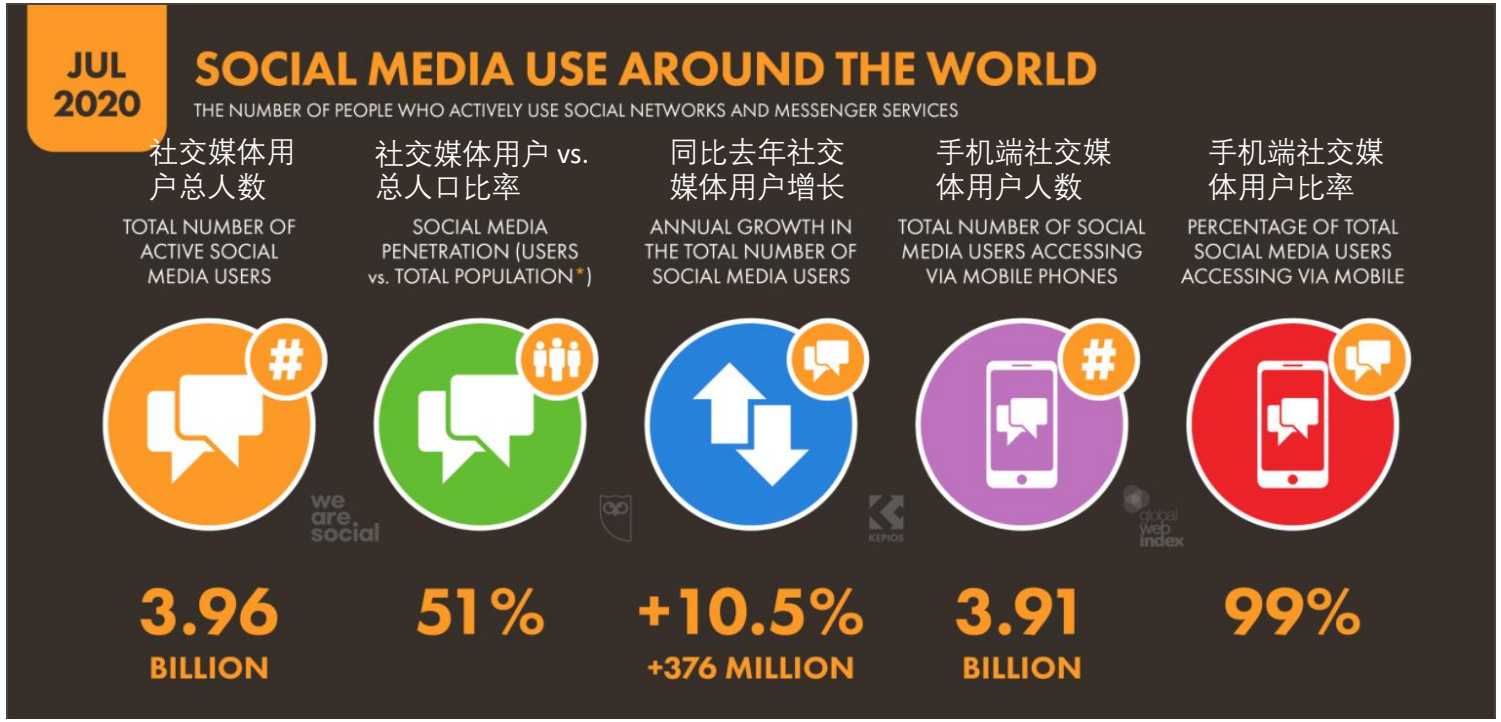
网络社交媒体虚假信息监测

姜也
英国谢菲尔德大学
计算机科学学院
自然语言处理组

信息疫情 infodemic

- 信息疫情是指在疾病爆发期间出现过多的虚假或者误导性信息 ---- WHO(2020)



https://datareportal.com/reports/more-than-half-the-world-now-uses-social-media

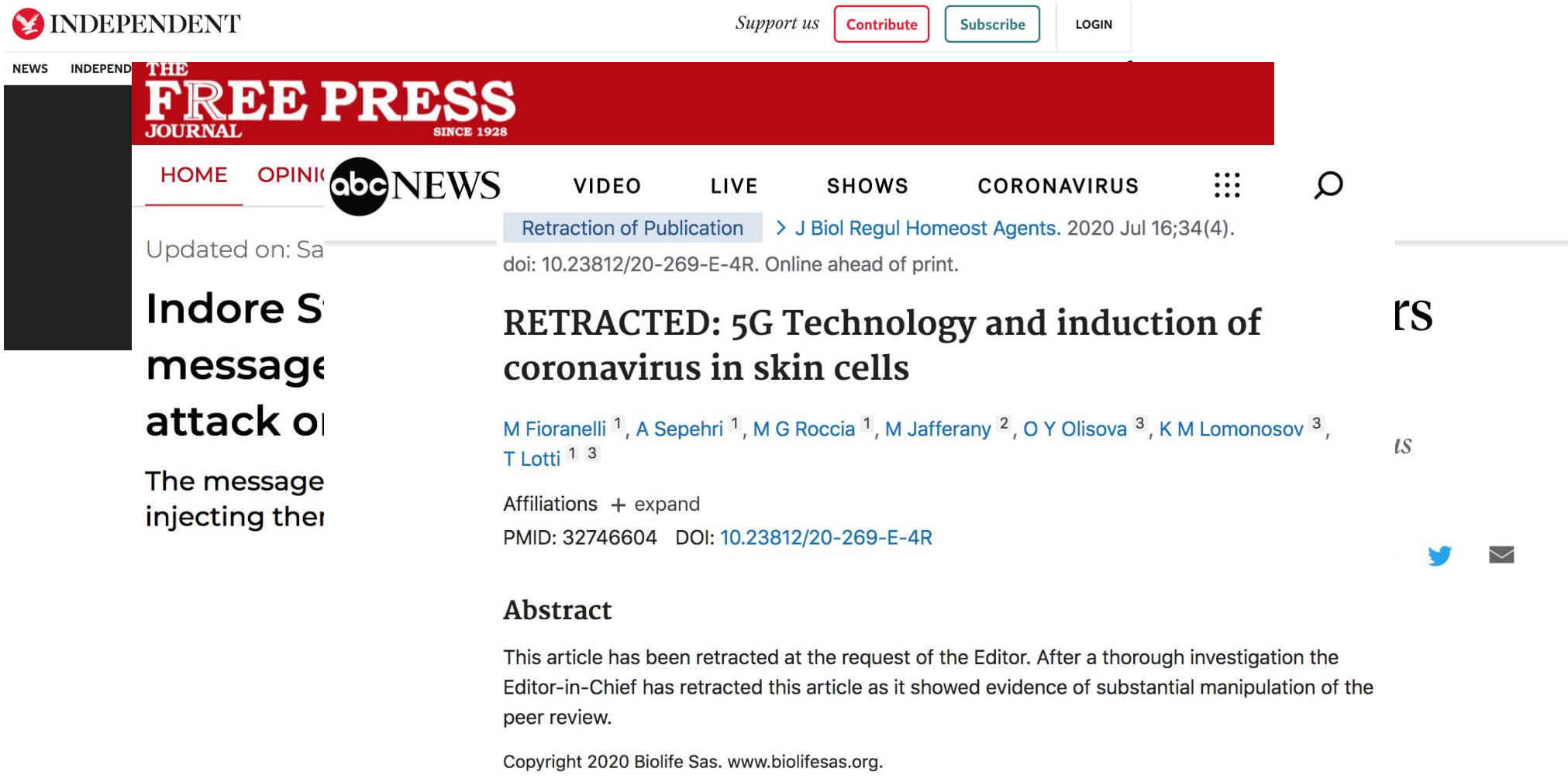
信息疫情 infodemic

Facebook中虚假信息网站链接的阅读量大大高于权威网站链接的阅读量。

Figure 1.4: Comparing the total estimated views to content from the top 10 health misinformation sharing websites with equivalent content from 10 leading health institutions by country/region.

https://secure.avaaz.org/campaign/en/facebook_threat_health/

信息疫情 infodemic 案例



INDEPENDENT

Support us Contribute Subscribe LOGIN

NEWS INDEPEND

THE FREE PRESS JOURNAL SINCE 1928

HOME OPINI

abc NEWS VIDEO LIVE SHOWS CORONAVIRUS

Updated on: Sa

Indore S
messag
attack o

The message
injecting ther

Retraction of Publication > J Biol Regul Homeost Agents. 2020 Jul 16;34(4).

doi: 10.23812/20-269-E-4R. Online ahead of print.

