Psychological Analysis of Social Media Data: The Lincui Analysis System

Authors: Zhu Tingshao, Zhu Tingshao

Date: 2025-04-19T00:00:00+00:00

Abstract

With the development of technology, the era of big data has quietly arrived. The emergence of big data has brought tremendous convenience to scientific research, enabling researchers to enhance the efficiency of their work through the analysis of large-scale datasets. This paper introduces the LinCui Analysis System, which we have developed and implemented to assist researchers with minimal or no programming background in utilizing existing Python programs for data collection and analysis, requiring zero programming expertise. The LinCui Analysis System follows the conventional workflow of data collection and processing in research. First, it filters data meeting specified criteria from the collected dataset to form data groups; this filtering process can be multi-step. Subsequently, the filtered data is segmented into individual data units, which are then computationally processed to obtain various psychological semantics or psychological indicators of users. Users may employ data collected via web crawlers or self-collected data, and through filtering and segmentation, obtain individual behavior data. Here, "individual" refers not only to each user but also to data from a region or entity within a specified time period. Based on these data, dictionary-based psychological semantic analysis (word frequency statistics) and psychological indicator prediction can be performed. Upon these computational results, cross-sectional analysis or panel data analysis can be conducted according to research requirements. This paper demonstrates the entire process of data analysis using the LinCui Analysis System through a specific case study, illustrating that the system can provide valuable assistance to scientific research in data acquisition and analysis.

Full Text

Psychology Data Mining on Social Media: The Implementation of PsyAnalytics

Tingshao Zhu[1],[2*1] Institute of Psychology, Chinese Academy of Sciences, Beijing 100101, China [2] Department of Psychology, University of Chinese Academy of Sciences, Beijing 100049, China

Abstract

With technological advancement, the era of big data has arrived, bringing unprecedented opportunities to scientific research by enabling large-scale data analysis. This paper introduces PsyAnalytics, a psychological analysis system designed to assist researchers with limited or no programming background in completing data collection and analysis using existing Python programs. The system follows conventional research workflows for data acquisition and processing. First, it filters collected data to form datasets according to specified criteria—a process that can involve multiple iterative steps. Next, the filtered data is segmented into individual-level data, which is then processed to extract various psychological semantics and indicators. Users can employ web crawlers or self-collected data, filtering and segmenting it to obtain individual behavioral data. Here, "individual" refers not only to single users but also to aggregated data for a region or entity within a specified time period. Based on these data, dictionary-based psycholinguistic analysis (word frequency statistics) and psychological indicator prediction can be performed. These computational results can then support cross-sectional or panel data analyses according to research needs. Through a concrete case study, this paper demonstrates the complete data analysis process using PsyAnalytics, illustrating how the system facilitates scientific research in data acquisition and analysis.

Keywords: Internet big data; psycholinguistic analysis; psychological indicator prediction; PsyAnalytics

1 Introduction

The exponential growth of data generation has ushered society into the big data era, profoundly impacting learning, work, daily life, and social structures. This data explosion also provides researchers with unprecedented opportunities for exploratory analysis. As one of the primary carriers of massive datasets, the Internet contains vast amounts of underutilized data that urgently require mining and application. However, effectively harnessing these data remains a significant challenge for many non-specialists.

PsyAnalytics is a comprehensive psychological analysis system encompassing the entire pipeline from data acquisition and cleaning to analysis. It was specifically designed to enable researchers without computer science backgrounds to obtain research data from the Internet without writing code. The system assists faculty

and students with minimal or no programming experience in completing data collection and analysis using existing Python programs. Users need only edit plain-text configuration files and know how to run Python programs, making it accessible even to those with zero programming foundation.

2 Processing Flow and System Architecture

PsyAnalytics emulates conventional research workflows for data collection and processing. Consider a multi-item questionnaire survey of Weibo users in Beijing and Shanghai: researchers first filter and select eligible participants, then administer questionnaires to each subject. Users complete the questionnaires through self-reporting, and scores are calculated to obtain multiple psychological indicators. This conventional workflow is illustrated in [Figure 1: see original paper].

