Characters of Web Rumors : Analyzing the 2010-2014 Period of Weibo with LDA and Sentiment Analysis

Introduction

(23447575,ZHI Meilin)

In this program, our objective is to analyze the general characteristics of rumors on the Weibo platform. Based on the data set of “Weibo_rumors2.csv”, our investigation will be on three main aspects of rumors: message, source, and receiver, and taking the data set as research samples.Message will be related to the post text, source will be related to rumor account features. The receiver aspect is related to three dimensions: likes, reposts, and comments.

Here are the research questions that we are raised:

RQ 1: What concerns were reflected in rumors?
RQ 2:What are the relationship of rumor emotion and user response?
RQ 3:What are the features of rumor account?

Data Processing and Methodological Framework

(23447575,ZHI Meilin)

1 Data Preprocessing

We used the “re” regular expression package to removed the text noise include punctuation marks, symbols, all the non-Chinese words as well as unimportant marks like “展开 c”, “全部 c”.Then we used jieba in tokenization. To make it more accurate at word segmentation, we added the Chinese emotion dictionary (”dict.txt.big.txt”) as a user dictionary to help it identify the emotion-related words we needed to analyze later. Then, we remove stop-words according to the file ”stop_words_chi.txt”.

2 Key word analysis and Topic Modeling Using LDA

In this part, we want to find out what words and what kinds of topics that rumors are most talking about since 2012 to 2014.

3 Sentiment Analysis

In this part ,we want to analysis the complex relationships of rumor emotion and topic and user responses.We adopted dictionary based sentiment analysis to calculate the emotional value of each posts and then tagged it the “positive” “negative” label for later analysis.

4 Descriptive Statistical graphs for different variables

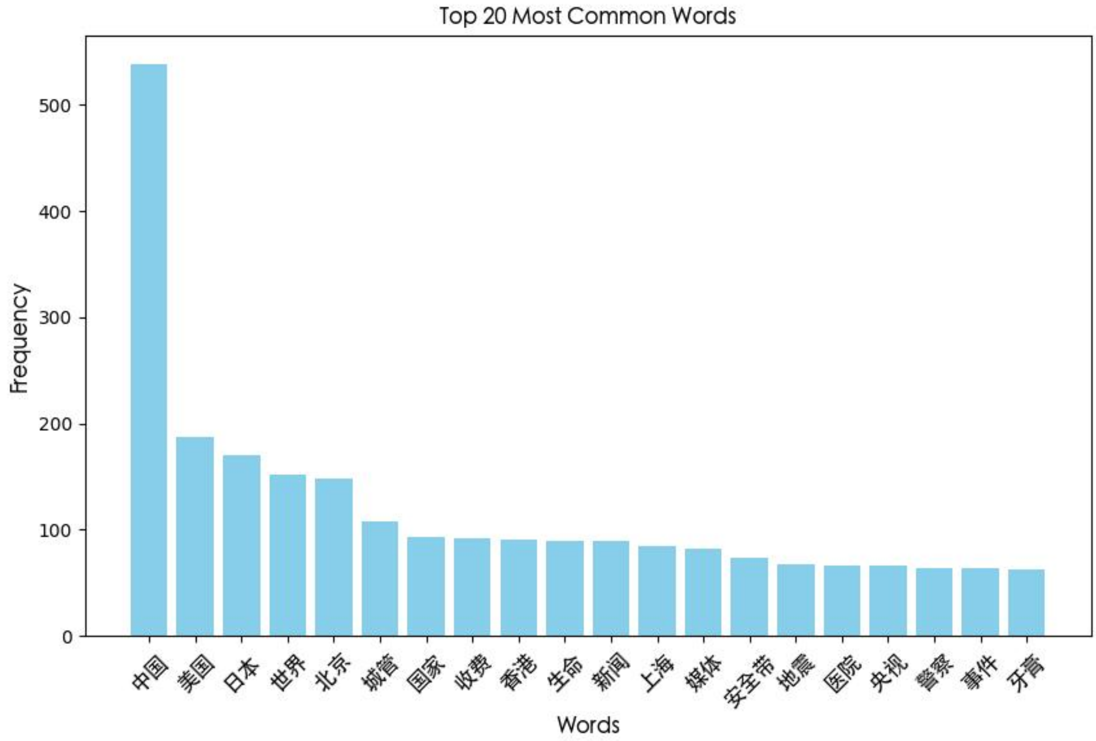
We also visualized the comprehensive information of variables such as time, geographical location, emotional valence, and topic in order to provide intuitive answers to our research questions.These graphs includes word clouds, grouped bar charts, bar charts in timeline and Sankey Diagram.

RQ 1: What concerns were reflected in rumors?

Task 1: Wordcloud & Word frequency (23466073,ZHI Yiran)

1. Methods
1.1 Data preprocessing
In order to improve the representativeness of text data, we use the jieba function to decompose the Chinese text of the post, replace traditional Chinese with simplified Chinese with multiple codes, delete the English and numbers of each post, and use regular expressions to remove all non word characters and characters that are neither word nor space from the text. In addition, due to the subjectivity and complexity of the text used for rumor dissemination, we used the existing list of Chinese stop phrases for filtering, and through filtering attempts, we added words that we found difficult to represent the theme in the original Chinese stopwords list, such as: 结束，只能，特别,年内，决定，一场，继续，感觉，完全，愿意，特别，东西, etc. Because these words have no actual meaning in Chinese expression, we believe they can be deleted.
1.2 .Matching rules
Based on the cleaned data, we have generated a word cloud. As shown in the figure, words related to urban countries such as China, Beijing, the world, the United States, and Japan are particularly prominent. Secondly, there are also many words related to social events such as rape, arrest, and media. In addition, there are words related to emotions such as surprise, happiness, and beauty.