RETRACTED: 5G Technology and induction of coronavirus in skin cells

rs

ts

M Fioranelli [1], A Sepehri [1], M G Roccia [1], M Jafferany [2], O Y Olisova [3], K M Lomonosov [3], T Lotti [1 3]

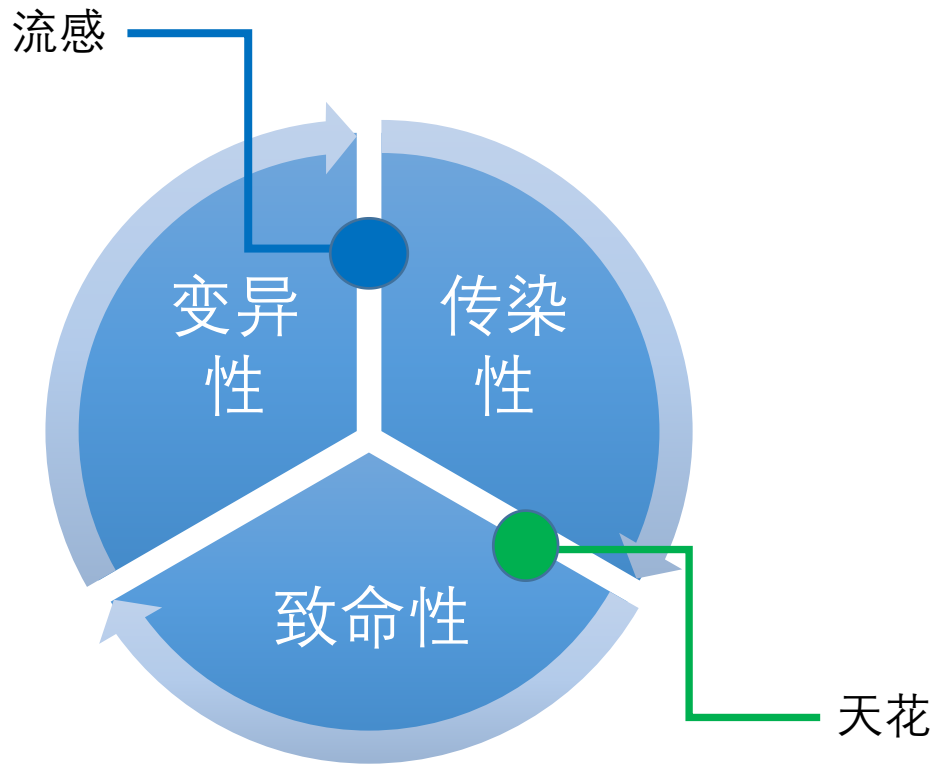
Affiliations + expand

PMID: 32746604 DOI: 10.23812/20-269-E-4R

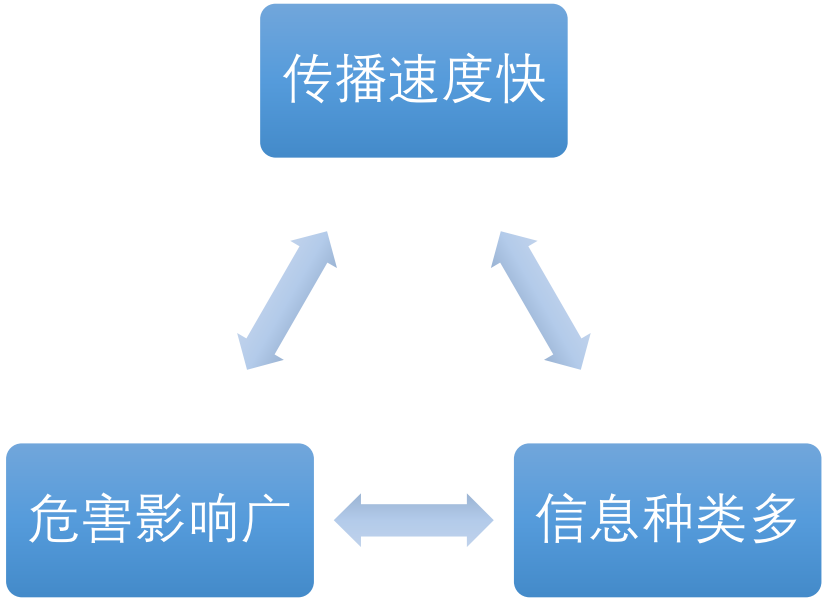
Abstract

This article has been retracted at the request of the Editor. After a thorough investigation the Editor-in-Chief has retracted this article as it showed evidence of substantial manipulation of the peer review.

信息疫情 infodemic vs. 病毒疫情 pandemic



病毒的三角平衡



信息疫情的特点

全球factcheking组织

Factcheck的困难与应对方案

- 人工查核耗时耗力
- 网络信息更新迅速
- 查核信息关注量低

- “半”自动虚假信息识别
- 提高揭露虚假信息主题多样型
- 提高揭露虚假信息关注度

虚假信息检测总体流程

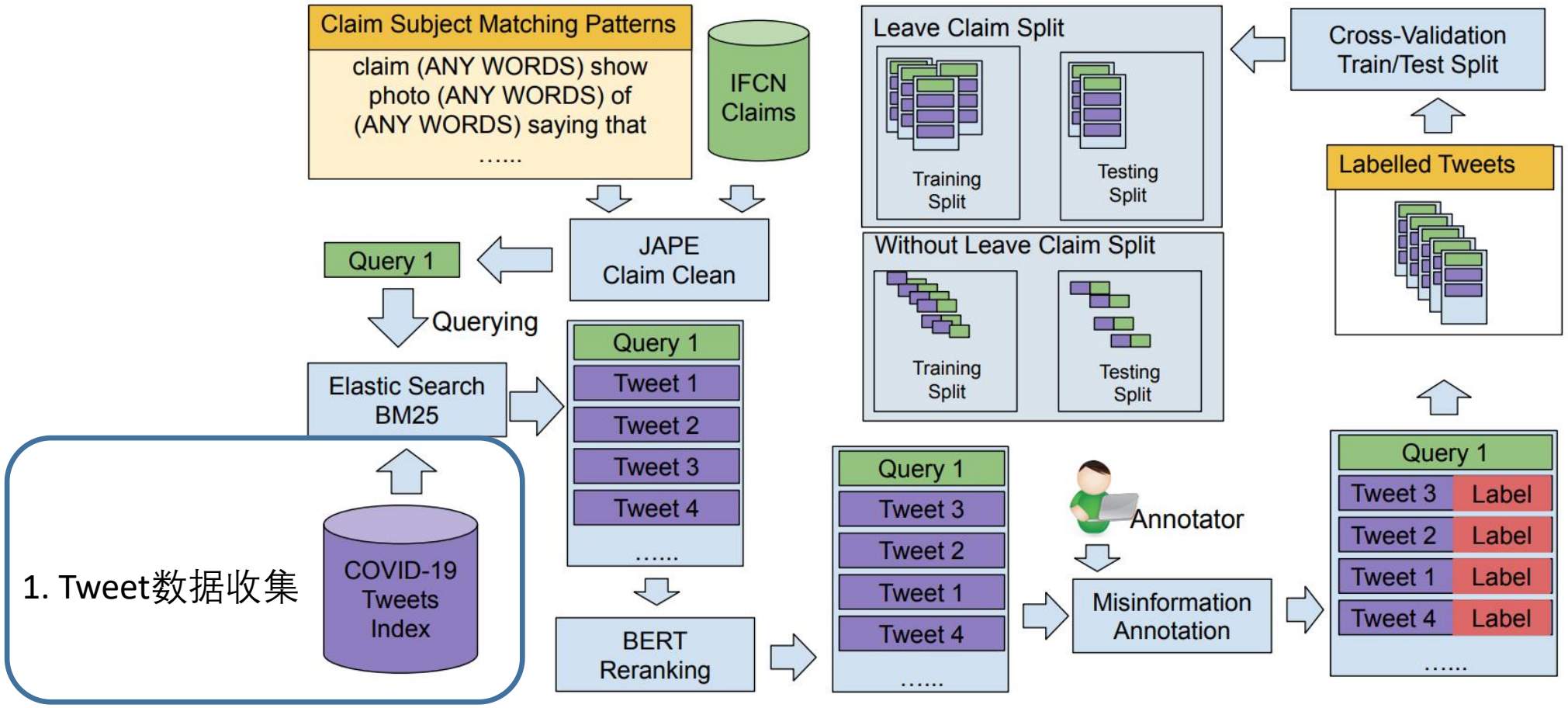


Figure 1. Overall pipeline

Step 1. twitter 数据收集

- Twitter Stream API (https://developer.twitter.com/en/docs/tutorials/consuming-streaming-data)；
- 关键词定义： Covid19, Coronavirus, Covid等；
- 收集数据从2020年3月至今；
- Json数据结构，包含用户基本数据，文本数据，引用和转发量等；