When using Weibo data to measure user characteristics, the technical workflow is shown in [Figure 2: see original paper]. The process begins by filtering collected data to form datasets according to specified criteria—a filtering process that may involve multiple steps. Next, the filtered data is segmented into individual-level data, which is then processed to obtain various psychological semantics or indicators for each user.

The overall system architecture is depicted in [Figure 3: see original paper]. PsyAnalytics supports multiple data acquisition methods, including crawlers or self-collected data, as long as the data conform to the required format. Through filtering and segmentation, the system generates individual behavioral data. Here, "individual" refers not only to each user but also to data aggregated for a region or entity within a specified time period. Based on these data, dictionary-based psycholinguistic analysis (word frequency statistics) and psychological indicator prediction can be performed. These computational results can then support cross-sectional or panel data analyses according to research requirements.

3 Data Acquisition and Processing

PsyAnalytics supports diverse data acquisition methods, including exported Weibo data, custom crawlers, or direct text input, provided the data conform to a unified format.

3.1 Data Format

The source data format consists of seven fields per line, separated by tab characters (`\t`). User identifiers, location, publication time, and content are typically required, while other fields can be customized as needed:

- **User or author identifier:** Can be a username or Weibo user ID
- **Location:** The author' s registered location or publication location, in either "Province/Municipality/Autonomous Region" format

or "Province/Municipality/Autonomous Region Prefecture-level City" format (space-separated). Examples: "Beijing" or "Beijing Chaoyang"
- **Content identifier:** Can be a Weibo ID or webpage URL
- **Publication time:** Must use Unix timestamps; other date-time formats require conversion
- **Month count:** Customizable field
- **Week count:** Customizable field
- **Text content:** Can be Weibo text or crawled webpage content. If saved as a text file, must begin with `FileRef::`, e.g., `FileRef::c:\psyanalytics\data\1.txt`

Individual data consist of multiple lines of plain text, which can be generated through segmentation or other user-defined methods.

PsyAnalytics currently provides three data acquisition methods: a custom crawler (`spider\ccpl_{spider}`), a specialized Weibo crawler (`spider\Weibo Crawler Program`), and Weibo2209 exported data. All acquired data conform to the source data format, which is also the format of filtered datasets. After segmentation, individual-level data are generated for subsequent psycholinguistic analysis and psychological indicator prediction.

3.2 Data Processing

PsyAnalytics data processing comprises three main steps: `rawFilter`, `rawSplitter`, and `getWordFreq`/`swls-predict`. Each step runs according to its configuration file (`.ini`) without requiring any code modification—users simply execute the corresponding Python program.

Users specify parameters in the `.ini` files according to their processing requirements, then run the appropriate `.py` program to complete each step. Generally, direct modification of program settings is not recommended.

rawFilter `rawFilter` filters data from `weibo2209` or `weibo2209-samples` to extract required datasets. Weibo data can be combined by region and time, with region granularity including national/province/prefecture-level city/user levels, and time granularity including year/quarter/month/week/day/hour. Different combinations enable data export and processing at varying spatiotemporal granularities, effectively creating individual-level datasets.

Each individual dataset is a plain text file, with all individual files stored in a single directory where each file represents one sample. For example, aggregating Weibo data from December 1-31, 2022 by province yields 31 sample files—one per province/municipality—stored together as individual data for each provincial region.

