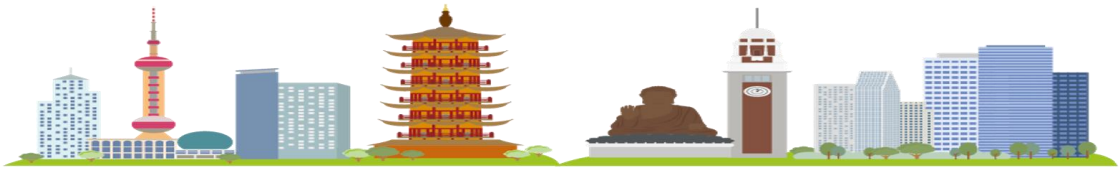
2022 Shanghai-HK Interdisciplinary Shared Tasks



The shared tasks on social media analysis is jointly organized by **the Hong Kong Polytechnic University**, **the Chinese University of Hong Kong**, **Fudan University**, and **Tongji University**. Datasets are collected and annotated by **SMART Group** (PolyU) and **DISC Lab** (Fudan) to explore the NLP application in interdisciplinary domains such as **rumor verification**, **poll generation**, **emotion analysis**, **user profile modeling** and **code-switched text analysis**. Besides, the **WRD Big Data Institute** will provide large-scale weibo corpus to help the training of general pretrained language models.

The aim of this competition is to promote the application of interdisciplinary social media analysis and provide communication platform for researchers and practitioners of relative fields. Welcome to register for the shared tasks if you are interested in social media processing!

The registration website is available now: http://fudan-disc.com/sharedtask/social22/

1. Introduction of Tasks and Datasets

Task 1: Trigger Identification in Rumor Verification

Nowadays, rumors tend to spread quickly and widely on the Internet, and automatically verifying rumors has become an urgent need for individuals and society. From social platforms (such as Twitter and Weibo), we can crawl information cascades that consist of source posts and corresponding reposts.

The task of rumor verification aims at classifying rumor cascade as true, false or unverified. However, predicting at cascade level is too coarse to organize massive messages and opinions. Therefore, Fudan DISC Lab supplement message-level annotations to exiting rumor corpus PHEME and propose a sub-task named trigger identification, which aims at identifying messages that have prominent effects on rumor proliferation and dominate the judgment of cascade credibility. We summarize message roles into 4 categories, i.e. amplify, deny, clarify and null. Amplify indicates tweets that initiate new concerns or enlarge the discussion scale related to the social event. Deny means presenting doubt or rejection towards previous messages. Clarify introduces factual or substantial information. Other messages are left as null which means they are insignificant for rumor propagation or verification.

The extended **dataset** contains 1,929 cascades and 26,871 messages annotated. Train, validation, test (phase 1) and test (phase 2) sets are split randomly with a proportion of 7:1:1:1.

The **baseline model** applies bert-base-uncased pretrained model to obtain message representation and employs mean pooling to acquire cascade representation. We adopt macro F1 score as **evaluation metrics**.

Task 2: Poll Question Generation

This task aims to generate poll questions for social media posts. As we know, social media is a crucial outlet for people to exchange ideas, share viewpoints, and keep connected with the world.

It allows us to hear the public voice for decision-making and better understand our society. Nevertheless, for the silent majority, they tend to read others' messages instead of voicing their own opinions with words, possibly because of the introverted personality, busy schedule, and others.

To automatically generate a poll question for a social media post will encourage public users, especially those reluctant to comment with words, to input their reflections via voting and thus have a better engagement.

The **dataset** contains the source post (src), context (conv), generation targets (trg) and candidate choice. Textual content has already been processed by jieba.

Evaluation metrics: Rouge and Bleu.

Task 3: Sentiment Analysis for Code-Switched Text

In social media, the text is becoming increasingly important due to its effectiveness in disseminating information in highly individualized and opinionated contexts. Affective analysis has been studied using different Natural Language Processing (NLP) methods from various linguistic perspectives such as semantic, syntactic, and cognitive properties. Social media text is often written in a code-switching style in certain parts of the world, such as Mainland China and Hong Kong.

This task aims to identify the users' sentiment from their comments. In this dataset, the instances are mainly written in Cantonese. Other languages are used, including English, French, Japanese, etc. The label classes are from 1-star to 5-star, forming a multi-class task.

Each record in the datasets is presented in the form of "TEXT\tLABEL\n". The evaluation metrics include accuracy and macro F1 score.