虚假信息检测总体流程

2. IFCN数据收集与整理

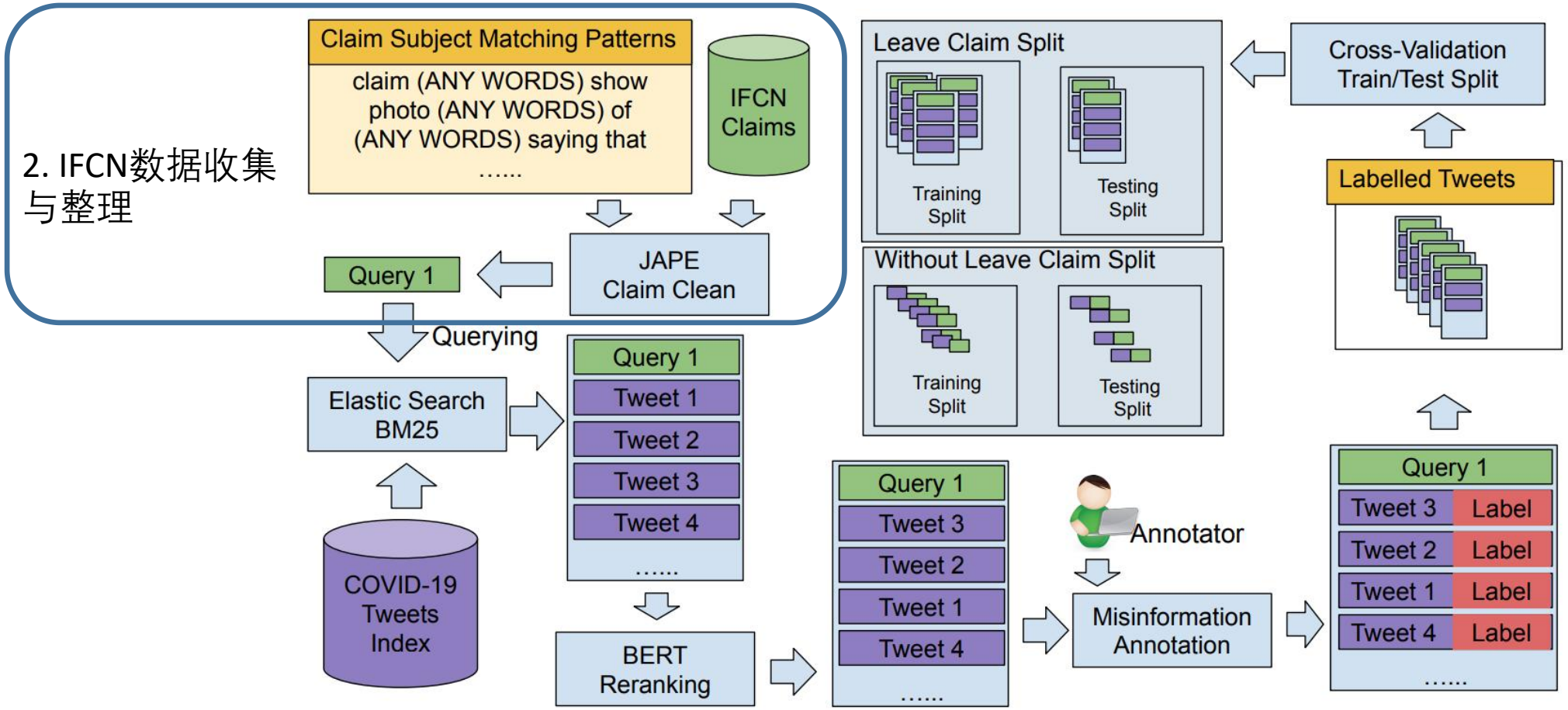


Figure 1. Overall pipeline

step 2. ifcn数据收集与整理

- 爬虫获取IFCN网站上关于新冠病毒相关的数据;
- 爬取数据清理(beatutifulsoup4+rule based)

虚假信息检测总体流程

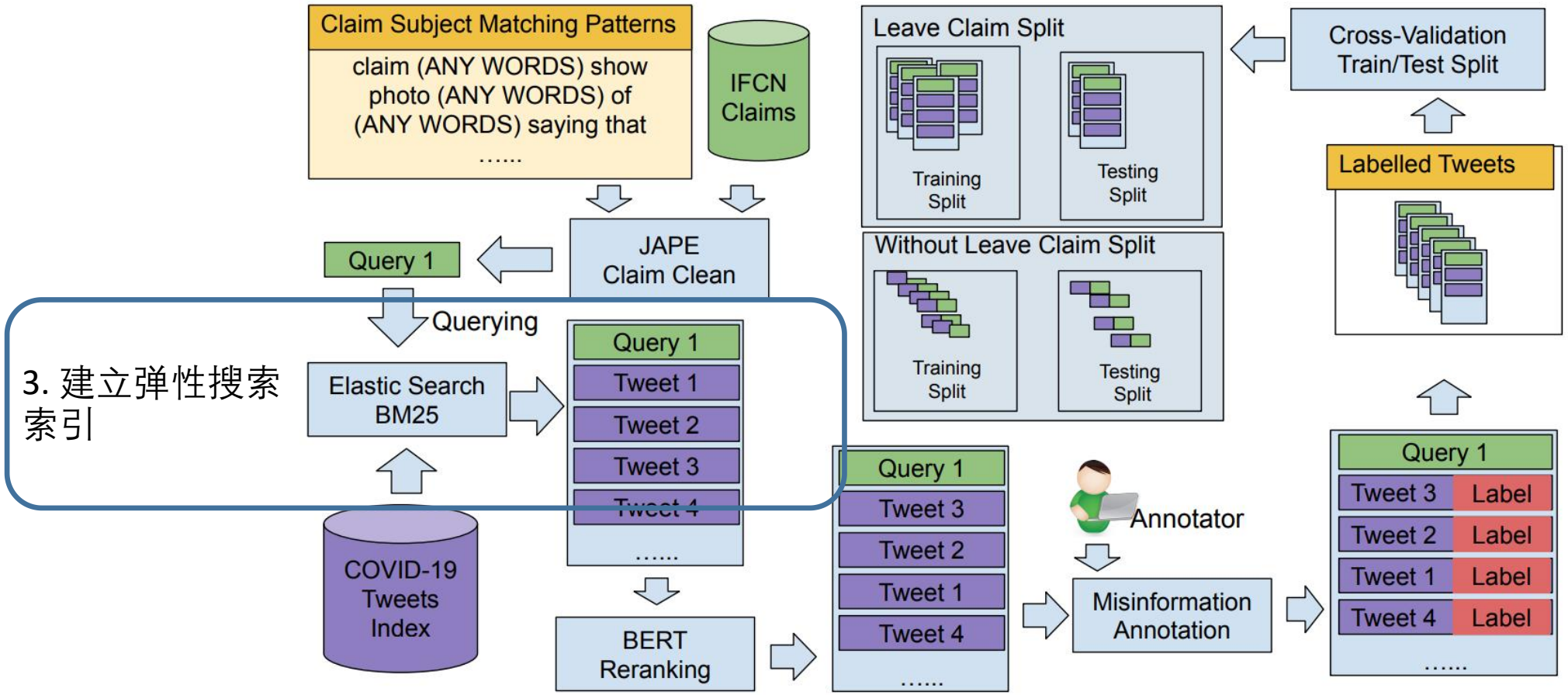
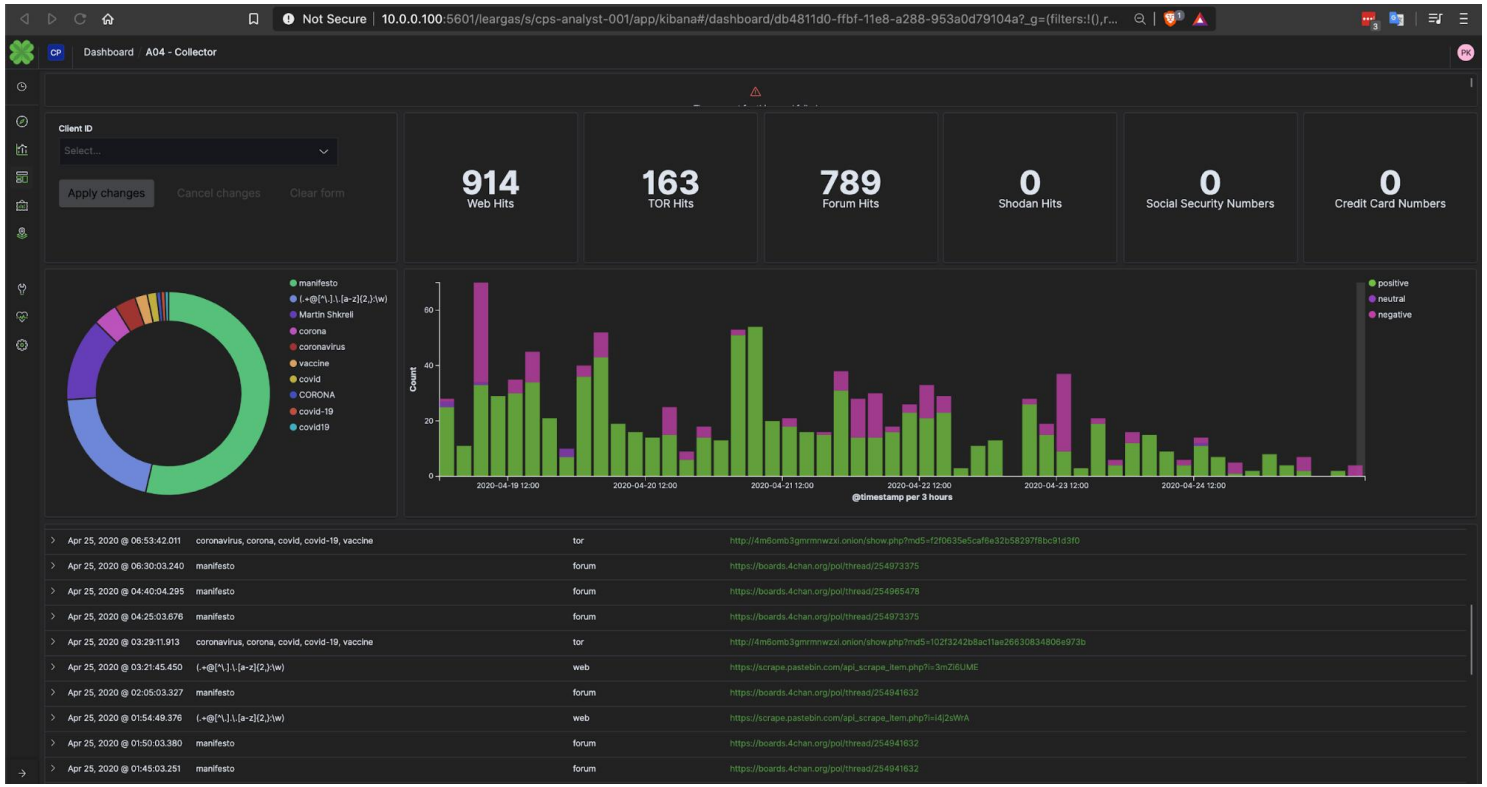


