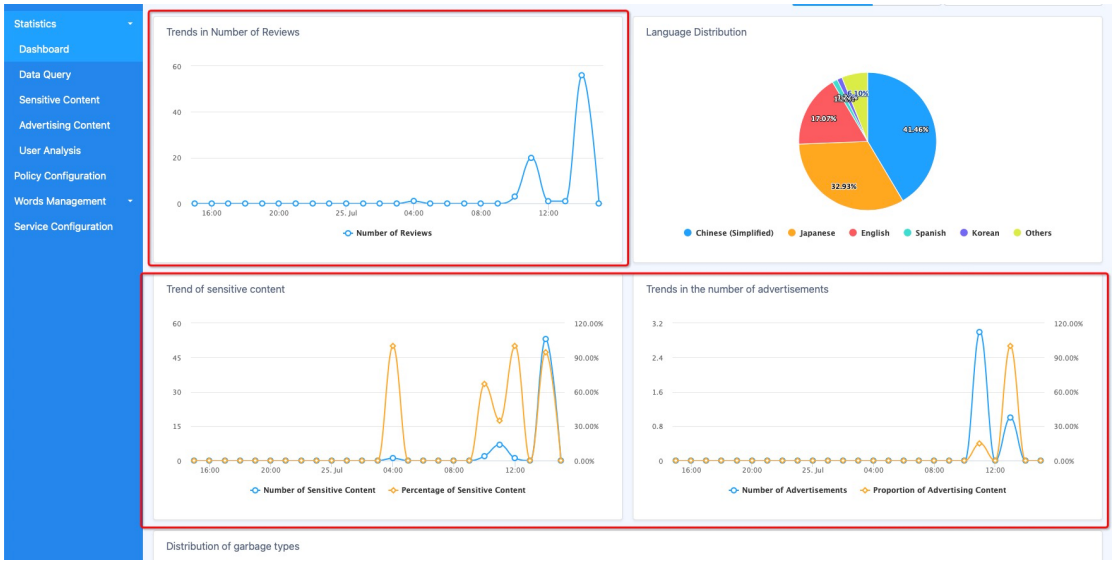
Here you can get an overview of the detected trend changes, detected language distribution, and detected type distribution of the project over a period of time.

1.1.1 Today's Overview



The number of detections, the number of sensitive content, the number of advertising content, the proportion of sensitive content, the proportion of advertising content, and the month-on-month change of each data item compared to the previous natural day is displayed here.

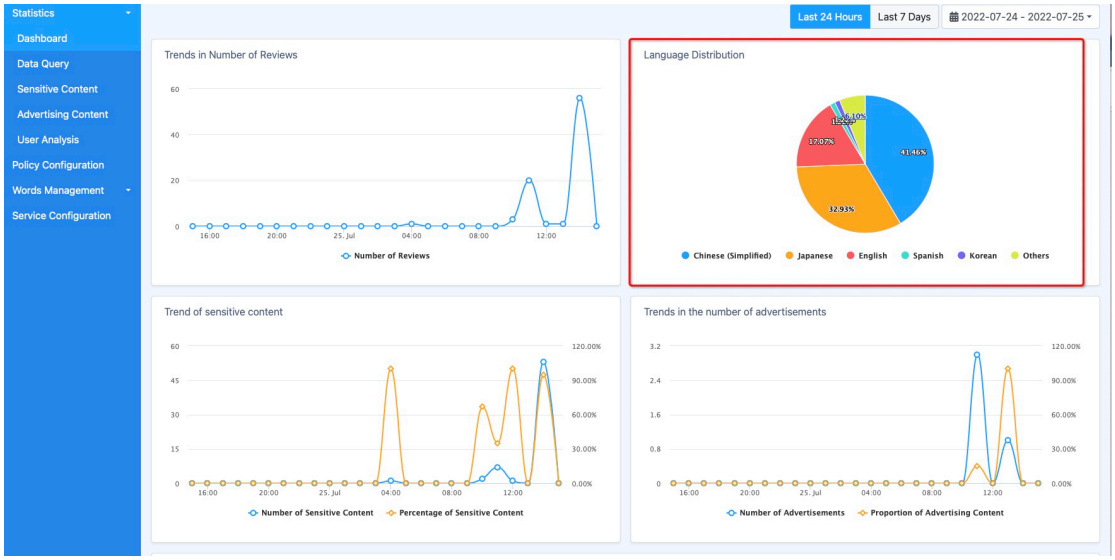
1.1.2 Trends in Number of Reviews&sensitive content&advertisements



The above three modules display the data trend of the number of detections, the number of sensitive content, and the number of advertisement content over time in the form of a line chart.

You can view the specific data at the corresponding time point by moving the mouse to a point on the polyline.

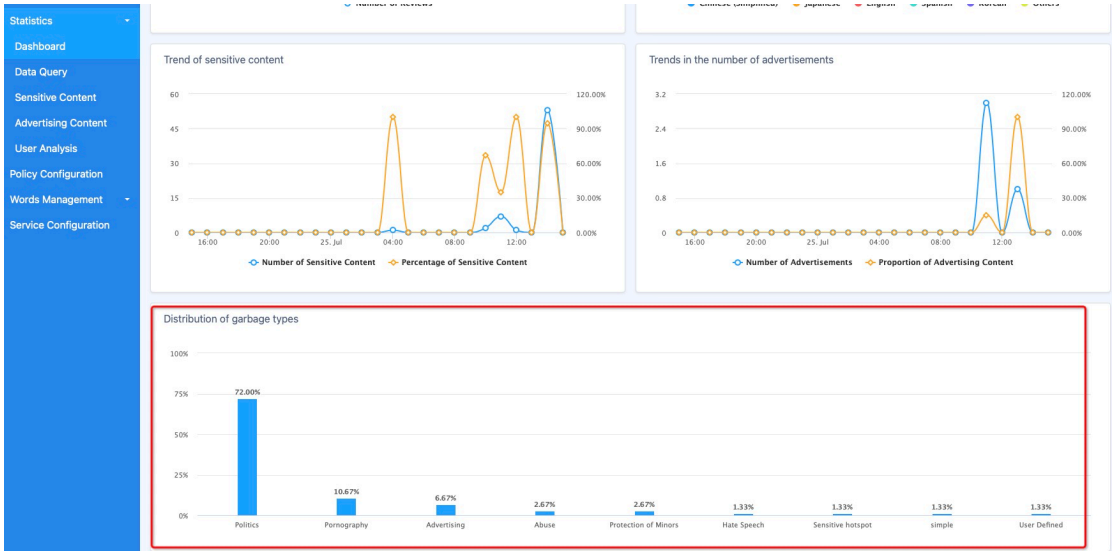
1.1.3 Language distribution



Here, the language distribution of the detected text is displayed in the form of a pie chart within your filtering time range.

You can view the specific number of detection bars for the corresponding language by moving the mouse to a fan-shaped area on the pie chart.