So, we conducted word frequency statistics on the cleaned data and selected the top 20 most frequently occurring topic words. They are: China, the United States, Japan, the world, Beijing, urban management, national, Hong Kong, life, toll, media, news, seat belts, hospitals, Shanghai, police, government, gang rape, toothpaste, South Korea. Among them, 'China' has appeared a total of 538 times, while 'the United States, Japan, the world, Beijing, and urban management' have appeared in the range of 100-200 times, and the rest of the rankings have appeared around 100 times.



Task 2: LDA (23431202, U Chanpong)

For this batch of data, we hope to understand the general theme content and distribution of the posts through the LDA model in the early stage, so as to better understand the various characteristics and attributes of the rumor content from a broader perspective. The following will introduce the process and conclusion of this LDA processing.

1. Methods

1.1 Data preprocessing

In order to improve the representativeness of text data, we used the jieba function to break down the Chinese text of the post, replaced traditional Chinese with simplified Chinese using multiple codes, deleted the English and numbers of each post, and used Regular Expression to remove all non-word characters and characters that are neither word characters nor spaces from the text. We also used the existing Chinese stopwords list for filtering, and added vocabulary that we thought was difficult to represent the theme. Finally, we filtered about 24,021 words. Afterwards, we implemented the LDA model using "LdaModel" and set up 10 topics, each containing 10 words.

1.2 Matching rules

Based on the keywords extracted by LDA and the feature extraction of related posts, the following 10 topics were obtained. Two of the themes are relatively consistent, so they are collectively referred to as "民生".

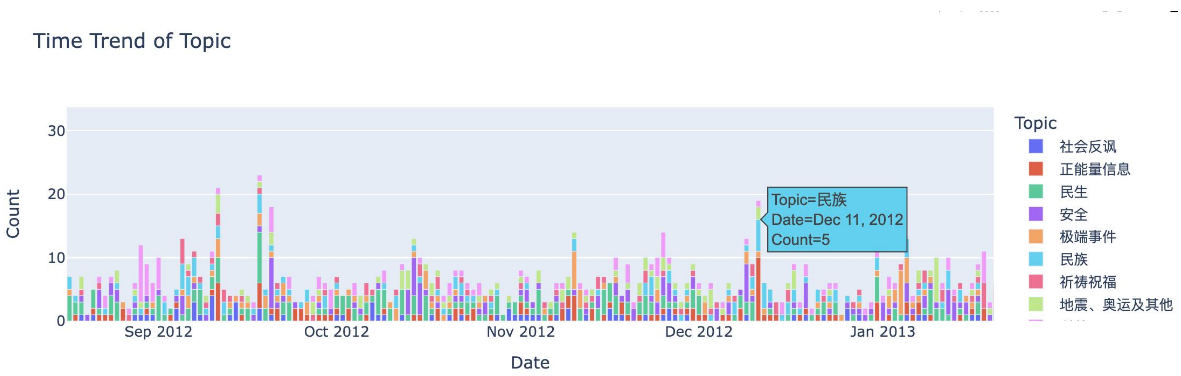
Topic	Key Words	Proportion
民族	中国、日本、美国、票房、贞子、上映	11.03%
安全	安全带、行驶、飞机、交通、处罚、抽烟、酒驾	9.27%
民生	中国、广州、视频、免费、大闸蟹	8.12%
民生	中国、火车、公司、盒饭、高考	12.85%
社会反讽	现实、父母、成都、黑匣子、蛟龙	8.79%
科普	牙膏、天然、动物、化学、黑色、绿色、红色	10.52%
祈祷祝福	蜡烛、刘翔、生命、死亡、健康	9.33%
极端事件	轮奸、女子、农村、大学、影响	8.24%
正能量信息	结婚、幸福、鼓掌、帮忙、谢谢	10.52%
地震/奥运相关	地震、奥运、现场、孩子、熊猫、加油、生活	11.33%

At the same time, we will showcase iconic posts for each topic.