Figure 1. Overall pipeline

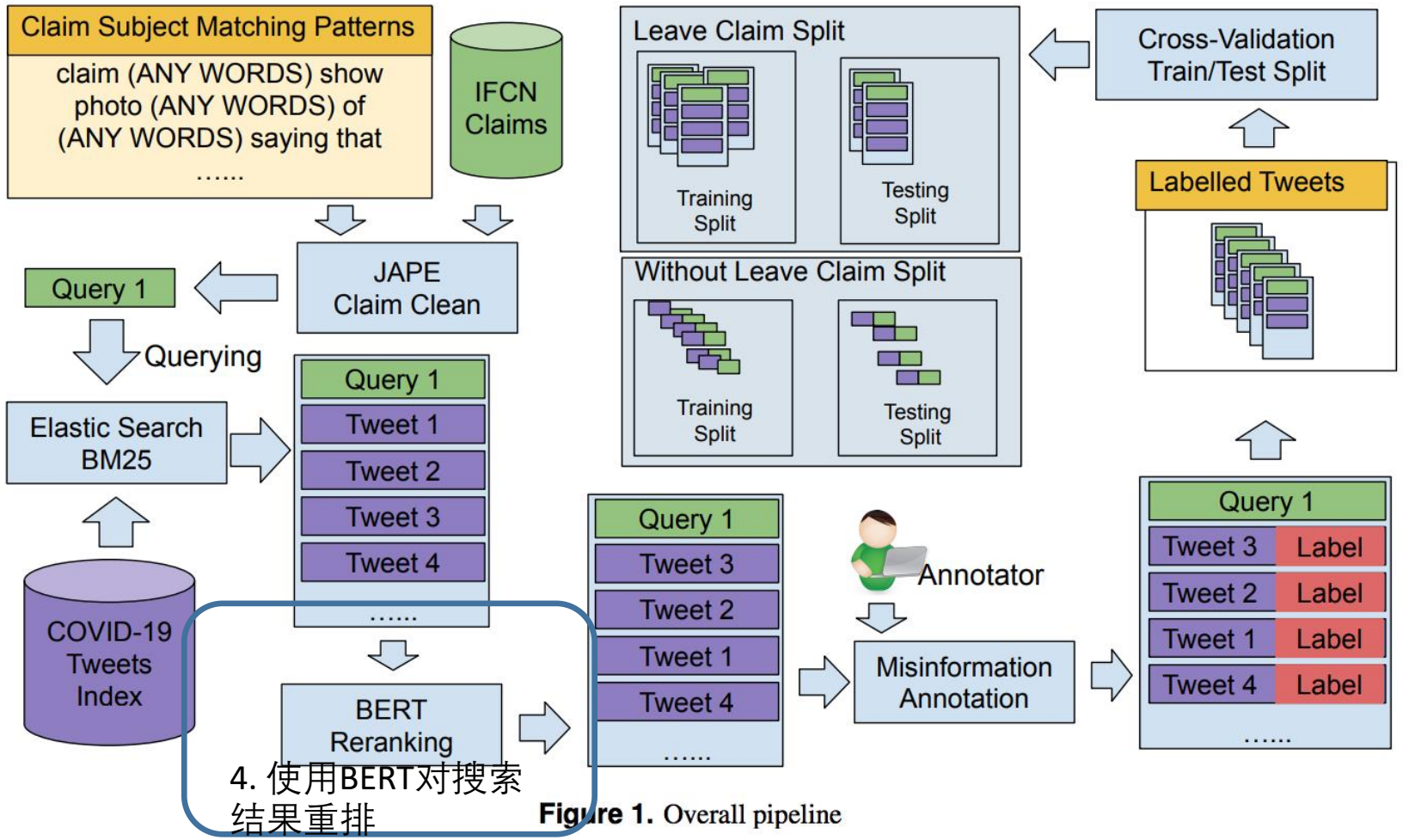
STEP 3. Elastic search（弹性搜索）索引

1. 建立弹性搜索索引，实时监控数据变化趋势；



2. 使用IFCN数据作为搜索条件，使用tfidf权重+BM25算法对twitter数据进行初步搜索；

虚假信息检测总体流程



4. 使用BERT对搜索结果重排

Figure 1. Overall pipeline

虚假信息检测总体流程

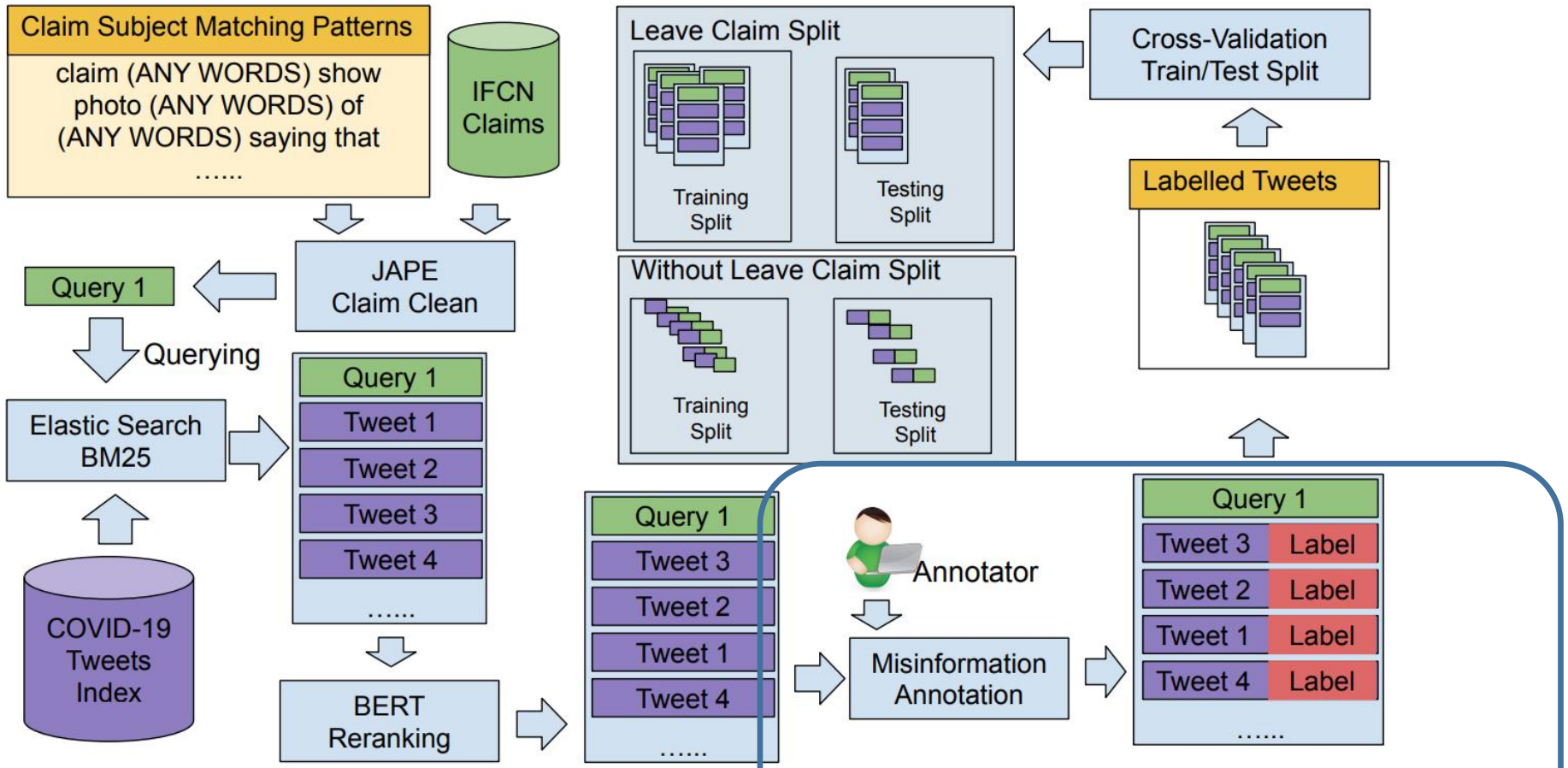


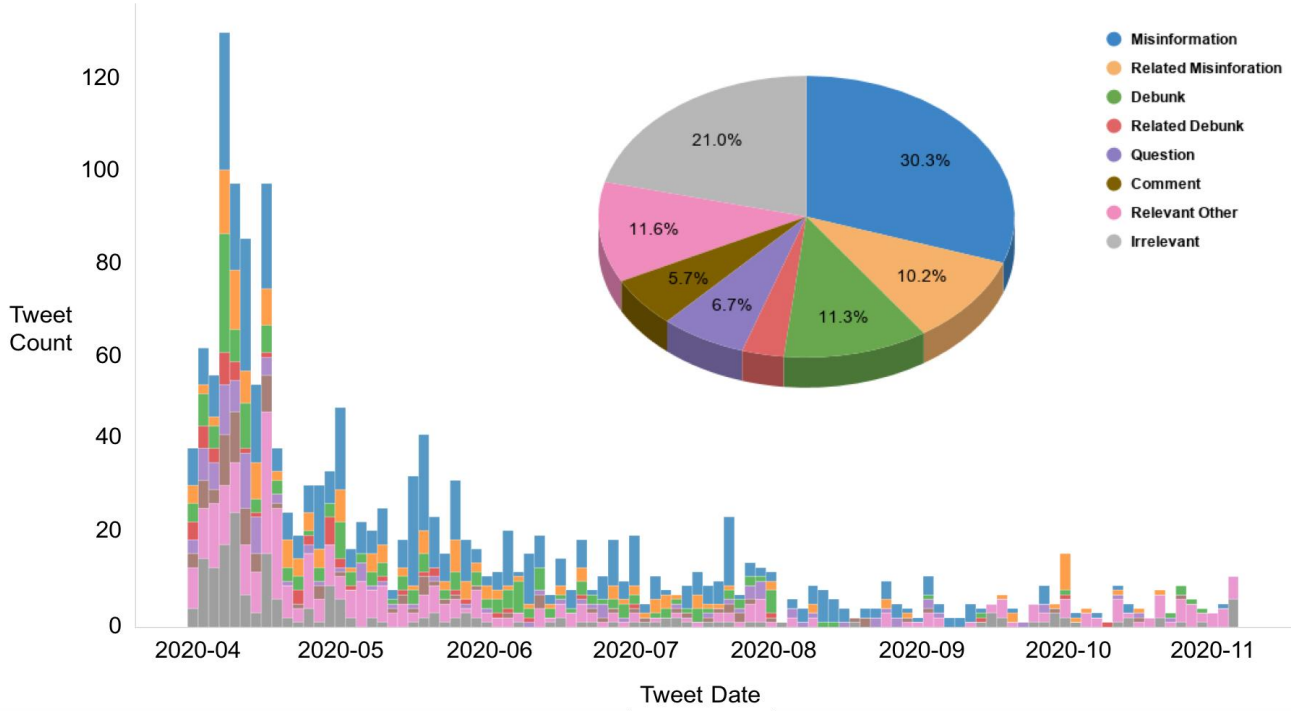
Figure 1. Overall pipeline

5. 对语义搜索结果进行人工标注