1.1.4 Distribution of garbage types

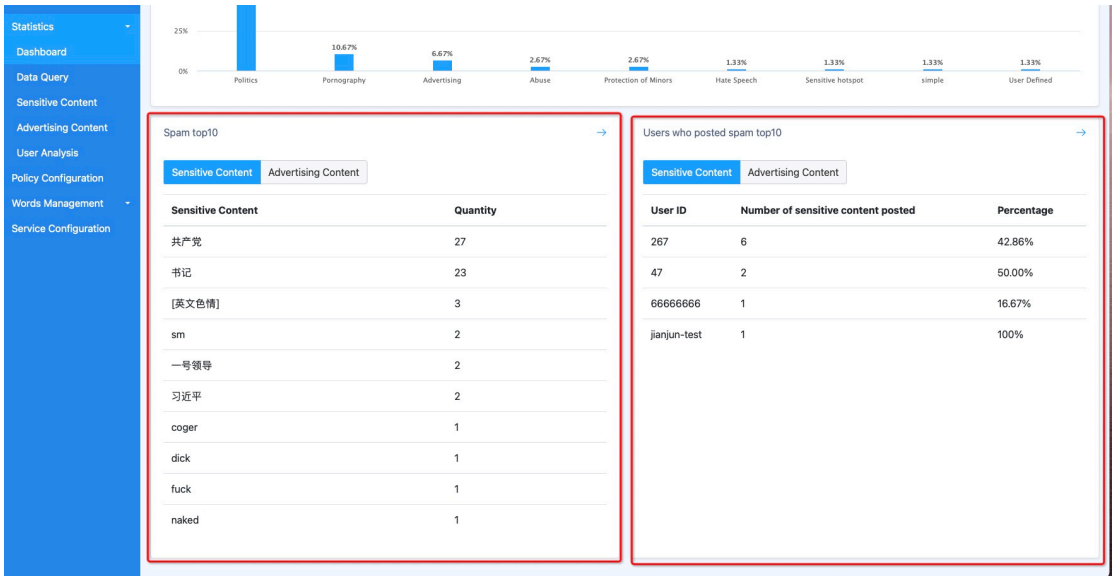


Here, the distribution of the types of sensitive content detected within your filtering time range is

displayed in the form of a bar chart.

You can view the specific detected number of the corresponding type by moving the mouse to a rectangle.

1.1.5 Spam&Users who posted spam top10



（1）Spam top10

- Sensitive Content：Here are the top 10 sensitive words with the most hits detected by sensitive content within your filtering time range and the corresponding number.

- Advertisements Content：Here are the top 10 most frequently appearing ad content and the corresponding number among the ad content detected within your filtering time range.

（2）Users who posted spam top10

If the user id is passed in when the project calls the text review interface, then the ids of the top 10 users who sent sensitive content and the largest number of advertising content within the screening time range, the number of users sent, and the content that accounts for the total sent content will be displayed here. proportion.

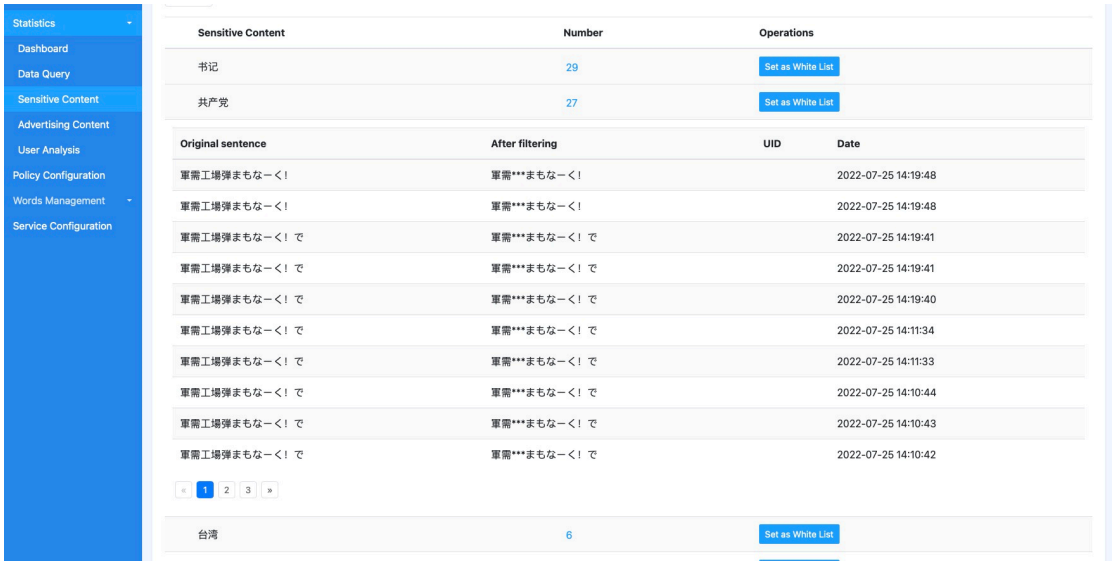
（3）“—>” button

- Click "—>" on the right side of "Spam top10" to enter the "Sensitive Content" page;
- Click "—>" on the right side of "Users who posted spam top10" to enter the "User Analysis" page.

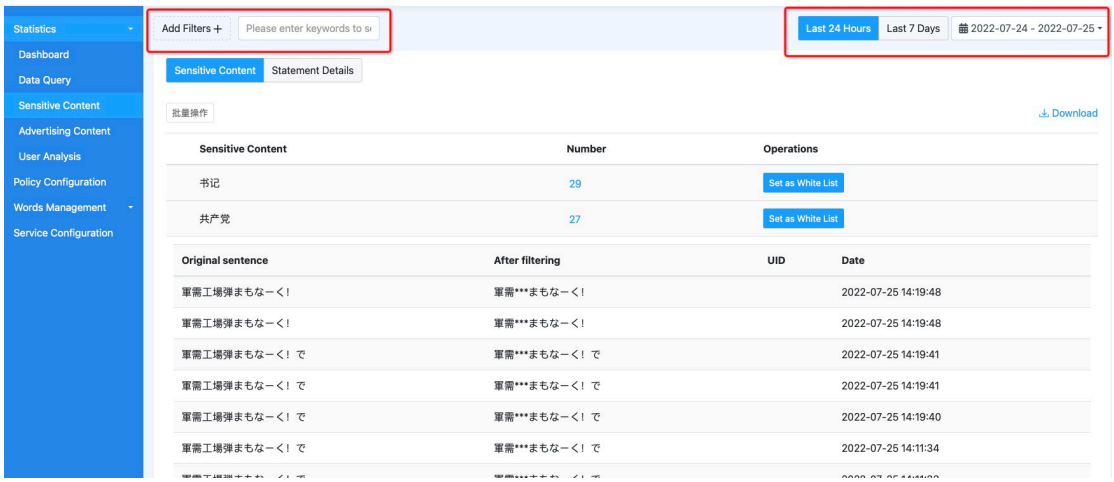
1.2 Sensitive Content

该页面按照检出敏感内容命中的敏感词分类统计数量。您可以在此快速浏览筛选时间段内命

中不同敏感词的数据。



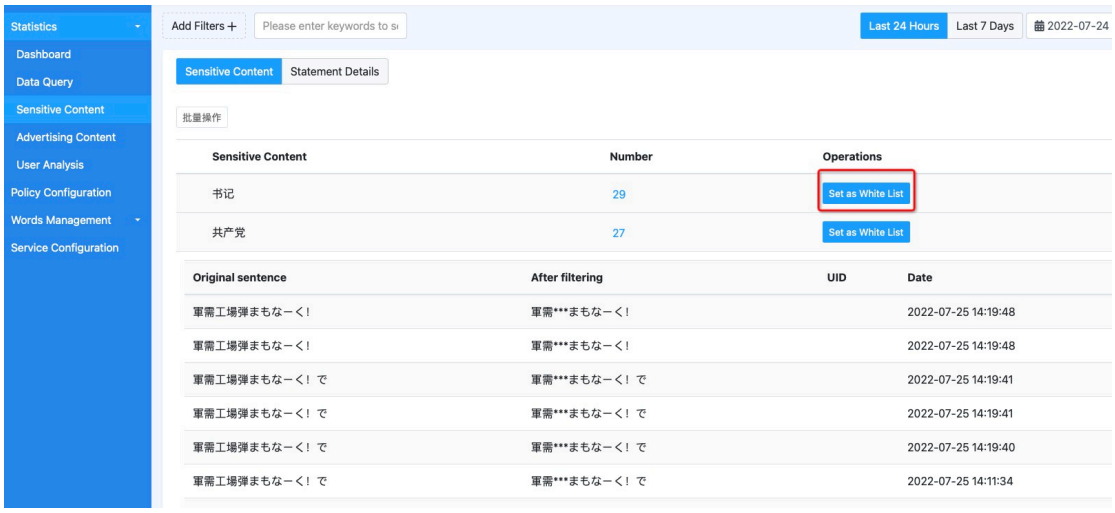
1.2.1 Filter Condition



You can filter the sensitive content you want to view by checking the language, hitting a sensitive category, entering a keyword, and selecting the four dimensions of time.