The `filter.ini` configuration file primarily specifies paths and parameters:

```
[path]
inpath = ../data/weibo2209-samples
```

```
outpath = ../data/shanghai

[para]
HAS_{PROVINCE} = False
fProvinces = ../config/myprovinces.txt
HAS_{CITY} = False
fCities = ../config/mycities.txt
HAS_{UID} = False
fUids = ../config/myuids.txt
HAS_{TIME} = True
startTime = 2019-12-28 09:16:00
endTime = 2020-01-02 09:16:00
HAS_{KEYWORDS} = False
fKeywords = ../config/mykeywords.txt
KEYWORD_{EXCLUSIVE} = False
HAS_{REGEX} = False
fRegex = ../config/myregex.txt
EXPORT_{{{WEIBO}}_{{ONLY}}} = False
HAS_{LOCTIME} = False
locTimeType =    plt
fLocTimes = ../config/myloctimes.txt
```

All logical values in the configuration file (`HAS_{PROVINCE}`, `HAS_{CITY}`, `HAS_{UID}`, `HAS_{TIME}`, `HAS_{KEYWORDS}`, `KEYWORD_{EXCLUSIVE}`, `HAS_{REGEX}`, `EXPORT_{{{WEIBO}}_{{ONLY}}}`, `HAS_{LOCTIME}`) default to `False`. If not specified in the `.ini` file, they are treated as `False`, equivalent to explicitly setting them to `False`.

This example configuration filters all Weibo data published between December 28, 2019, 09:16:00 and January 2, 2020, 09:16:00, saving the exported data to the `shanghai` directory.

rawSplitter `rawSplitter` segments filtered data into individual datasets for subsequent psycholinguistic analysis and psychological indicator prediction (e.g., life satisfaction). Its configuration file is `splitter.ini`:

```
[path]
inpath = ../data/weibo2209-samples
outpath = ../data/timeloc/hour

[para]
locType = r
twType = h
EXPORT_{{{WEIBO}}_{{ONLY}}} = False
```

The `inpath` and `outpath` parameters function identically to those in `rawFilter`, specifying input and output directories. `locType` defines the geographic/user

granularity for segmentation:

```
class LocGranularity(Enum):
    PRC = "r"
    PROVINCE = "p"
    CITY = "c"
    UID = "u"
```

Thus, `locType` can only be `r`, `p`, `c`, or `u`. This segmentation aggregates source data (filtered or unfiltered) by specified granularity and outputs individual data files. `r` represents national-level data, `p` provincial-level, `c` prefecture-level city, and `u` individual user.

After geographic aggregation, temporal segmentation can be applied:

```
class TimeGranularity(Enum):
    HOURLY = "h"
    DAILY = "d"
    WEEKLY = "w"
    MONTHLY = "m"
    QUARTERLY = "q"
    YEARLY = "y"
    ENTIRE = "e"
```

`e` indicates no temporal segmentation; `y/q/m/w/d` segment data by year/quarter/month/week/day respectively; `h` segments all data by 24-hour periods within a day.

getWordFreq `getWordFreq` performs word frequency statistics on individual data using specified dictionaries. Each dictionary contains one or more word categories, and the statistical results represent the frequency ratio for each category.

```
[path]
inpath = ../data/timeloc/bs
outpath = ../data/out

[para]
filenameHeader =  市  月
dict_{name} = dic_{weibojibenqingxu}
MIN_{WEIBO} = 10
numProc = 5
```

`inpath` specifies the directory containing individual data files for batch processing. `outpath` defines the output directory for statistical results. `dict_{name}` specifies a dictionary from the `dict` directory. `MIN_{WEIBO}` sets the minimum number of Weibo posts required for word frequency analysis. `numProc` determines the number of processes for parallel processing.

This configuration performs word frequency analysis on the `bs` dataset in `timeloc` using the Weibo Basic Emotion Dictionary, excluding samples with fewer than 10 Weibo posts.

swlsPredict `swls/swlsPredict` predicts life satisfaction by extracting features from each dataset and applying a pre-trained prediction model. Note that sample data must be derived from individual Weibo users, not aggregated provincial or municipal data.

```
[path]
inpath = ../data/timeloc/shanghai
outpath = ../data/out

[para]
filenameHeader = ID
MIN_{WEIBO} = 20
numProc = 8
samplePoolSize = 1000
```

`inpath` specifies the directory containing Shanghai individual data files, where each file represents one sample. Samples with fewer than 20 Weibo posts are excluded from prediction; others undergo feature extraction and life satisfaction prediction using the trained model.