Topic	text
民族	各位注意：5月12日不要去电影院。 请大家一定别进影院，大家一起为《贞子》票房为零，做努力！ 中国人拍的《金陵十三钗》在日本小鬼子票房为零。小日本拍的《贞子》3D将于5月12日在中国大陆上映。而5月12日既是南京大屠杀纪念日，又是国难日。勿忘国耻！！作为中国人，敢不敢让贞子3D 5月12日票房为零美国护照中写着：不管你身处何方，美国政府都是你强大的后盾。在中国护照中写着：请严格遵守当地的法律，并尊重那里的风俗习惯。微评：美国说：出去了有人欺负你，招呼一声咱修理他！中国说：出去了老实点，听人家话，少给老子惹麻烦！
安全	今天下午六点开始， 高清.探头全部启动，副驾驶室不系安全带相同处罚，开车时打电话罚款50元，闯黄灯罚200，越线停车罚100，今天起晚六 点半至深夜二点，为期60天，全国交警集中查处酒驾，一经查获，一律拘役六个月，五年内不得考证。每月10、20、30日由省厅带队检查，每月3、7、13、17、23日今天下午六点开始,太原市区高清.探头全部启动,副驾驶室不系安全带相同处罚,开车时打电话罚款50,闯黄闪罚200,越线停车罚100,今天起晚六点半至深夜二点,为期60天,全国交警集中查处酒驾,一经查获,一律拘役六个月，五年内不得考证。
民生	【爆炸消息：中国移动9月1日起对微信微博收费】两种收费标准：1，包月10元/月；2、按条数收费，100条以下免费，600条5元，600至1200条10元，超过1200条每条收两分钱。联通牵手百度制衡微信，电信与网易合作以翼信对抗微信；转发自@当代章鱼哥【敢吃大闸蟹的人，是世界上最“勇敢”的人】《壹周刊》记者把12只大闸蟹送去化验，结果发现11个样本有土霉素，6个样本有氯霉素。土霉素属“过时”抗菌素；而氯霉素属香港违禁物质，因会压抑骨髓功能，导致贫血、抵抗力下降和凝血困难问题。孕妇吃了含有土霉素的毒蟹，胎儿的骨质会变灰、变脆。
民生	【新闻联播：火车盒饭只要5元！ 你吃过吗？】《新闻联播》记者亲身登上火车，报道火车上的东西一点都不贵。网友吐槽：“央视记者是怎么买到火车票的？”、“能不能别逗了？5元太能扯了！最便宜矿泉水都要五块以上，还能买盒饭？播新闻联播的人难道没做过火车？PS：继你幸福吗，央视又一大亮点~[围观]捡到粗心小朋友的高考准考证，谁认识通知下，别耽误了高考大事。 白娅倩 考点市一中 考场013 座号11 准考证号204101311 / 请大家帮忙互相转发，别耽误了孩子高考！联系电话138-3046-8131 转发一下 求扩散@洛阳容威 @张晓理-老牛 @车炳恒 @火箭 我在这里
社会反讽	【出来混迟早是要还的】那个MH370的黑匣子不是被探测到在南太平洋4000到6000米的海沟里嘛？当我听到这个消息的时候马上就兴奋了起来，因为我依稀记得蛟龙号第五次下潜最大深度达7062米。那搜这个黑匣子还不是小意思吗？刚想@蛟龙深潜 突然发现蛟龙成了乌龙了微博都被删光了一刚！（转）一名利用假期卖菌的四川万源学生,遭到数名城管殴打,原因不明。学生被迅速送往医院,目前仍在昏迷。马上要转往成都。万源不是一个小县城，安居乐业挺好的,。这些城管谁他们给你们这么打的权利可以掌控人的生死。临武的悲剧才过去多长时间啊?伤疤还没好,就忘了疼？《 ...
科普	【让电脑提速的小方法】开始菜单里点击"运行"，输入gpedit.msc指令后确定，会出现一个小屏幕，然后单击计算机配置--管理模板--网络--QoS数据计划程序，选“限制可保留带宽”，“设置”选“已启用”，将“带宽限制”中的20改为0，“应用”，就可以使用100%的网速,因为平时Windows XP自动保留20%的网速。刚学了个知识，和大家分享一下：如何判断你的牙膏是否健康。在每个牙膏袋的尾部都有个色块，绿色表示牙膏成份纯天然，蓝色表示天然且含药物，红色表示含化学成份，黑色表示纯化学成份合成。建议大家使用绿色和蓝色的，快看看你家的牙膏健康吗? via 桃桃琳
祈祷祝福	真相：【刘翔在医院图】最后一张慎点！！！看到心都碎了！！！[伤心]那些骂刘翔的人看清楚了！【请为这位小女孩点上一支[蜡烛]】这位戴着眼镜努力往前跑的1035号小朋友，今年8岁。在美国东部时间4月15日下午2点30分，波士顿马拉松爆炸事件中失去了生命。小宝贝，天堂的路没有恐怖袭击，天堂的路可以自由奔跑……（人民日报）[蜡烛][蜡烛]
极端事件	人间惨剧：今天下午约14点，宁波妇儿医院，一妇女携带一婴儿在住院楼跳楼，后抢救无效死亡。具体情况有关部门正在调查。据现场网友称妇女因小孩病重，加上负担不起昂贵的医疗费，带着刚满月的宝宝从12楼跳楼身亡。[蜡烛] 底层民众的医疗费用猛于虎，国人的性命其何等脆弱！[泪]【炒房赔进5000万：美女光身子跳楼自杀！】为炒房，她7万一平在绿城鹿城广场买了28套房子，现在5万一平也脱不了手，一套房就要赔200多万，因买房的钱多是向亲戚借的，昨晚，她跳楼自杀。——鹿城广场是温州最高房价水平的代表，2010年巅峰时二手房报价已经接近10万一平
正能量信息	【保护小动物们的好消息来了！】上海已开通全国首个卖猫狗肉的举报电话！只要拨打食品药品安全投诉热线12331就可举报，而且政府奖励¥500！！！已经有网友举报成功并拿到了奖金！这可谓是新年最好的消息之一了，爱猫狗的亲们赶紧转起来吧！！！[good][good][good]给即将过去的2012年说声感谢 谢谢那再也回不来的过去 谢谢时间给我留下最好的朋友 最爱我的家人、最棒的嘉园～怀着感恩的心迎接2013平安
地震/奥运相关	#狮子月助力刘翔夺冠#如果刘翔最终获得伦敦奥运会金牌，凡转发本微博并关注@给狮子座的999封信 的网友，所有人送一台三星 I9300 GALAXY SIII。记住，不是一个人，而是所有人。信誉保证，决不骗人，到期兑现！[爱心传递][爱心传递][爱心传递]感动又被萌翻，地震发生时，一只害怕的熊猫死死地抱住警察叔叔的腿。#众志成城、抗震救灾、雅安加油#[爱心传递][爱心传递]爱心传递！为雅安祈福！[蜡烛][蜡烛]

Task 3:The trending topics posted on the timeline (ZHI,Meilin)

We classified all the rumors based on their topics and presented them on a daily basis. This allows us to identify significant time periods and observe the changing frequency of mentions for different topics within those time periods. It also helps us formulate research questions for further in-depth statistical analysis.