Edit the selected time period here, and the sensitive data corresponding to your selected time period will be displayed below.

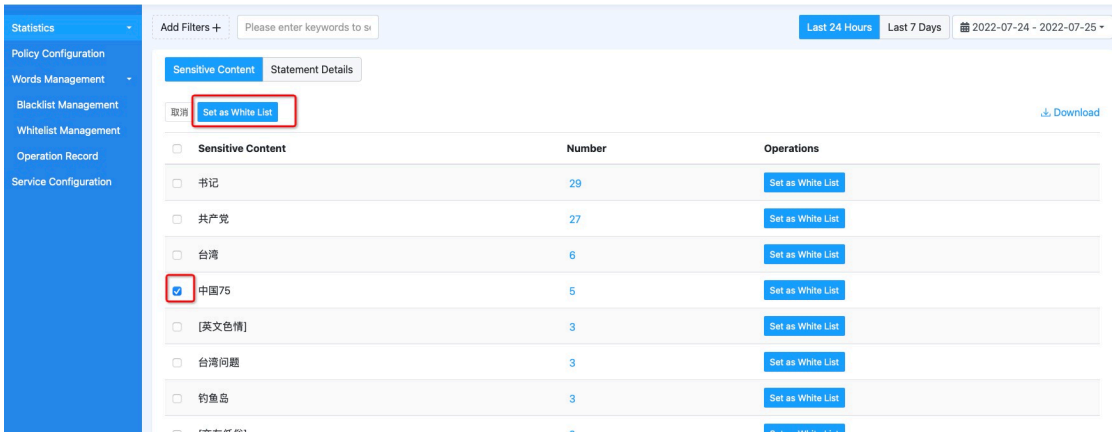
1.2.2 Set as White List



You can use this function to set sensitive words as "whitelist" words.

Click the "Set as White List" button after each word, and the system will prompt you to confirm the operation. After your confirmation, the corresponding sensitive words will enter the whitelist, which you can view on the "Whitelist Management" page. At the same time, the system will no longer detect texts that only hit this sensitive word.

1.2.3 Bulk operations

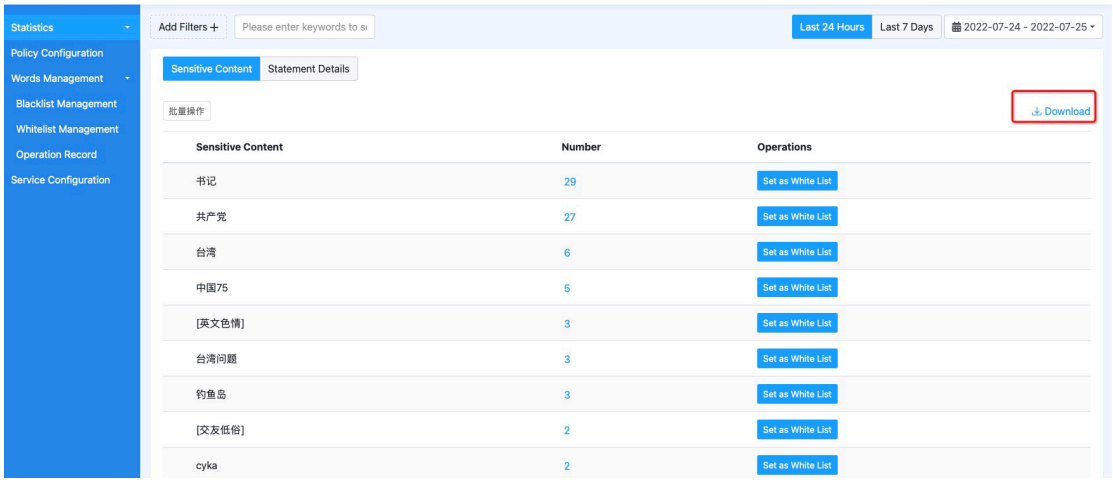


This function can batch set sensitive words as "whitelist" words.

Click the "Batch Operation" button, the page will enter the state as shown above, you can check multiple words you want to set as whitelist. After checking, click the "Set as whitelist" button and confirm that the processing logic after the operation is consistent with 1.2.3, and the page automatically exits the state shown in the figure above.

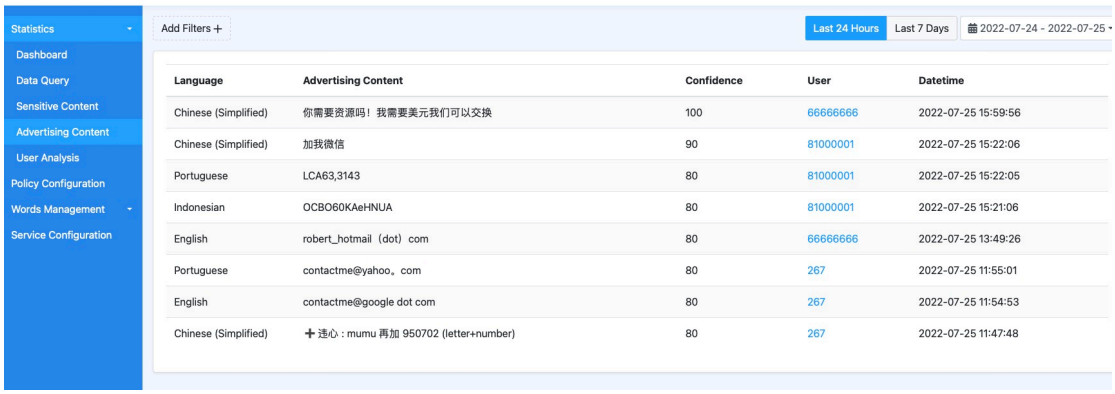
In the state shown in the figure above, click the "Cancel" button to exit this state directly.

1.2.4 Download



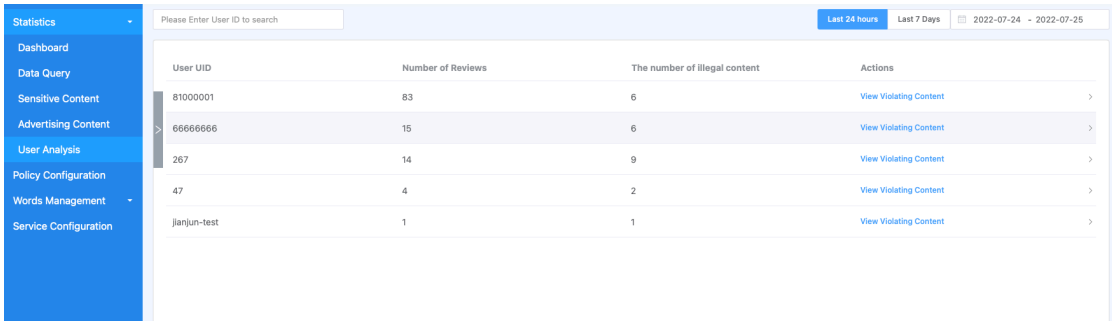
This function can download the file corresponding to the specific text content in .csv format to the local according to your filtering results.