Step 5. 人工标注

Twitter数据类别
Misinfomation (虚假信息)
Related Misinformation （相关虚假信息）
Debunk （揭露）
Related Debunk （相关揭露）
Question （提问）
Comment （评论）
Irrelevant （无关信息）
Other （其它）

1. 标注人员：9位专业标注志愿者；
2. 标注时长：3周；
3. 验证基准：Krippendorff’s Alpha (三重验证)+majority vote
4. 校正得分：0.67 （substantial agreement ）



虚假信息检测总体流程

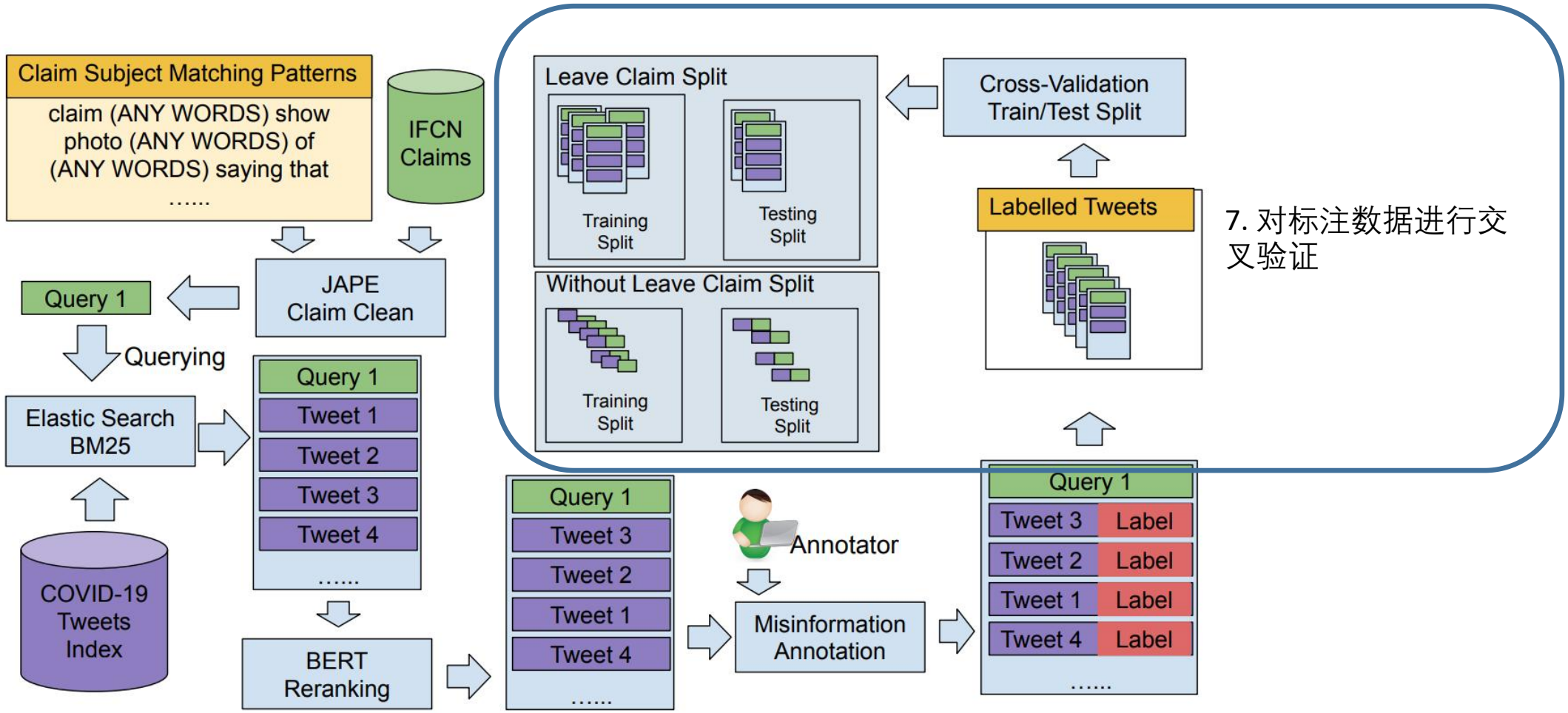
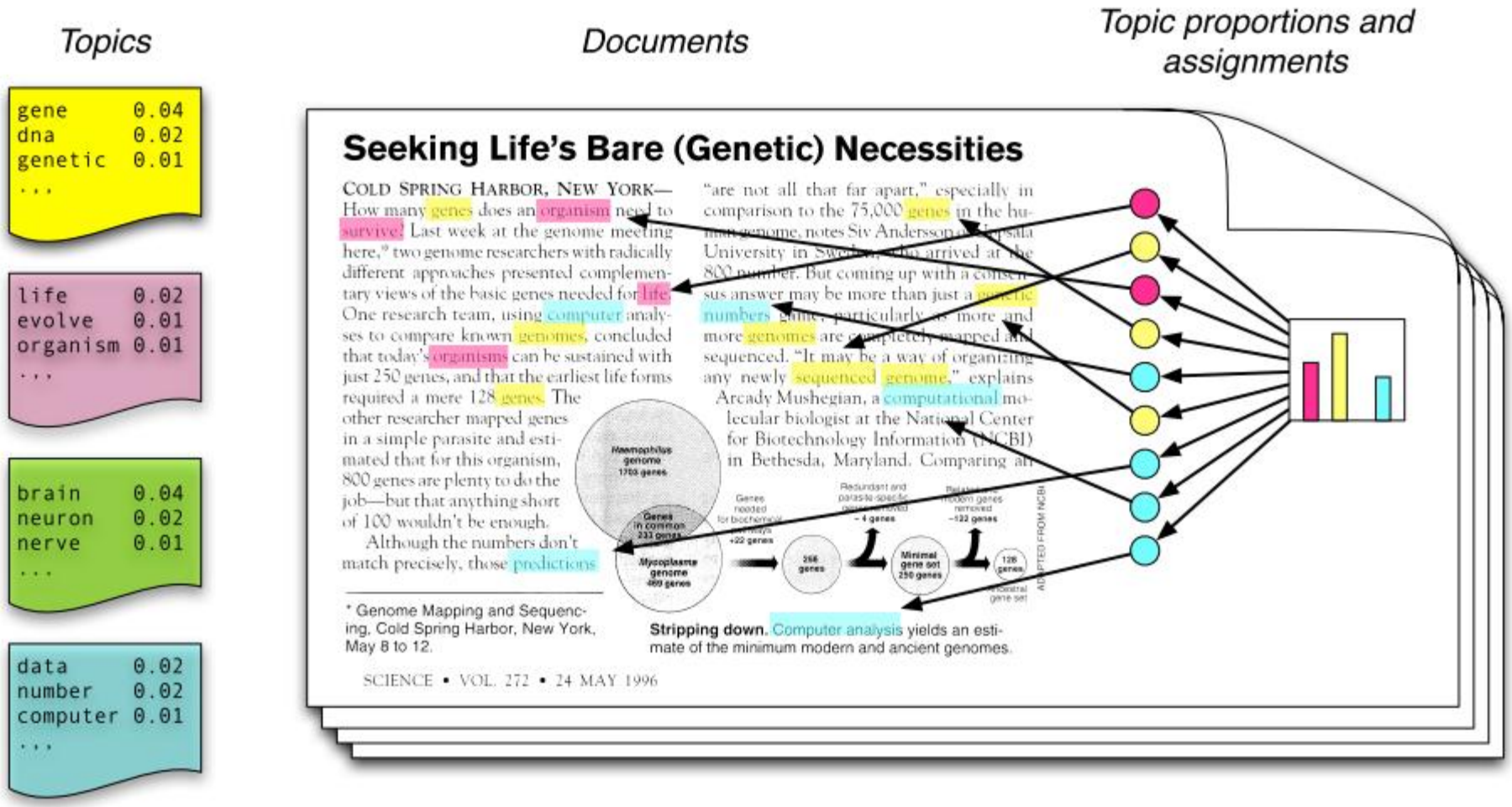