Findings of topic analysis

From the statistics of word cloud and word frequency, the content of Weibo rumors mainly focuses on countries such as China, the United States, and Japan, as well as regions such as Beijing, Hong Kong, and Shanghai. Other frequently occurring words have a high correlation with social events, such as life, seat belts, and police.

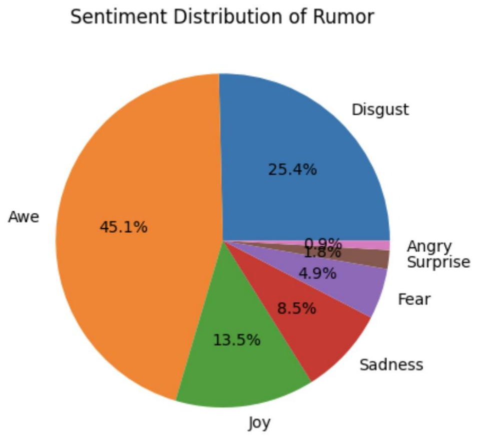
As can be seen from the above posts, the content sources are mainly news and ordinary users, while the rumor themes mainly focus on national issues, popular science knowledge of social and livelihood issues, among which livelihood issues account for the highest proportion (> 29.76%).

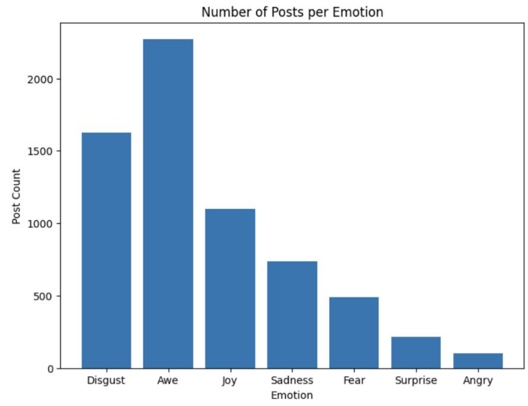
At the same time, it should be mentioned that the theme recognition ability of LDA for this dataset is relatively insufficient. Even using Hyperparameter Tuning pages cannot accurately recognize most posts. Therefore, in the process of improving accuracy, we can only try to improve its accuracy by adding more stopwords and fixed recognition words.

RQ 2:What are the relationship of rumor emotion and user response?

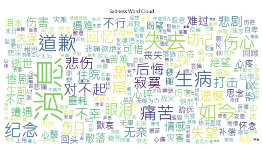
1.Descriptive statistical Analysis (23447575,ZHI Meilin)

To answer this question, we primarily used a dictionary-based sentiment analysis approach. Firstly, we generated several descriptive statistical graphs to provide general visual insights.By going through the words in each post,and matching the sentiment type for that word in the sentiment dictionary, we calculated the sum of the various sentiment type in each post.Meanwhile, we count the number of post of each sentiment. The analysis revealed that the rumor texts dominantly conveyed the emotions of awe, disgust, and joy, ranking as the top three emotions expressed.

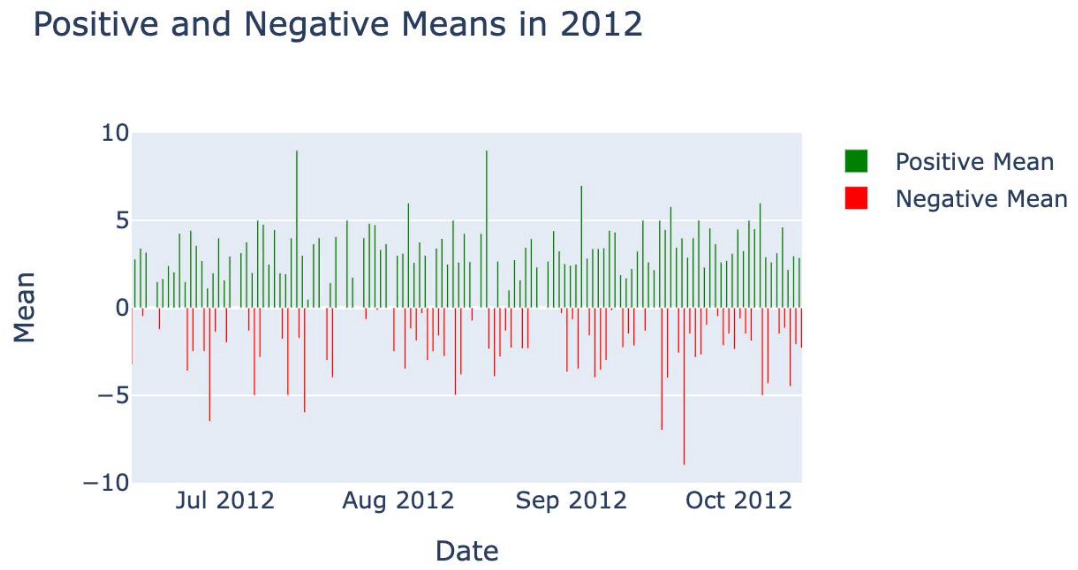




Secondly, we created word clouds to identify specific emotion-related words mentioned in each post. This approach is helpful in extracting emotion-related words from the post text, which can be further utilized to train machine learning models for contextual analysis. The word clouds also provide insights into the choice of words used by rumors to convey their emotions.