1.3 Advertisements Content



This page displays ad content data detected during the filter period. It is basically the same as the "Sensitive Content" page. It supports three dimensions of selecting the language, hitting the sensitive category, and selecting the time to filter the advertising content to be viewed.

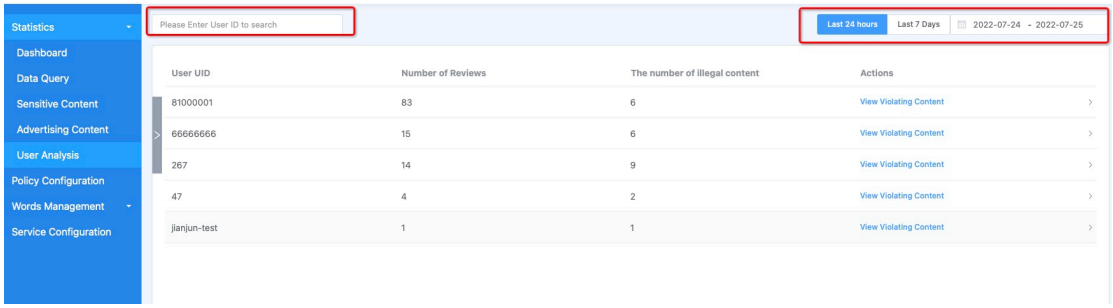
1.4 User Anyalysis



By default, the page uses the user id as the main dimension to display the number of detected content, the number of illegal content, and the specific text content of the corresponding user within the

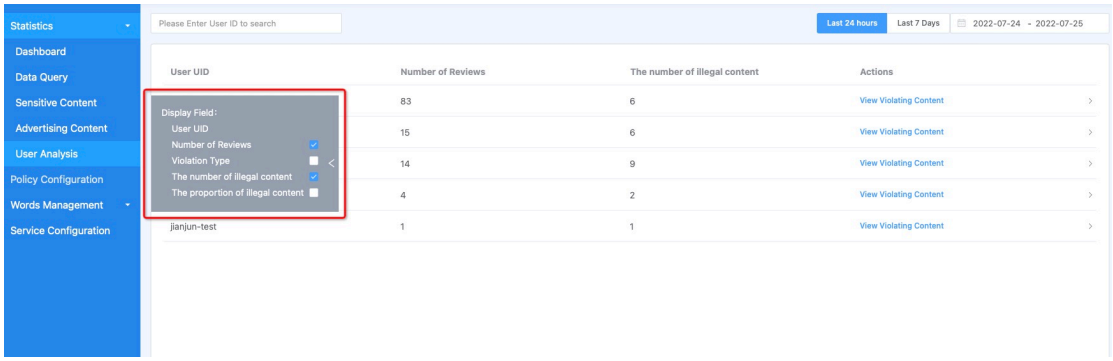
filtering time range.

1.4.1 Filter Condition



You can filter the content to be displayed by entering the user uid and selecting the time.

1.4.2 Display field selection

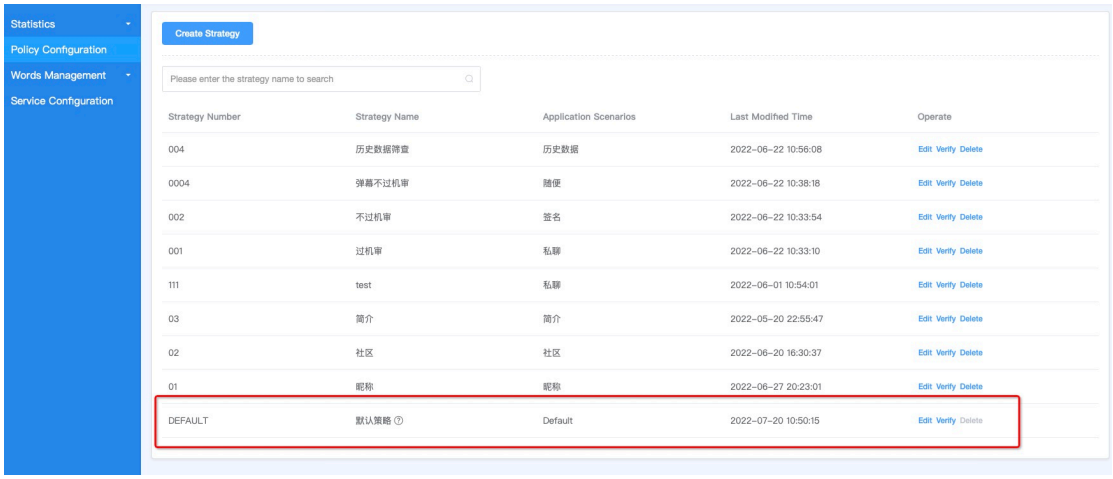


You can choose which fields to display via the Display Fields Configurator shown in the image above. When selecting to display "Ratio of Violating Content", the system will add a column of "Ratio of Violating Content" after the column of "Number of Violating Content"; The results of filtering data are displayed in separate rows.

2、Policy configuration

It is used to configure policies with different levels of audit tightness for different scenarios in the same project.

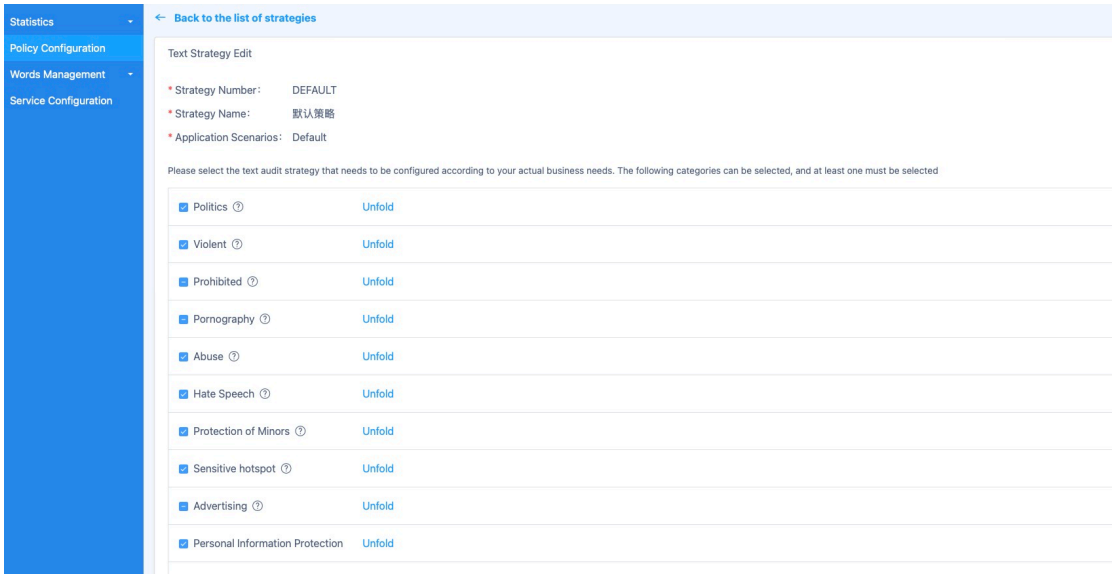
2.1 Default policy



A policy created automatically by the system when the project is created. When the input parameter strategyId of calling the text audit interface is null, the strategy will be executed by default.