Feature extraction and prediction results are saved as two separate files in `outpath`: - `shanghai_{WordFreqResults}.csv` - `shanghai_{{{result}}_{{SWLS}}}.csv`

System and Custom Dictionaries PsyAnalytics includes the following dictionaries: `dic_{bodyimage}` (body image) [1], `dic_{cmfd2}` (moral foundations) [2], `dic_{daodedongji}` (moral motivation) [3], `dic_{getijitizhuyi}` (individualism-collectivism) [4], `dic_{mfdwu623}` (moral foundations) [5], `dic_{scliwc2024}` (LIWC 2024 Simplified Chinese revision) [6], `dic_{sleepquality}` (sleep quality) [7], `dic_{suicide}` (suicide dictionary) [8], and `dic_{weibojibenqingxu}` (Weibo basic emotion) [9].

Users can also construct custom dictionaries according to research needs. Dictionary development is a complex and meticulous process involving linguistics, computer science, education, and psychology. Conceptualization is the first step, requiring thorough literature review to define the psychological construct, its connotations, and dimensional structure, thereby establishing the dictionary' s framework. Next, initial terms are selected from authoritative measurement tools, standardized psychological tests, questionnaires, and relevant lexical databases, supplemented by validated literature such as academic papers and professional books. Irrelevant terms are filtered according to established criteria, often through independent evaluation by psychology professionals who assess each term' s relevance to the target construct. Since initial terms from authoritative sources tend to be formal and written, lacking everyday usage, ex-

pansion is necessary. Finally, the validity of word frequency as an indicator of the psychological construct is assessed by computing consistency between word frequency results and human ratings.

4 Psychological Semantic Changes Before and After Car Purchase

To compare psychological changes before and after car purchase, we first identify Weibo posts explicitly mentioning "car purchase" and extract their timestamps as purchase time points. We then collect one month of Weibo data before and after each purchase event as pre-test and post-test samples.

The `filter.ini` configuration is:

```
[path]
inpath = ../data/weibo2209-samples
outpath = ../data/buycar

[para]
HAS_{KEYWORDS} = True
fKeywords = ../config/mykeywords.txt
```

Search keywords in `mykeywords.txt` include terms related to car purchase.

Running `rawFilter.py` extracts all Weibo posts containing these keywords. Each filtered post undergoes manual analysis to confirm whether it truly indicates a car purchase event. For large datasets, multiple annotators can be invited to code and label the data. After manual inspection, confirmed purchase events serve as temporal markers.

User IDs and purchase timestamps (Unix format) are stored in a file with one user per line, separated by tabs.

The `genLoctimes.py` program reads this file and generates time intervals before and after each purchase event based on the configuration in `config/loctimes.ini`:

```
[path]
infile = ../data/list.txt
outpath = ../data/out

[para]
timeUnit = m
offSet = -1
```

`timeUnit` specifies the time unit:

```
class TimeUnit(Enum):
    HOUR = "h"
    DAY = "d"
```

```
WEEK = "w"
MONTH = "m"
YEAR = "y"
```

`offSet` is the offset value; negative values indicate the period before purchase (e.g., -1 = one month before), while positive values indicate the period after. After configuring `loctimes.ini`, running the program generates a `myloctimes.txt` file with time intervals.

The `toDateTime.py` program can convert Unix timestamps to standard date-time format.

With confirmed purchase time points, we can filter Weibo data for one month before and after each event using event alignment. For pre-purchase data, the `filter.ini` configuration is:

```
[path]
inpath = ../data/weibo2209-samples
outpath = ../data/pre1m

[para]
HAS_{LOCTIME} = True
locTimeType =    ult
fLocTimes = ../config/myloctimes.txt
```

Running `rawFilter.py` extracts pre-purchase Weibo data. Due to sampling, exported data volumes may be small for statistical analysis. Running on the full `weibo2209` dataset yields more substantial results.