Finally, we also took the time variable into consideration. We transformed the time column into the "year-month-day" and "year-month" patterns and labeled them as new columns to facilitate efficient data extraction. Then, we calculated the mean emotion value for each post and depicted the results on a daily basis.By analyzing the resulting graph, we can identify the date when emotions peaked and observe any interesting temporal changes in the rumors. This helps us understand the specific time periods during which the rumors discussed certain emotions for our further study. Here is an example result from 2012.



2.The Relationships between rumor emotion and user response (ZHI Meilin)

Firstly, we wanted to find out whether different sentiment types can lead to different tendencies to share, comment, and like.To make different types comparable, mentioned times are used as weights and a weighted average is calculated for each response.Comparing the results, we found that Awe and joy in the rumor post are the most likely to get users toresponse and are the most likely to be reposted.

The formula is :

*sum(average_each * mentioned_times_each)/sum(mentioned_times_total)"*

Sentiments	Mentioned times	Avg Comments	Avg Shares	Avg Likes
Awe	6045	155.761989	414.839859	43.748790
Disgust	3404	143.190652	408.178967	34.569496
Joy	1809	171.083560	416.267938	57.841054
Sadness	1135	154.691475	426.442490	38.894452
Fear	655	125.816701	393.716904	27.798371
Surprise	246	147.187793	394.347418	38.075117
Angry	122	128.615385	447.173077	16.182692

The sentiment dictionary provide each emotion word with sentiment type,intensity and valence.After the comparison of different sentiment types,we calculated and labeled the valence of each posts.We wanted to find out the correlation of rumor valence and user responses.To make it comparable, mentioned times are used as weights and a weighted average of each sentiment type is calculated.

The formula is :

emotion_value=

*sum(specific_senti-word_times * intensity * valence) / total_senti-word_times*

According to the rule that negative emotion is negative, positive emotion is positive and neutral is 0, we put the corresponding tag of valence according to the emotion value of the posts.Then we did a correlation test.The result shows that the higher the emotional value (positive) of a post, the fewer responses (comments/shares/likes) it receives from users. Conversely, more negative posts tend to elicit a greater response. There is a weak correlation between emotional polarity and reposts, while comments and likes exhibit a strong positive correlation, which is statistically significant.

	Correlation	P-value
comments	0.036523	0.035906
reposts	0.026195	0.132459
likes	0.065468	0.000168

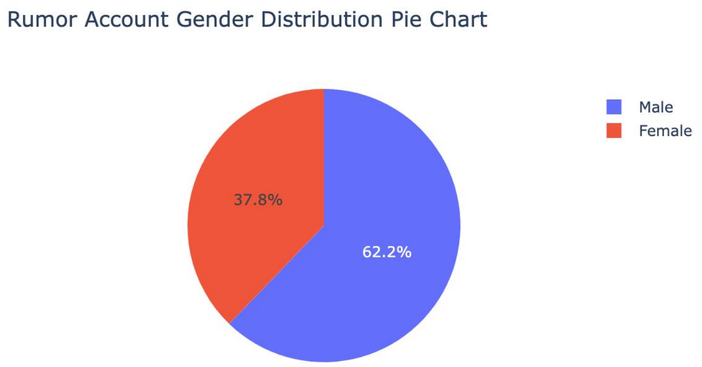
RQ 3: What are the features of rumor account?

Rumor Account Analysis (SHI Hantong)

To understand how these rumor sources influence user , we need to examine both the identity of communication subject and how they frame their posts on socail internet. The former is mainly focused on its basic information about the accounts, from which we can figure out their gender distribution, geographical distribution, verified type, and number of followers.

Gender Distribution (ZHI Yiran)

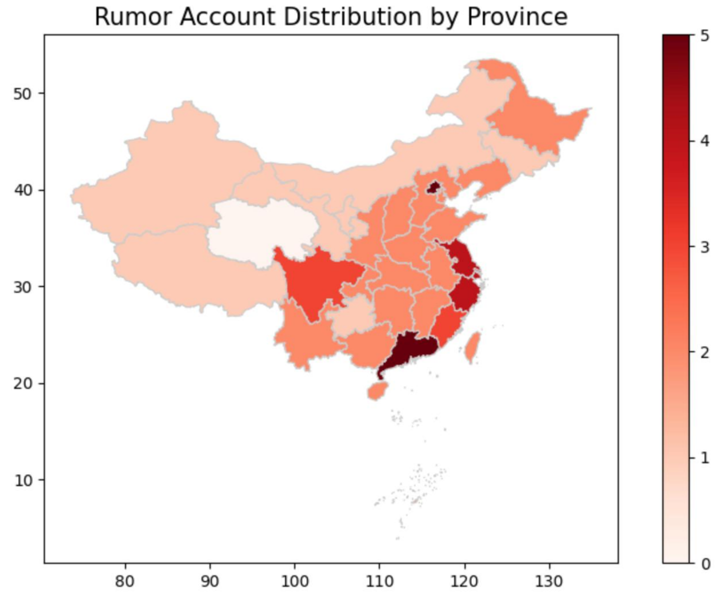
In the form of a pie chart, visual analysis was conducted on the gender of accounts involved in Weibo rumors. According to the results, among the groups involved in rumors on Weibo, males accounted for 62.2% and females accounted for 37.8%.