Task 4: User Profile Modeling

User profile modeling tasks aim to predict user attributes, such as gender, age, education, region, occupation, income, etc., which can be used to help characterize potential users in various applications such as e-commerce and social networks. This task predicts the user's emotional state and recent work in the past month by analyzing the historical texts posted by the user. Using the method of user modeling, we can track the user attributes in the process of event propagation, helping us to grasp the propagation law of social hot events.

Each record in the dataset represents a user, containing the id, historical weibo content, and the emotion label in the past month.

The baseline model utilizes keyword extract techniques and pre-trained word vectors to implement classification. Macro F1 score is adopted as evaluation metrics.

Task 5: Social Emotions to Online Topics

This task aims to predict social emotion to online discussion topics. While most prior tasks focus on emotions from writers, we investigate readers' responses and explore the public feelings about an online topic.

A large-scale dataset is collected from the Chinese microblog Sina Weibo with thousand of trending topics, emotion votes in 24 fine-grained types from massive participants, and user comments to allow context understanding.

Each record in the dataset contains the hashtag, relative comments and expected top 3 attitudes.

The baseline model utilizes keyword extraction techiniques and SGM to implement classification. The evaluation metrics is macro and micro F1 score.

2. Schedule

7 Mar, 2022: Open for registration

10 March - 25 May, 2022: Upload prediction results (phase 1). The website will automatically compute and record the evaluation metrics. The format of prediction results should follow the requirements in the task page. The maximum number of uploads per day is 10.

26 May - 31 May, 2022: Upload trained models & running instructions (phase 2). The organizers will run the model and obtain the evaluation metrics as the basis for final ranking. Each team has 3 opportunities to run models successfully.

1 Jun - 4 Jun, 2022: Submit reports for winning teams

6 Jun - 8 Jun, 2022: Online symposium

3. Awards

The **top three** teams for each task (15 in total) will be awarded prizes.

The champion: **HK$ 5,000**

The first runner-up: **HK$ 3,000**

The second runner-up: **HK$ 2,000**

***** Awards will be delivered in the form of consumption coupons. Winning teams will also have electronic certificates and be invited to attend the workshop.

4. Organization

Organizer

The Hong Kong Polytechnic University (PolyU)

The Chinese University of Hong Kong (CUHK)

Fudan University

Tongji University

General Chair

Kam-Fai Wong (CUHK)

Program Chair

Jing Li (PolyU)

Zhongyu Wei (Fudan)

Haofen Wang (Tongji)

Organization Chair

Lei Chen (Fudan)

Organization Comittee

Zexin Lu (PolyU)

Rong Xiang (PolyU)

Kun Wu (Fudan)

Keyang Ding (HIT)

Advisory Comittee

Qing Li (PolyU)

Wenjie Li (PolyU)

Xuanjing Huang (Fudan)

Baohua Zhou (Fudan)

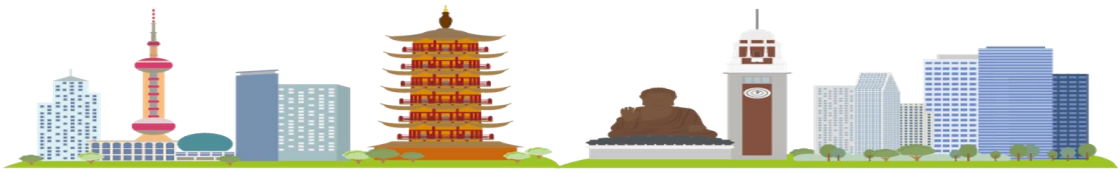
Yidong Liu (WRD)

Organization Support

Shanghai-Hong Kong University Alliance

University Grants Committee

2022 “上海-香港跨學科社交媒體分析聯合評測”開啟報名



本次社會媒體分析聯合評測由**香港理工大學**、**香港中文大學**、**復旦大學**、**同濟大學**共同承辦，由香港理工大學 SMART **研究組**和復旦大學 DISC **實驗室**提供標注數據集，通過收集和標注社交媒體上豐富的數據，來探索自然語言處理技術在**謠言檢測**（rumor verification）、**投票生成**（poll question generation）、**情緒識別**（emotion analysis）、**用戶畫像**（user profile modeling）以及**多語言語境下的情感分析**（sentiment analysis for code-switched text）5 項任務上的應用。此外，**新浪微熱點大數據研究院**將提供大規模數據助力**通用社交文本理解模型**的訓練。此次評測旨在推廣跨學科的社會媒體分析應用，為相關領域的研究者和從業人員提供良好的溝通平臺，歡迎對社交媒體分析處理感興趣的個人和團體積極報名參賽。