Next, `rawSplitter.py` segments the data with this `splitter.ini` configuration:

```
[path]
inpath = ../data/pre1m
outpath = ../data/timeloc/pre1m

[para]
locType = u
twType = e
```

This aggregates each user' s pre-purchase Weibo data into individual sample files.

Word frequency analysis is then performed using `getWordFreq.py` with this `wordfreq.ini` configuration:

```
[path]
inpath = ../data/timeloc/pre1m
outpath = ../data/out

[para]
```

```
filenameHeader =  人  月
dict_{name} = dic_{weibojibenqingxu}
MIN_{WEIBO} = 5
numProc = 5
```

The output `pre1m_{{{dic}}_{{weibojibenqingxu}}}.txt` contains emotion word frequency ratios for each user during the month before car purchase.

Life satisfaction (SWLS) scores are then predicted using `swls/swls-predict.py` with this `swls.ini` configuration:

```
[path]
inpath = ../data/timeloc/pre1m
outpath = ../data/out

[para]
filenameHeader = ID
MIN_{WEIBO} = 5
numProc = 8
samplePoolSize = 1000
```

The resulting life satisfaction scores enable various comparative analyses. By specifying appropriate start and end times, researchers can analyze the impact of specific events on psychological states.

The advent of big data has provided researchers with powerful analytical capabilities. However, the technical barriers to big data analysis remain challenging for many. To facilitate psychological research using Internet data, we developed PsyAnalytics to enable researchers without programming experience to complete data collection and analysis, thereby advancing scientific exploration.

References

[1] Xinyu Ji, Taotao Zhan, Tingshao Zhu. *(2024) Impact of COVID-19 on negative body image: Evidence based on social media data.* Social Science & Medicine*, Volume 340, 2024. DOI: 10.1016/j.socscimed.2023.116461

[2] Calvin Yixiang Cheng and Weiyu Zhang. (2023) C-MFD 2.0: Developing a Chinese Moral Foundation Dictionary. *Computational Communication Research*, 5(2), 2024. DOI: 10.5117/CCR2023.2.10.CHEN

[3] Zhang, Y., & Yu, F. (2018). Which Socio-Economic Indicators Influence Collective Morality? Big Data Analysis on Online Chinese Social Media. *Emerging Markets Finance and Trade*, 54(4), 792-800.

[4] Ren, X., Xiang, Y., Zhou, Y., & Zhu, T. (2017). A psychological map of Chinese individualism/collectivism based on Weibo big data. *Journal of Inner Mongolia Normal University (Philosophy and Social Science Edition)*, 06, 59-64.

[5] Wu, S., Yang, C., & Zhang, Y. (2019). Introduction and preliminary analysis

of the Chinese version of the Moral Foundations Dictionary. *ChinaXiv.org*, 2019. DOI: 10.12074/201911.00002V1

[6] Cui, X., Chen, S., Zhao, N., Liu, X., & Zhu, T. (2024). Revision and validation of the Simplified Chinese LIWC2024 (SCLIWC2024) dictionary. *ChinaXiv:202404.00159*.

[7] Lin, J., Sun, A., Chen, J., Li, H., & Zhu, T. (2023). Regional differences in sleep quality and their impact on life satisfaction. *ChinaXiv*. doi:10.12074/202307.00067V1

[8] Meizhen Lv, Ang Li, Tainli Liu, Tingshao Zhu. (2015) Creating a Chinese suicide dictionary for identifying suicide risk on social media. *PeerJ* 3:e1455 https://doi.org/10.7717/peerj.1455

[9] Dong, Y., Chen, H., Lai, K., & Yue, G. (2015). Measurement and validity test of basic social emotions on Weibo. *Psychological Science*, 8, 521-528.

Author Contributions

Tingshao Zhu: System design and implementation, manuscript writing

Note: Figure translations are in progress. See original paper for figures.