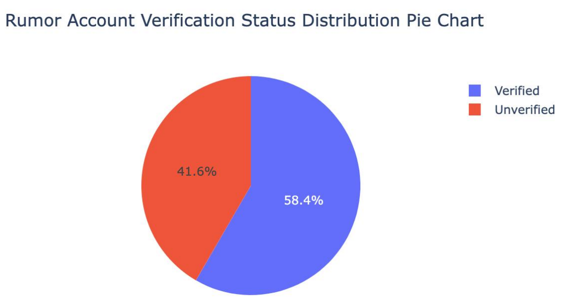
Geographical Distribution (SHI Hantong)

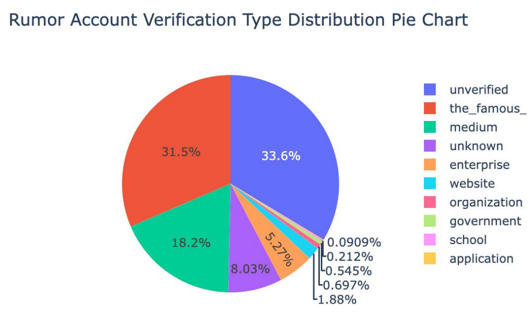
Based on the geographical distribution data of rumor accounts, we created a distribution map of rumor accounts in different provincial-level administrative regions in China. The color shades indicate the density of the number of accounts. Dark colors indicate a larger number, and light colors indicate a smaller number. Because the data is too large, we divide each order of magnitude into 6 gradients, where level 0 means no account distribution, level 1 means there are 0-10 accounts, level 2 means there are 10-50 accounts, and level 3 means there are 50-100 accounts. For accounts, level 4 means there are 100-500 accounts, and level 5 means

there are more than 500 accounts. We observed the following:
Significant regional differences: By comparing the color differences in different regions, the quantitative differences between provinces can be visually displayed. Among them, the darkest areas are Beijing and Guangdong, which have the highest order of magnitude, with more than 500 accounts each. Beijing is the area with the highest concentration of rumor accounts in the country, with more than 1,000 accounts. The accounts are mostly distributed in the southeastern coastal provinces, and the northwestern region is lighter. Among them, Qinghai Province is the least with no account distribution.
The account distribution density in the coastal provinces is higher: places such as Guangdong, Zhejiang, and Shanghai show darker colors, indicating that there are more rumor accounts in these areas. This may be because these areas have developed economies, dense populations, and high Internet penetration. The use of social media is also relatively high.



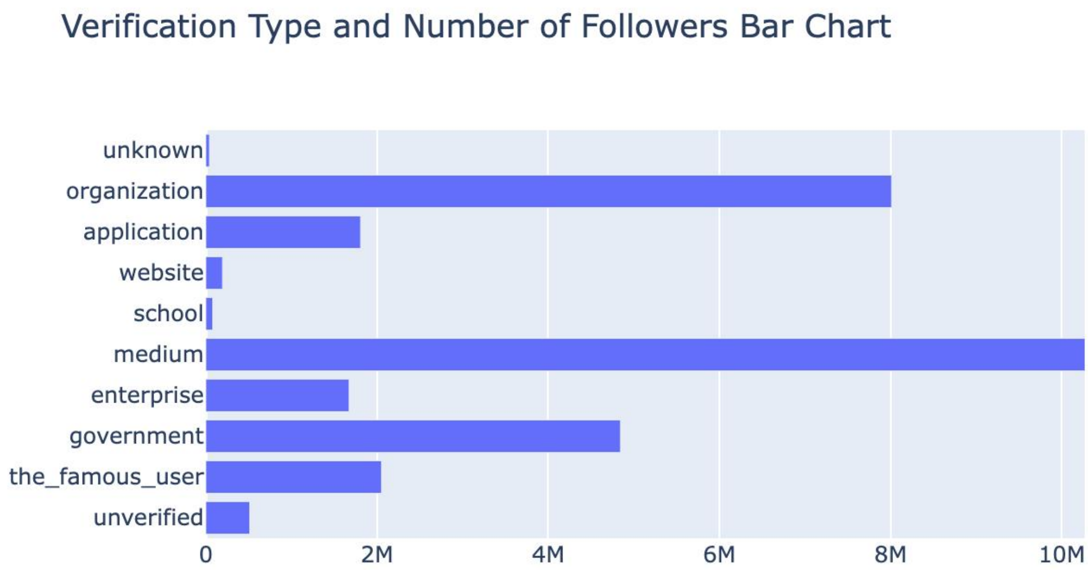
Verified Type (SHI Hantong)
According to the verified type of the account, we found that rumor accounts can be divided into two categories: verified and unverified. The verified types include the famous user, medium, enterprise, website, organization, government, school and application, wheras the unverified types include unverified and unknown. Surprisingly, among all rumor accounts, verified accounts accounted for more than half, reaching 58.4 percentage points. Among all types, unverified accounts are the largest subcategory, with 33.6% of accounts being unauthenticated. Among verified accounts, the famous users or opinion leaders are the largest, accounting for 31.5% of the total, followed by media accounts at 18.2%, followed by enterprise, websites, organizations, governments, schools, and applications.
To analyze, we can understand that the individual accounts are dominant among all the accounts because most of the social media users are individuals. Compared to the regular organizations, media has found to contribute a prominent part of romors because they play important roles on information dissemination. Account types like government and enterprize are relatively cautious because they are authentic and consider more about their reputation.