Figure 1. Overall pipeline

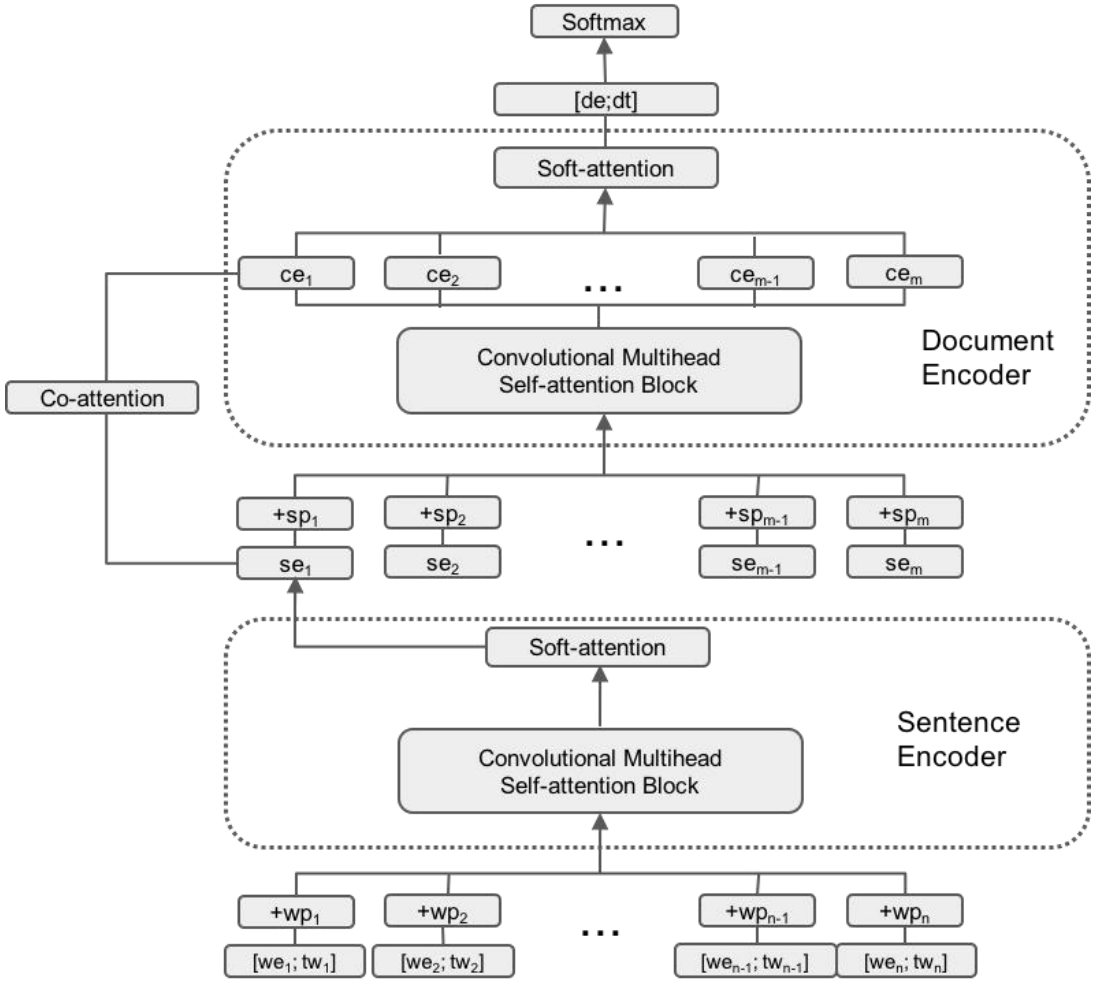
Step 7. 交叉验证

- 基于5-fold cross-validation 的训练集与测试集;
- 两种cv分集方式:
  - Without leave claim out：将tweet数据与ifcn数据相结合，随机分配至训练集与测试集;
  - Leave claim out：将tweet数据与ifcn数据相结合，按照ifcn的主题将数据分配至训练集或测试集。

基于主题模型分布与多种注意力机制的结构化模型



基于主题模型分布与多种注意力机制的结构化模型



注：

1. we代表词向量；
2. tw代表主题-词分布；
3. wp代表词位置向量；
4. se代表句子表征；
5. sp代表句子位置向量；
6. ce代表自注意力机制后的句子表征；
7. de代表文本表征；
8. dt代表文本-主题分布；

基于主题模型分布与多种注意力机制的结构化模型

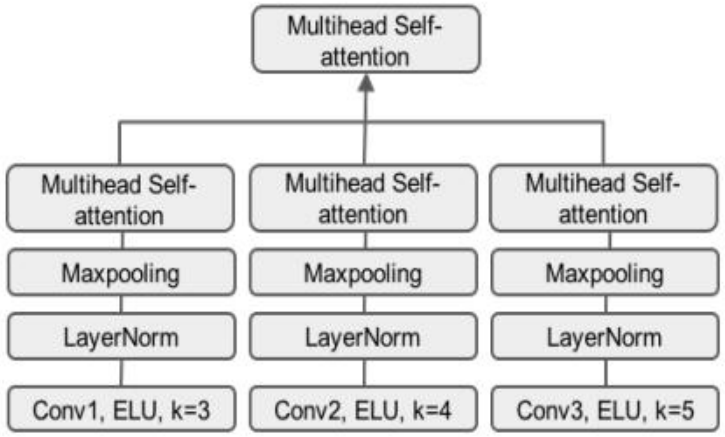


Figure 4: Convolutional multihead self-attention block

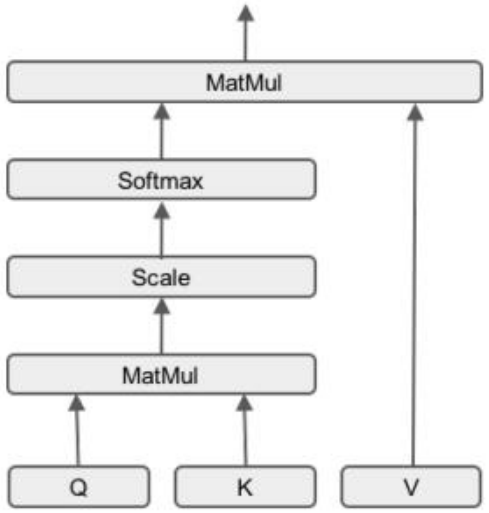


Figure 2: Scaled dot product attention

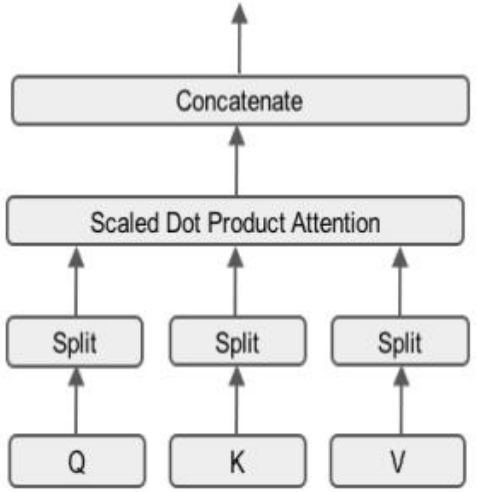


Figure 3: Multihead self-attention

$$c^k = ELU(Conv1D(EW^k + b^k))$$
$$Q, K, V = Maxpooling(LayerNorm(c^k))$$

$$Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d_k}})V$$

$$Multihead(Q, K, V) = [head_1, ..., head_h]$$
$$where\ head_i = Attention(Q_i, K_i, V_i), i \in [1, h]$$