報名通道已於 3 月 7 日開啟，各任務數據及基線模型可通過賽事網站獲取：http://fudan-disc.com/sharedtask/social22/

1. 任務與數據介紹

任務一：謠言檢測中的資訊引爆點識別（Trigger Identification in Rumor Verification）

對社交網路中消息的真實性進行檢驗，可以借助信源的評論轉發作為輔助特徵。根據消息的轉發消息可形成資訊傳播樹，如何組織和利用相關文本特徵是目前的一大難點。復旦大學數據智慧與社會計算實驗室基於現有的謠言檢測語料集 PHEME，完成了消息級別的標注，以期在消息級別上對消息的引爆點類別進行預測，來輔助後續的謠言檢測。

任務一的**數據**是 csv 格式，每行包括每條消息的“所屬傳播樹編號（cid）、消息編號（mid）、轉發父節點編號（pid）、消息內容（content）、發佈時間（time）、消息引爆點標籤（trigger）、所屬傳播樹標籤（verify）”，共涉及 1929 個傳播樹以及 26871 條消息，訓練、驗證、測試 1、測試 2 集合通過隨機方式劃分，比例為 7:1:1:1。其中 trigger 標籤的數字對應類別分別為 0-無實義（null），1-開啟話題（amplify），2-反對質疑（deny），3-澄清說明（clarify）；verify 標籤的數字對應類別分別為 0-虛假消息（false），1-真實消息（true），2-存疑消息（unverified）。

所採用的**基線模型**是基於 bert-base-uncased 進行消息表示，使用平均池化進行傳播樹表示的分類模型，所採用的**評價指標**為兩個分類任務上宏平均 F1 值（macro F1），排名依據為兩個任務上評價指標的加和。

任務二：投票問題生成（Poll Question Generation）

本任務為社交媒體的推文（比如微博的內容）生成問題，從而啟發人們的思考和參與討論。社交媒

體為人們提供了交換思想和獲取資訊的管道，從而也使得用戶連接在一起，這有利於社會群體針對社會話題進行意見表達。然而，有相當大一部分人，傾向於閱讀資訊，比如刷微博，刷微信朋友圈，而不是參與互動，比如點贊，評論等行為。這可能是由於內向的性格或者生活習慣等原因造成的。該任務希望提出一些問題來幫助人們更好地參與到社交媒體當中去，如參與社會話題的討論。

數據檔案格式為 txt，每行為一個樣本。Train，valid，test 分別有 src，trg，conv，choice，對應帖子，問題，評論，選項答案。比如 train_scr.txt 第一個行是一個 post，該樣本對應的 question 在 train_trg 的第一行，對應的投票答案在 train_choice 的第一行，評論 conv 同理。valid 和 test 的格式和 tain 同理。數據已經用 jieba 做分詞處理。

評價指標：Rouge 和 Bleu.

任務三：多語言語境下的情感分析（Sentiment Analysis for Code-Switched Text）

在社交媒體中，由於網路用語的高度個性化和迅速傳播的特點，文本分析變得越來越重要。情感分析作為社交媒體文本分析的一個典型應用，已經在多種多樣的自然語言處理（NLP）領域中，從各種語言學角度（如語義，句法和認知屬性）展開研究。由於全球化的發展，在某些地區，例如中國大陸和香港，社交媒體文本通常伴隨著大量的的語言轉換，例如“今天我很 happy!”。

該任務旨在從伴隨著語言轉換的用戶評論中識別用戶的情緒。在此數據集中，評論主要語言為粵語。其他語言，包括英語，法語，日語等的使用，給情緒識別帶來了新的挑戰。

數據格式：訓練、驗證、測試數據集以 tsv 檔的形式提供。標注等級從 1 星到 5 星，故該任務為多標籤分類任務。在數據集中，每行提供了文本和相應的標籤（測試集除外），格式為 "TEXT\tLABEL\n"。**基線模型代碼**已提供，實現方式為中文 transformer (chinese-bert-base) 的 fine-tune。所採用的**評價指標**為 accuracy 與 macro F1 score。

任務四：用戶畫像建模（User Profile Modeling）

用戶畫像建模任務旨在預測用戶的屬性，如性別、年齡、教育程度、地區、職業、收入等，這些屬性可以用來說明描繪電子商務、社交網路等各種應用中的潛在用戶的特徵。準確刻畫用戶的身份、行為特徵，有助於在這些領域中更好地為用戶提供相關服務。復旦大學數據智慧與社會計算實驗室通過問卷調查的方式，基於微博文本構建了用戶畫像數據集，期望利用用戶的歷史微博資訊來推斷用戶近一個月的情感狀態及工作類別。