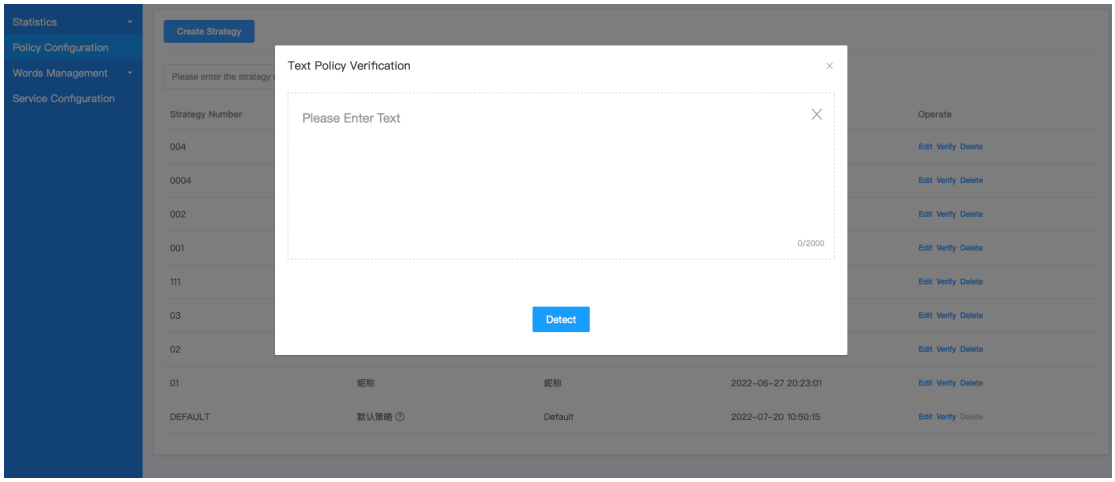
Note: The default policy does not support deletion.

2.1.1 strategy editor



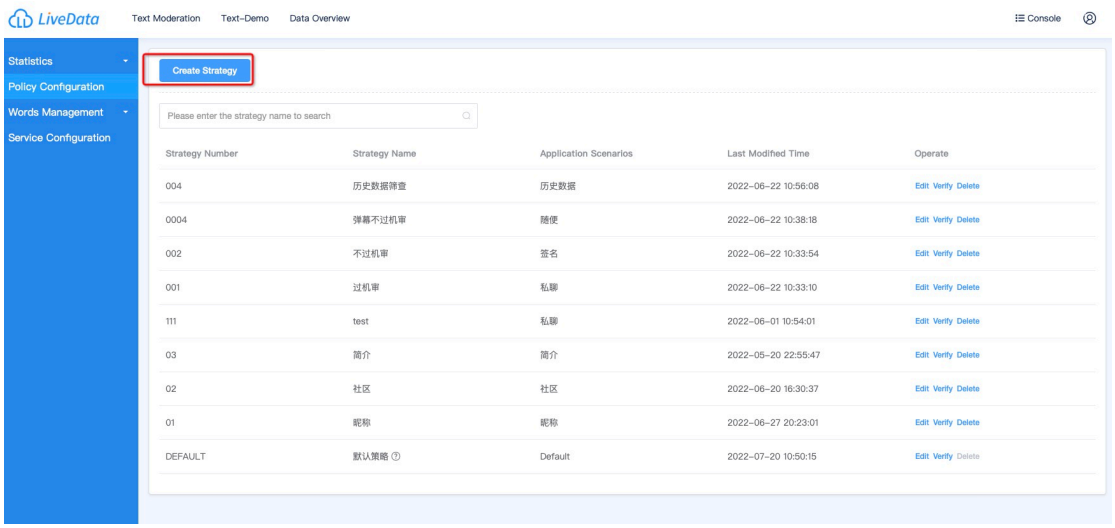
Click "Edit" to enter the "Text Policy Edit" page, where you can adjust the category of the corresponding policy to enable detection.

2.1.2 Policy verification



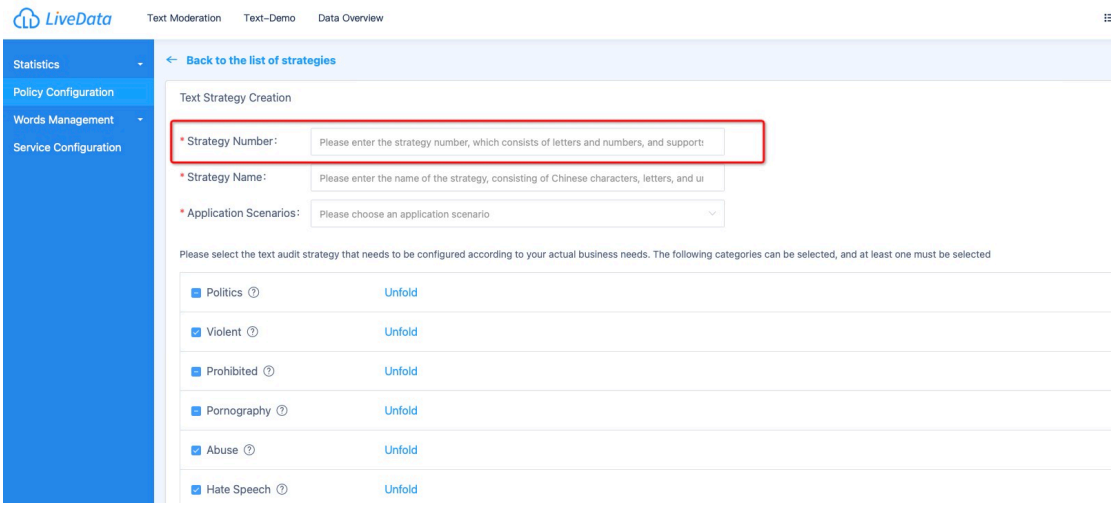
Click "Verify" to enter text to test the detection effect of your configured policy.

2.2 Create Strategy



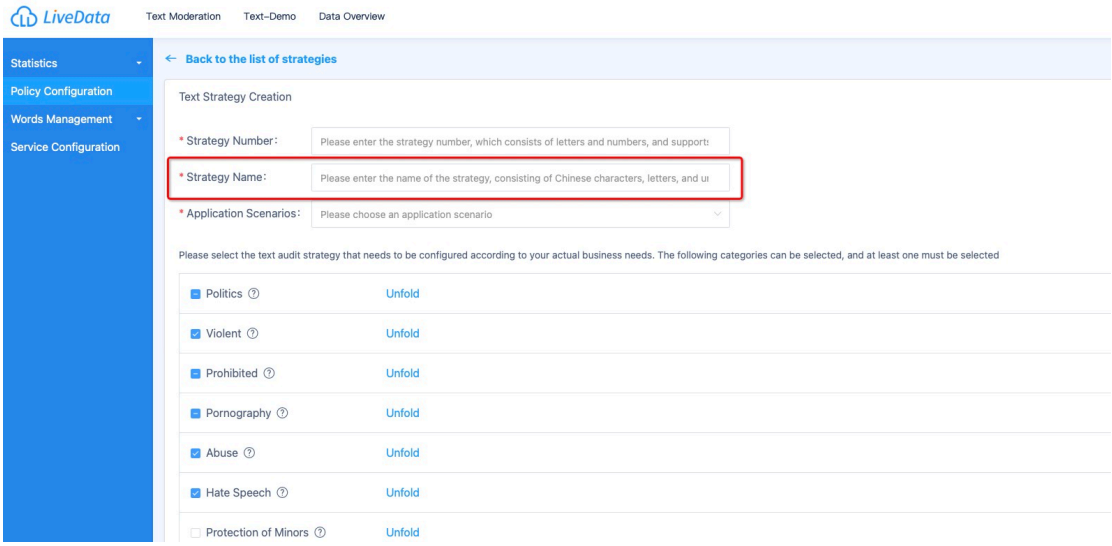
Click " Create Strategy " to enter the "Text Policy Creation" page.