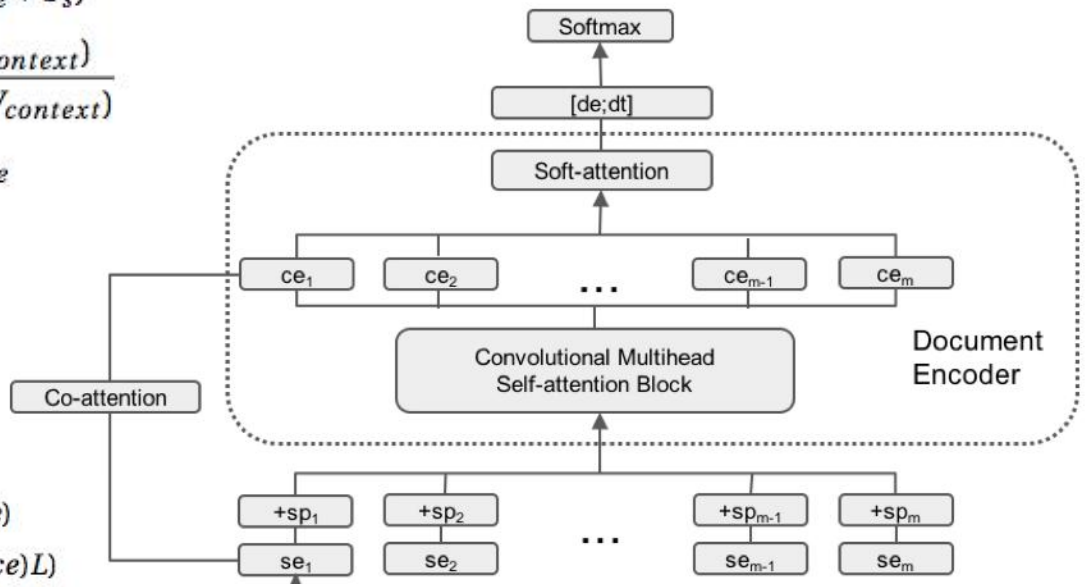
基于主题模型分布与多种注意力机制的结构化模型

1. 软性注意力机制（soft-attention）：

$$u = tanh(W_s c_e + b_s)$$

$$\alpha = \frac{exp(u^T W_{context})}{\sum_i exp(u^T W_{context})}$$

$$f = \sum_i \alpha c_e$$



2. 协同注意力机制（co-attention)：

$$L = tanh(ce^T W_l se)$$

$$H^s = tanh(W_s se + (W_d ce)L)$$

$$H^d = tanh(W_d ce + (W_s se)L^T)$$

$$\alpha^s = Softmax(W_{hs}^T H^s)$$

$$\alpha^d = Softmax(W_{hd}^T H^d)$$

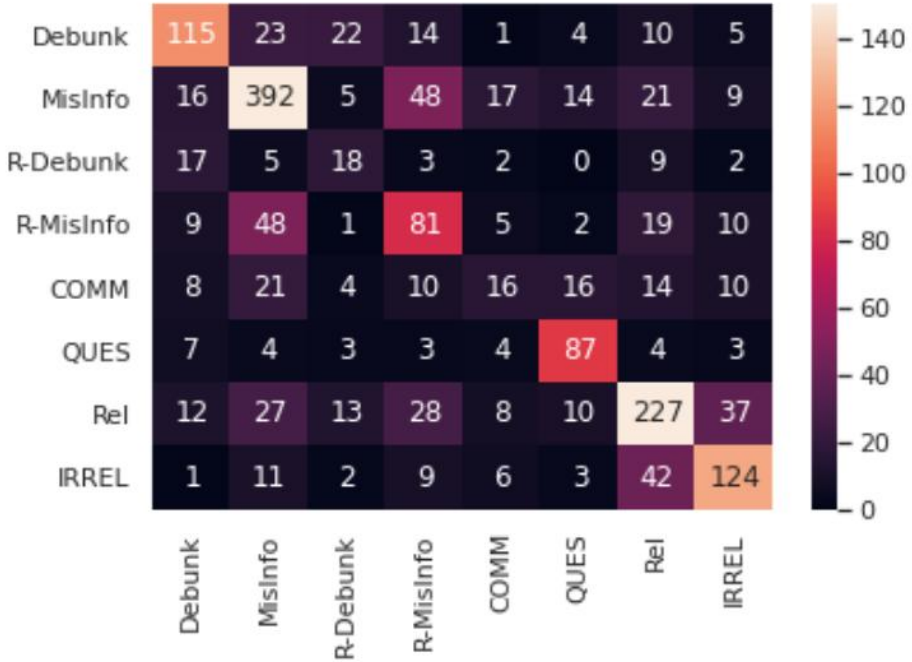
$$\bar{s} = \sum_m \alpha^s se^m$$

$$\bar{c} = \sum_m \alpha^d ce^m$$

实验结果

	Without Leave Claim					Leave Claim				
	BERT_CLS	CANTM	SBERT	BERT_Pair	T-HMAN	BERT_CLS	CANTM	SBERT	BERT_Pair	T-HMAN
Accuracy	0.584	0.621	0.639	0.615	0.667	0.310	0.349	0.353	0.370	0.427
F1	0.515	0.524	0.555	0.524	0.601	0.271	0.277	0.259	0.276	0.365
Debunk F1	0.622	0.638	0.630	0.602	0.624	0.333	0.312	0.361	0.382	0.410
MisInfo F1	0.671	0.736	0.757	0.742	0.777	0.373	0.476	0.535	0.495	0.434
R-Debunk F1	0.293	0.264	0.409	0.258	0.138	0.025	0.0	0.071	0.038	0.104
R-MisInfo F1	0.416	0.439	0.478	0.434	0.357	0.135	0.085	0.069	0.131	0.289
COMM F1	0.239	0.224	0.159	0.209	0.255	0.110	0.221	0.143	0.149	0.132
QUES F1	0.715	0.695	0.719	0.697	0.757	0.613	0.623	0.451	0.578	0.638
REL F1	0.595	0.624	0.646	0.635	0.641	0.335	0.343	0.309	0.320	0.220
IRREL F1	0.573	0.572	0.643	0.613	0.601	0.248	0.158	0.131	0.116	0.203

实验结果



Claim 1	Steam from boiling oranges kills COVID-19.
Tweet Text	#Fact: No scientific evidence to prove that inhaling hot water steam kills #Coronavirus
Prediction: DEBUNK	Label: RELATED_DEBUNK

Claim 2	Research proves that commercial mouthwash could protect against COVID-19.
Tweet Text	Mouthwash could prevent COVID-19 transmission, scientists say https://… via @... @... This is a reckless headline. It should read, "Scientists theorize mouthwash may prevent COVID-19, more research needed." #scicomm #covid19 cc @...
Prediction: MISINFORMATION	Label: COMMENT

数据分析

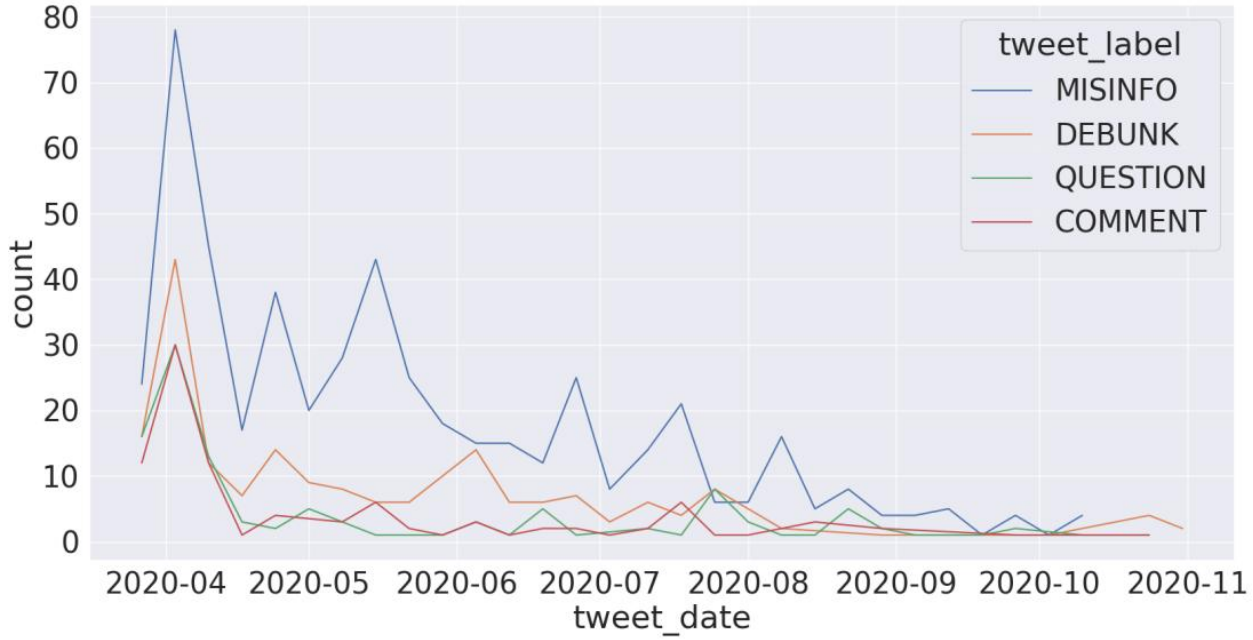


Figure 2. Misinformation, debunk, question and comment tweets volume over time (in weeks).