任務四的**數據集**是 csv 格式，每行表示一個用戶，其中包括用戶的問卷號(id)，用戶的歷史微博(weibo)，用戶填寫問卷的時間(paper_time)，用戶近一個月的情感狀態(emotion_label)，用戶的工作標籤(job_label)。訓練、驗證、測試 1 共 5177 個用戶，通過隨機方式劃分，比例為 6: 2: 2。

所採用的機器學習**基線模型**通過對用戶近幾個月發的微博進行關鍵字提取，再用公開的預訓練微博詞向量 public_opinion_word2vec 得到表示，使用 SVM / Logistic Regression 進行分類。所採用的深度學習模型將用戶近幾個月發的微博文本進行拼接，基於 bert-base-chinese 得到表示，最後通過線性層進行分類。所採用的**評價指標**為兩個分類任務上的宏平均 F1 值(macro F1)，排名依據為兩個任務上評價指標的加和。

任務五：面向話題的群體情緒識別（Social Emotions to Online Topics）

社交媒體已經是人們日常生活中不可分割的一部分，越來越多的互聯網用戶參與到社交媒體之中進行討論、分享自己的觀點和看法。這為研究者們提供了一個龐大的可以獲取流行話題、傾聽公眾針對熱點事件意見的資源。但是飛速發展的社交媒體所蘊含的資訊已經超越了人所能獲取資訊的極限，我們迫切的需要一些手段輔助我們進行情感分析。之前雖然有這方的研究，但是大多局限于作者的情感或類似新聞報導、標題等比較正式的文字，較少有工作涉及到公眾對於這種零散且口語化的線上話題的情感。因此針對集體所對話題所產生的社會情感，我們推出了此任務，其目標是通過分析來自社交媒體話題裡的簡短、口語化的文本，來分析預測公眾的情緒。

任務五的**數據格式**為 json 文檔，其中"hashtag"為話題的名稱，"comments"為該線上討論話題下的評論(最多 100 條)，"attitudes"為按照投票結果得到的 Top3 表情（已按投票數量排好）, 按照隨機方式進行劃分。

所採用的的**基線模型**是將話題名稱與基於 jieba 分詞包的 textrank 方法提取的關鍵字作為輸入的 SGM 模型，所採用的評價指標為多標籤分類任務上的宏平均 F1 值與微平均 F1 值，排名依據為二者的加權和(分別為 0.6 與 0.4)。

2. 時間安排

2022 年 3 月 7 日 報名通道開啟及數據發佈（第一階段測試結束前仍可報名）

2022 年 3 月 10 日-5 月 25 日 測試階段一開啟，參賽人員需提交在 test1 數據集上的標籤結果檔，系統將自動計算評價指標並記錄，系統將自動計算並記錄評測指標並對比當前結果和最優結果，最優結果將即時顯示在排行榜中，每個團隊每日最多上傳 10 次。

2022 年 5 月 26 日-5 月 31 日 測試階段二開啟，參賽人員需提交訓練好的模型及生成測試結果的代碼說明，由主辦方在 test2 數據集上進行測試，所得的評價指標將作為最終排名依據，每個團隊最多可提交 3 個運行成功的模型，記錄最優結果。

2022 年 6 月 1 日-6 月 4 日 任務報告提交

2022 年 6 月 6 日-6 月 8 日 線上研討會，並由獲獎團隊分享比賽經驗和模型

3. 獎項設置

五個任務分開排名，分開頒獎，分別設置以下獎項：
一等獎（1 名，獎金 5000 港幣）
二等獎（1 名，獎金 3000 港幣）
三等獎（1 名，獎金 2000 港幣）
* 獎金將以消費券的形式發放，另外將為優勝團隊頒發電子版獲獎證書，並邀請至研討會進行分享

4. 賽事委員會

主辦單位

香港理工大學

香港中文大學

復旦大學

同濟大學

新浪微熱點大數據研究院

大會主席

黃錦輝（香港中文大學）

程式委員會主席

李菁（香港理工大學）

魏忠鈺（復旦大學）

王昊奮（同濟大學）

評測主席

陳蕾（復旦大學）

評測委員

盧澤鑫（香港理工大學）

向榮（香港理工大學）

吳焜（復旦大學）

丁可陽（哈爾濱工業大學-深圳）

顧問委員

李青（香港理工大學）

李文捷（香港理工大學）

黃萱菁（復旦大學）

周葆華（復旦大學）

劉益東（新浪微熱點大數據研究院）

特別鳴謝

滬港大學聯盟

香港特別行政區大學教育資助委員會