2.2.1 Strategy Number



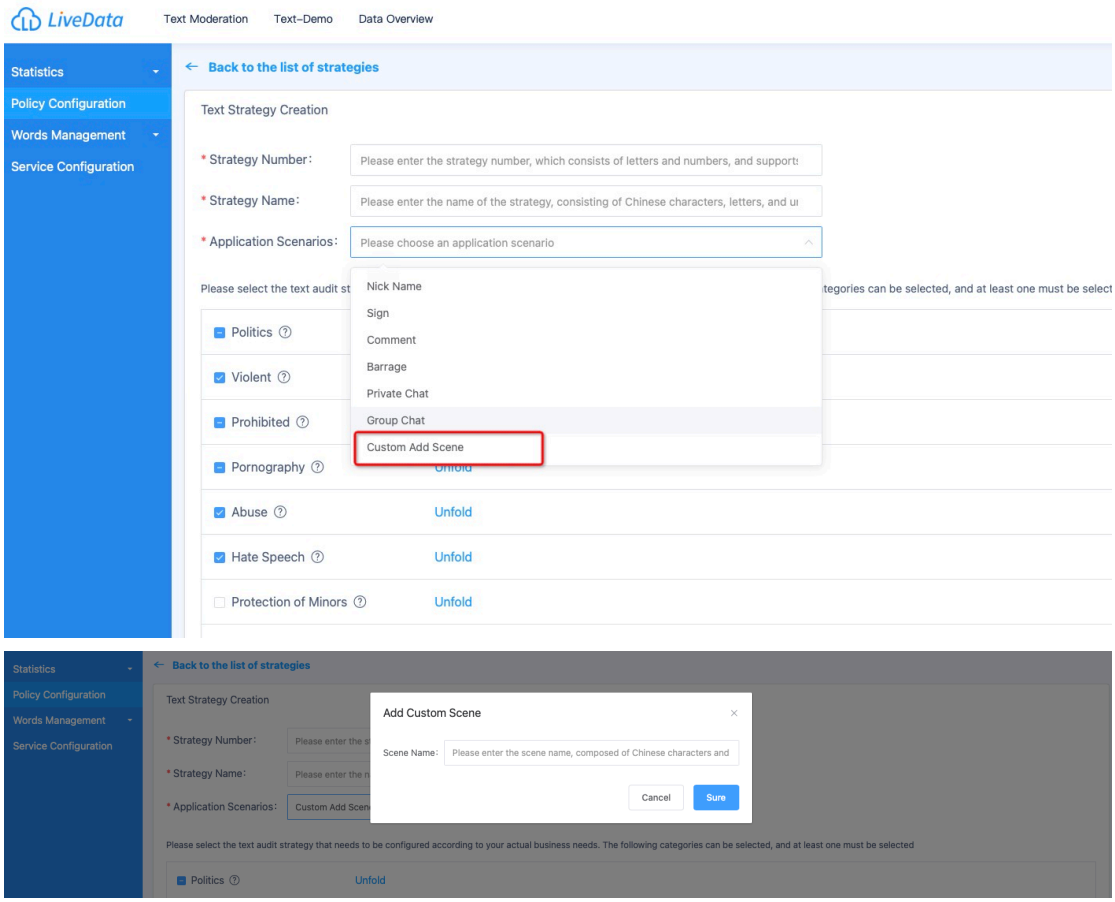
You need to define a unique number for this policy, which will be used as an identifier to distinguish different policies, so it cannot be duplicated with the number of an existing policy. When calling the text audit interface, the ID can be passed in as the value of the input parameter StrategyId, and the system will call the corresponding strategy to detect the text content.

2.2.2 Strategy Name



You need to enter the name of the strategy to be created here, so as to quickly query and distinguish different strategies after creation.

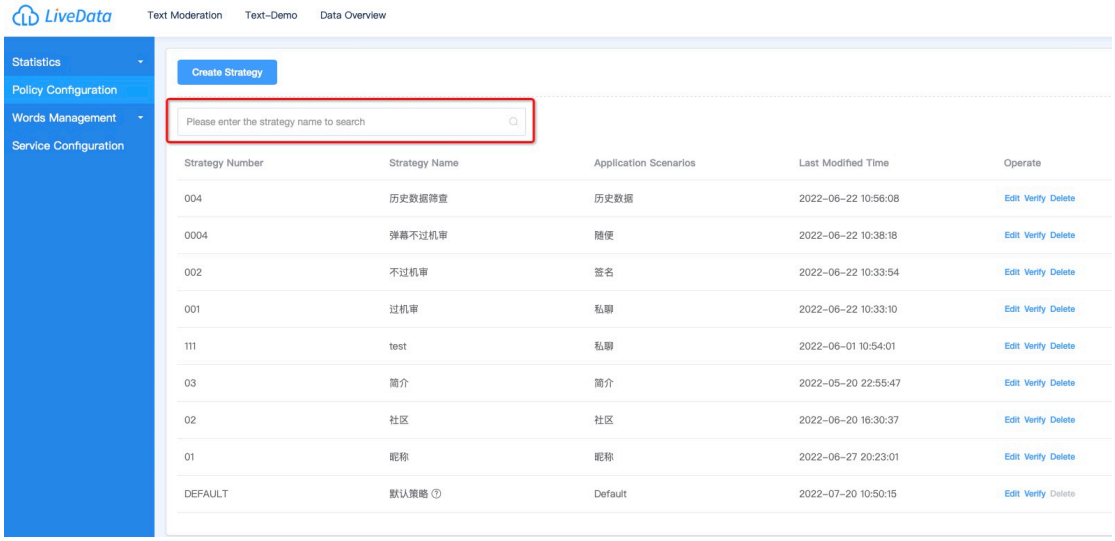
2.2.3 Application Scenarios



You can directly select from the system-defined scenes, or customize a scene name.

After filling in the above three items, you can continue to configure the categories to be detected by the policy as needed; after clicking "Save", the policy is created. Subsequent calls to the text audit interface can use the corresponding policy through the policy number.