1. 新冠疫情爆发与虚假信息传播以及揭露信息成正相关(Pearson correlation ρ = 0.55, p < 0.001);
2. 在第一次疫情爆发周期中更踊跃的提出疑问（Pearson correlation ρ = 0.58, p < 0.001）和表达自己观点（ Pearson correlation ρ = 0.45, p < 0.001 ）

数据分析

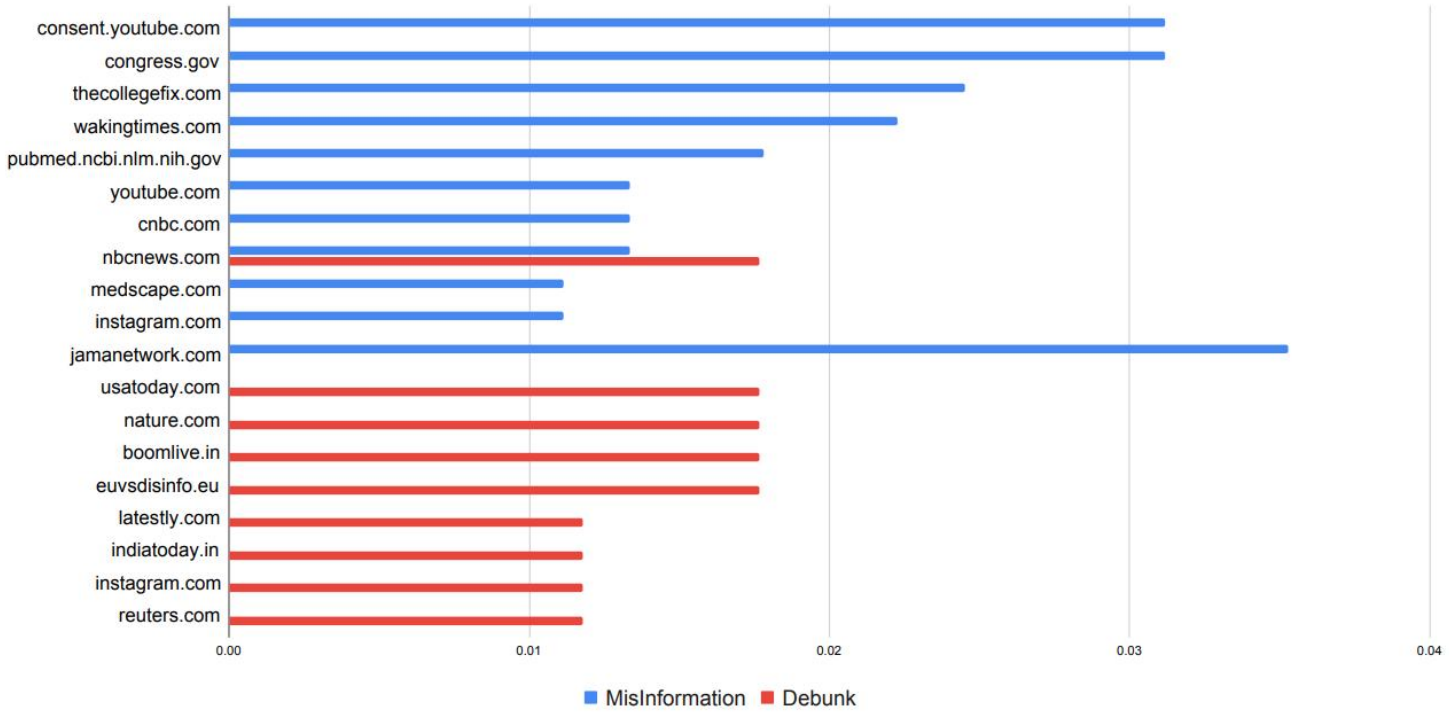
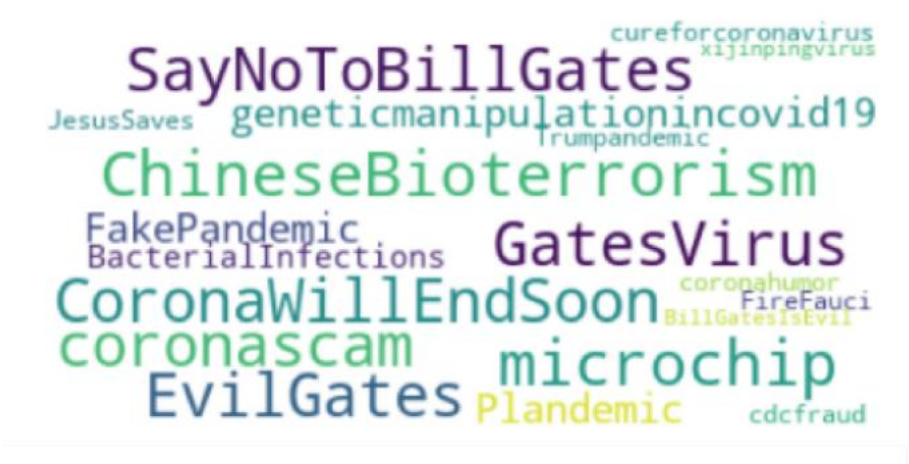


Figure 3. Top 10 frequent URLs found in misinformation and debunk tweets.

1. 虚假信息较比揭露信息更经常引用外部链接;
2. 虚假信息经常引用权威网站链接（xx.gov）;
3. 外部网站链接不能直接作为判断该信息是否是虚假信息或揭露信息的唯一标准

数据分析



Misinformation Hashtags

Debunk Hashtags

公共数据集实验分析

Models	IMDB	Yelp 2015	Amazon Reviews	Hyperpartisan News	AG News
Shallow-and-wide CNN	88.32	60.71	89.96	75.47	89.33
Attentive-RNN	87.89	60.15	90.21	76.01	88.78
HAN	91.09	63.81	92.45	76.56	92.48
HAHNN	90.24	64.71	92.68	77.00	92.57
HCAN	90.59	64.89	93.13	77.17	93.89
T-HMAN	**91.36**	**66.07**	**94.20**	**78.51**	**94.77**

基于测试集准确率（%）

公共数据集实验分析

A note to Steve Bannon : No matter what your beliefs are , the most dangerous thing you can do is cross Donald Trump . The fallout against the former White House political strategist and Trump confidant is continuing this week , in the wake of his comments appearing in Michael Wolff 's tell - all book , " Fire and Fury , " and as his political foes are celebrating his perceived demise and his political friends are starting to distance themselves from him . One of Bannon 's strongest allies ,Äî the Mercer family , which is funding Breitbart ,Äî issued a stinging rebuke on Thursday , but limited the scope of their attack to what Bannon said about the president . " I support President Trump and the platform upon which he was elected , " Rebekah Mercer told the Washington Post . " My family and I have not communicated with Steve Bannon in many months and have provided no financial support to his political agenda , nor do we support his recent actions and statements . " And CBS reported Friday that Mercer " cut all ties " with Bannon because that 's what Trump wanted her to do . . @CBSNews confirms Steve Bannon 's longtime benefactor Rebekah Mercer cut * all ties * with him at the request of White House officials in the aftermath of # FireAndFury . pic.twitter.com/8eICjpp6Qw ,Äî ryan kadro (@RyanKadro) January 5 , 2018 But Mercer is continuing to finance Breitbart , which has not only been Bannon 's political agenda , it 's been his political essence . Bannon made Roy Moore his candidate in the Alabama special election . When Moore 's major flaws were exposed , Bannon dispatched Breitbart 's staffers to attack the women who accused Moore of trying to date them when they were teenagers . After the election , Breitbart 's editor admitted that they were acting as political operatives in their decision to cover the election . So while Breitbart 's funders are wondering whether or not to cut ties with Steve Bannon , it 's important to realize that Bannon 's legacy will certainly live on . It will just be more Trump - friendly .

Run

note to steve bannon no matter what your beliefs are most dangerous thing you can do is cross donald trump .
fallout against former white house political strategist and trump confidant is continuing this week in wake of his comments appearing in michael wolff 's tell all book fire and fury and as his political foes are celebrating his perceived demise .
one of bannon 's strongest allies , äî mercer family which is funding breitbart , äî issued stinging rebuke on thursday but limited scope of their attack to .
i support president trump and platform upon which he was elected rebekah mercer told washington post .
my family and i have not communicated with steve bannon in many months and have provided no financial support to his political agenda nor do we support his recent actions and statements .
and cbs reported friday that mercer cut all ties with bannon because that 's what trump wanted her to .
.
@cbsnews confirms steve bannon 's longtime benefactor rebekah mercer cut all ties with him at request of white house officials in aftermath .

This document is Tendentious

注意力机制可视化分析（超党派新闻）

谢谢!