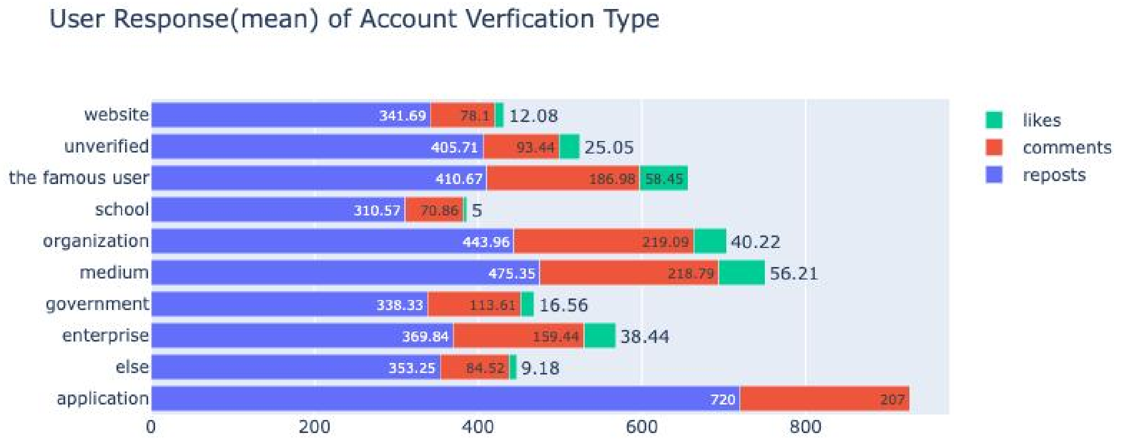
Number of Followers Under Different Verified Types (SHI Hantong)

After an analysis of the account verification types, we proceeded to make a statistical analysis of the number of followers under different types of accounts. As can be seen from the bar chart below, the number of followers of different types of accounts is significantly disparate. The account type with the largest number of followers is Media, with an average of 12.4 million fans. This is followed by the organization's 8 million fans. For personal accounts and website accounts, although there are many accounts, the average number of fans is relatively low.



User Response Under Account Verfication Type (ZHI Meilin & SHI Hantong)

Analyzing users' interactive responses under different types of accounts, we found that users are more inclined to engage in the reposting of information when discussing rumors. This represents a significant proportion of overall user interactions across all account types. In contrast, the less common forms of user engagement on rumor blogs are comments and likes. It can be seen that because users lack the ability to distinguish rumors, when faced with false information, actively spreading it is the most likely way to interact.



Findings (SHI Hantong)

The geographical distribution is uneven, rumor accounts mostly concentrated in the southeastern coastal provinces and economically developed areas.

Relative to institutional accounts, personal accounts (whether verified or not) account for the vast majority.

Although there are many individual rumour accounts, personal accounts have relatively few followers, while verified accounts such as media and organizations have huge numbers of followers.

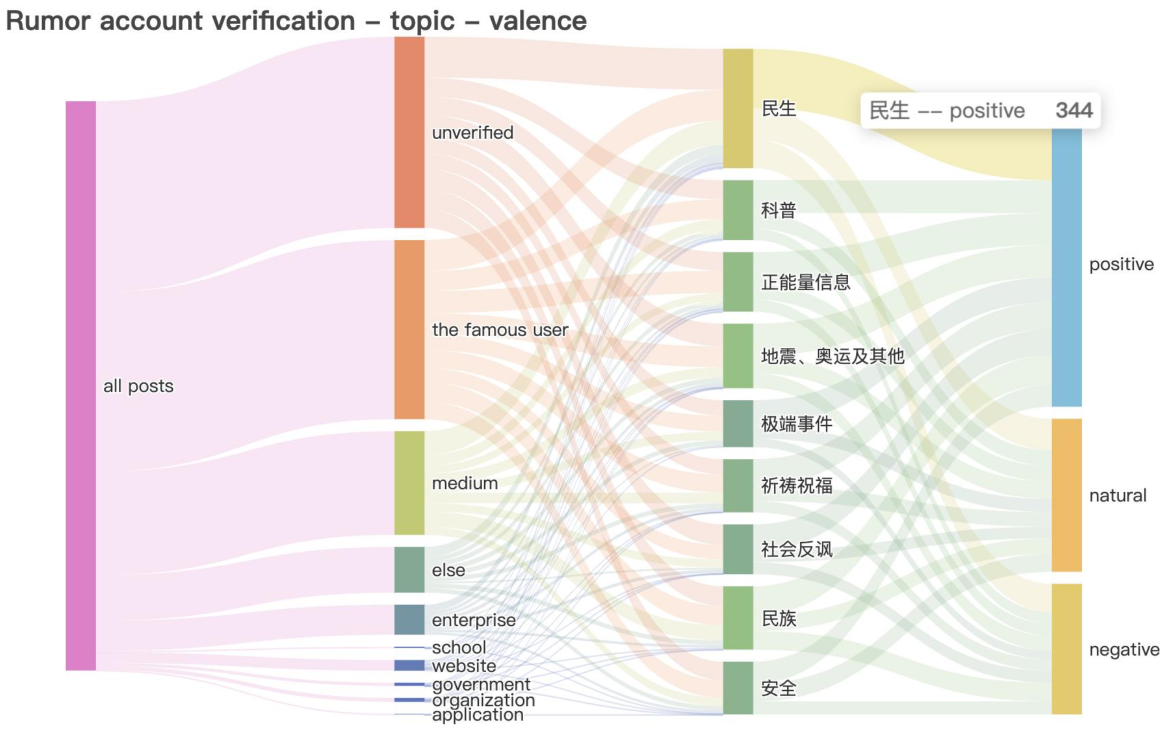
The vast majority of user interaction methods are reposting.