2.3 Strategy Search

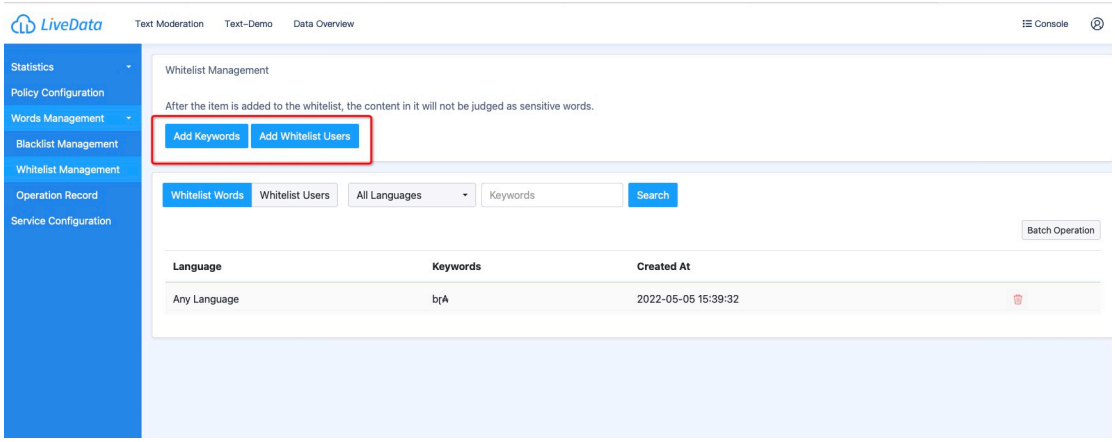


When there are many policies in the list, you can quickly find the policy you need to query by entering the policy name here.

3、Words Management

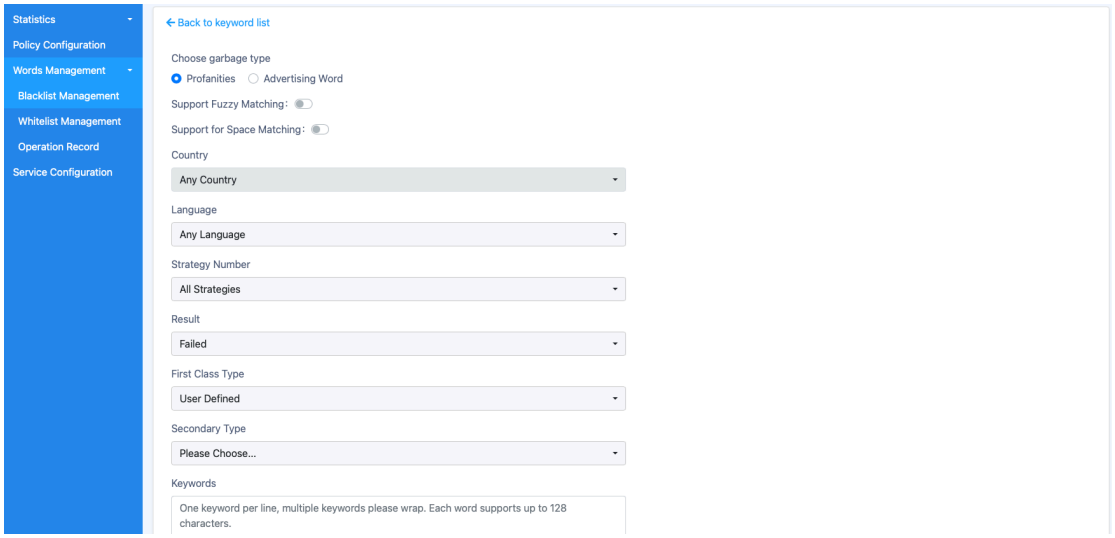
The blacklist is used to customize special sensitive words that need to be blocked;
The whitelist is used to customize the non-sensitive words that the system considers to be sensitive but do not need to be blocked for business purposes, and to customize the IDs of users who send content without text review.

3.1 Blacklist Management



You can add sensitive words and advertising words to the blacklist by directly adding or bulk uploading. The added sensitive words and advertising words will be classified and displayed in the list below.

3.1.1 Add keywords



Click "Add Keywords" to enter the vocabulary addition page.

(1) Select the type of garbage

It is divided into two types: common sensitive words and advertising words (Note: the customized advertising words in the blacklist can only be detected normally after the "user-defined advertising" category is enabled in the policy configuration.)

（2）Country

Here you can select the countries to which the sensitive words to be added are applicable. After the addition is successful, the corresponding sensitive word will only take effect when the value of the country field passed in the interface is the same as the configured country code.

（3）Language

Indicates the language to which the sensitive word (or advertising word) is applicable; that is, only when the sensitive word appears in the detected text and its language recognition result is in the applicable language selected by the sensitive word (or advertising word), the sensitive word (or advertising word)) will be detected.

（4）Strategy Number

Indicates the policy applicable to the sensitive word (or advertisement word); that is, the sensitive word will only be detected when the system uses the text to call the corresponding policy configured and the text contains the sensitive word.

（5）Result

Indicates the result returned by the system when the detected text hits the sensitive word, including two cases: failed and suspected.

（6）Type

Indicates the detection type returned by the system when the detected text hits the sensitive word. The first-level types include: user-defined, political, violent, prohibited, pornographic, abusive, hate speech, underage protection, sensitive hotspots, personal information Protection, Private Transactions, Violation Emojis; secondary types are subcategories of these primary categories. By default, the first-level type system automatically selects the "user-defined" class, and the second-level type does not support selection.

（7）Krywords

Blacklist words to customize.

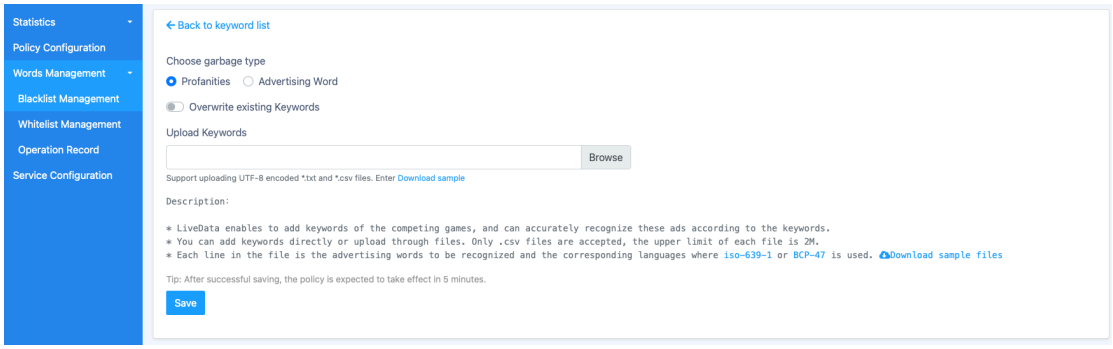
（8）Whether to call the police

If enabled, it means that when the detected text contains the custom vocabulary, the text audit

interface will return the warning field with the value true. (Note: This function can be configured only when "Advertisement" is selected for the garbage type)

Click "Save", the system will automatically save the added vocabulary and return to the blacklist list page.

3.1.2 Upload keywords



Click "Upload Keywords" to enter the vocabulary upload page.

(1) Select the type of garbage

Ditto.

(2) Overwrite existing Keywords

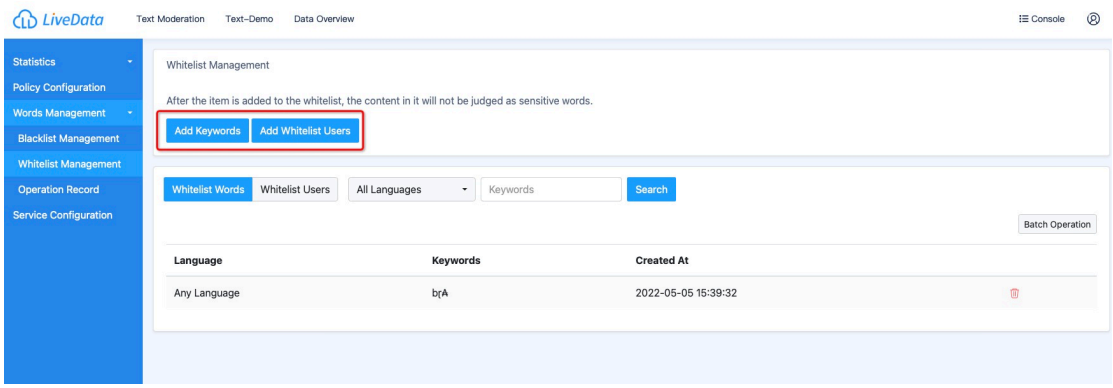
If it is enabled, after saving the uploaded sensitive words, the system will cover all the sensitive words added in the history with the sensitive words uploaded this time, and only keep the uploading results of this time.