Comprehensive Information of

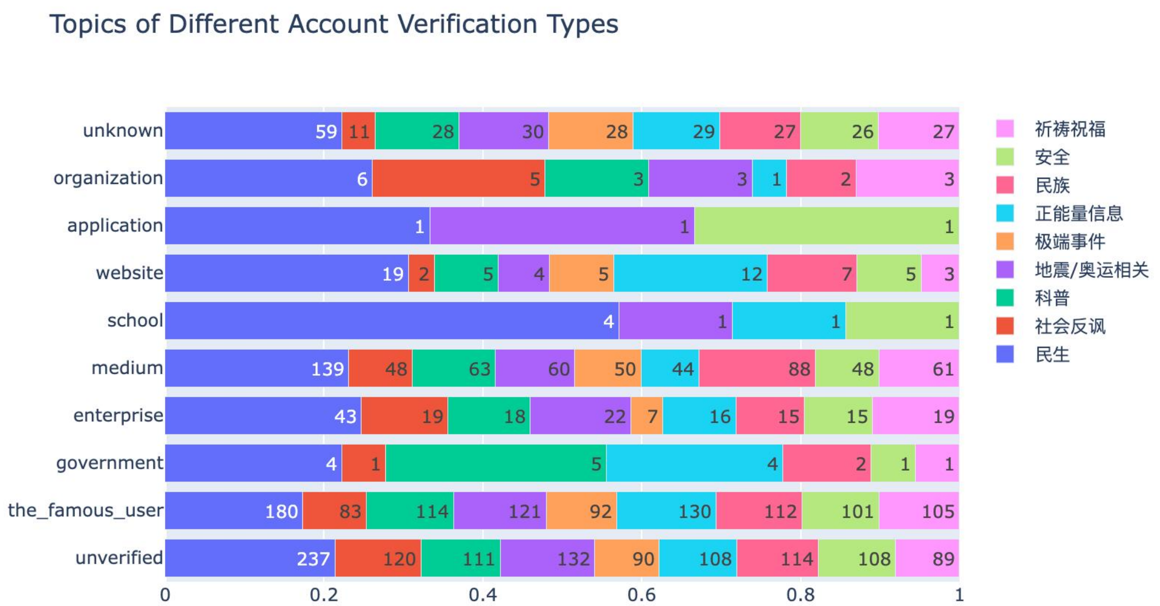
Rumor Accounts, Topics, Emotional Valence, and User Responses

Sankey Diagram (ZHI Meilin)

We visualized the relationships between variables related to rumors through a multi-level sankey diagram, including the verification types of rumor accounts, rumor topics, and the emotional valence of the rumors. This helps us in identifying interesting findings. Subsequently, we created bar charts for each relationship segment to examine the remaining characteristics of the rumors.

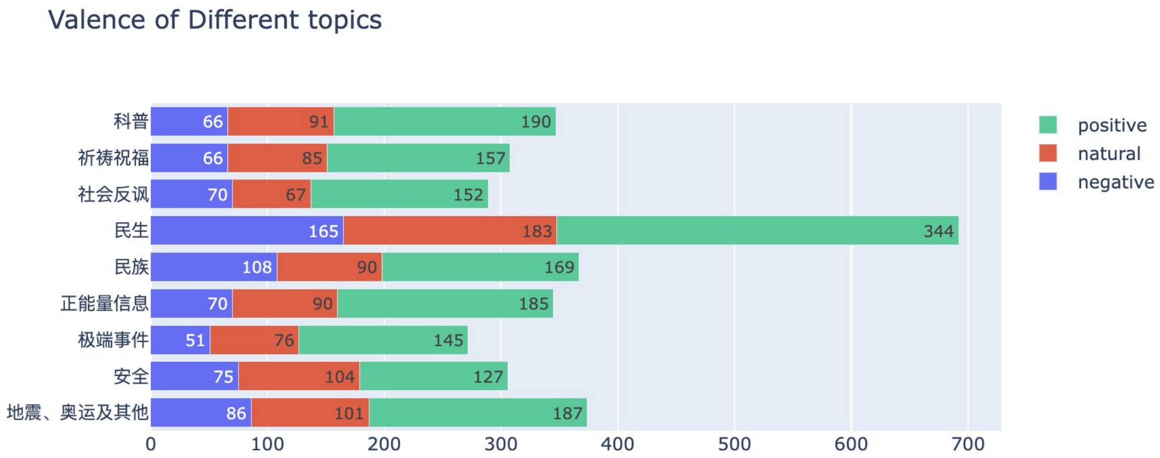


1 Topics of Different Account Verification Types (SHI Hantong)
From the bar proportion chart of topic types published according to different account types, we can intuitively find that among the content published by each type of account, rumors about people's livelihood topics account for the largest proportion. The remaining rumor topic types are evenly distributed.



2 Valence of Different Topics (ZHI Meilin)
From this chart, we can observe that rumors on various topics tend to convey more positive

emotions. Among these, rumors related to livelihood topics have the highest amount of both positive and negative emotional content.



3 User response of Different Topics (ZHI Meilin)

This comparison clearly tells us that the most common act people have to rumors is to express their response through reposting. The topics that most effectively garner discussion are those related to positive messages, extreme events, and issues concerning ethnicity and nation. Topic that receive the most expressions of likes is that involves prayers and blessings.