(3) Upload Keywords

Click the "Browse" button to select the file to upload (the file format must be the same as the file format in "Download Example")

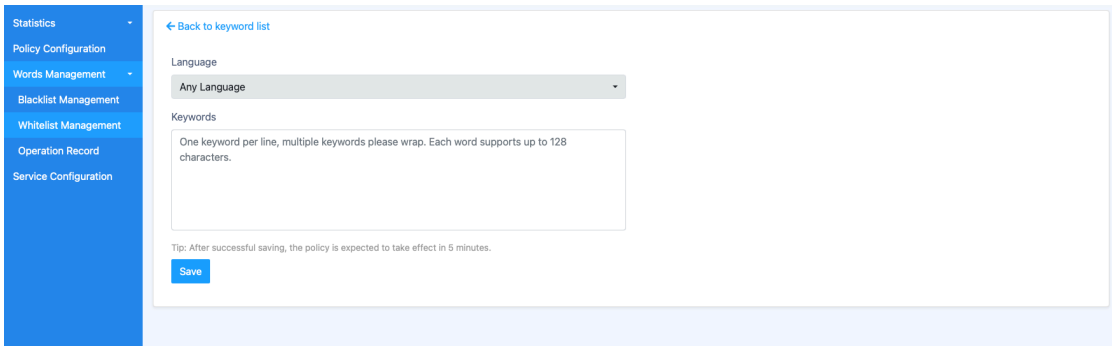
Click "Save", the system will automatically save the uploaded vocabulary and return to the blacklist list page.

3.2 Whitelist Management



You can add words or user IDs to be ignored to the whitelist by directly adding keywords. The added words and user IDs will be displayed in the list below.

3.2.1 Add Keywords



Click "Add Keywords" to enter the vocabulary addition page.

（1）Language

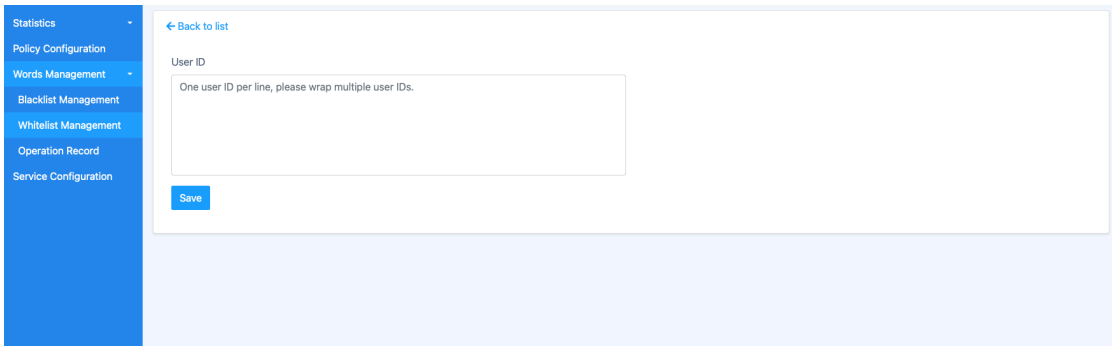
Indicates the language to which the whitelisted word is applicable; that is, only when the word appears in the detected text and is judged as "sensitive" by the system, and the language recognition result is in the applicable language selected by the word, the word will be corrected to "" again "normal" result.

（2）Keywords

Whitelist words to customize.

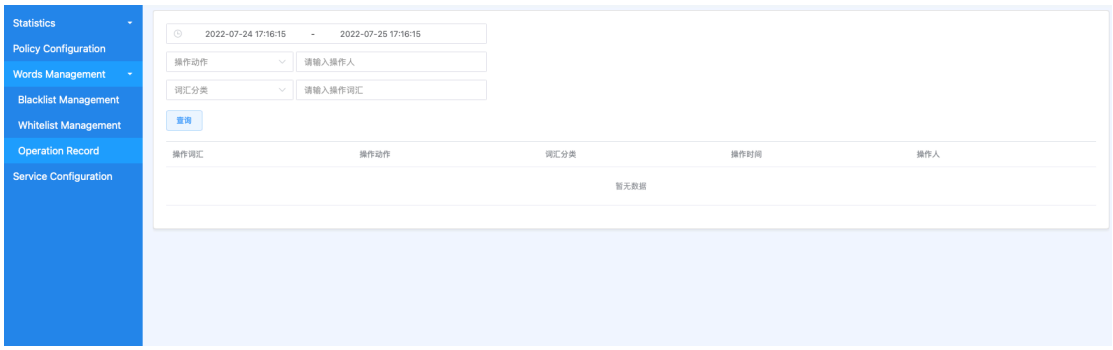
Click "Save", the system will automatically save the added vocabulary and return to the whitelist list page.

3.2.2 Add whitelisted users



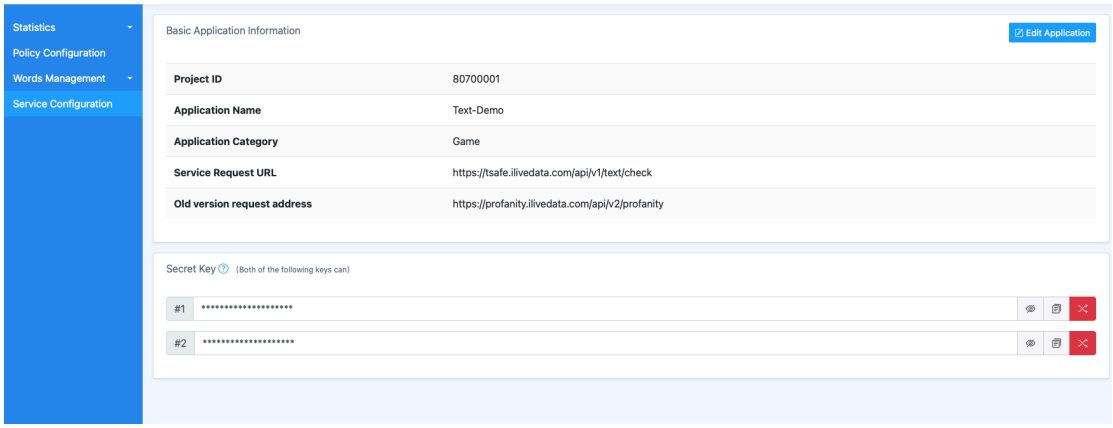
Click the "Add Whitelist User" button to enter the user id add page. You can add multiple whitelisted users at the same time in the form of one user id per line. After saving, the system will return a pass result to the content sent by these users.

3.3 Operation Record



The operation record records the user's operation data on the blacklist and whitelist. You can search for records by filtering conditions such as operation time, operation action, operator, vocabulary classification, and operation vocabulary.

4、Service Configuration



This page displays the item number, name, category, request address, key and other information. At the same time, you can modify the project name, classification and description information through

the "Edit Project Information" button.