04_preprocessing

February 22, 2026

1 04 — Preprocessing Pipeline

Apply preprocessing to all three datasets and save cleaned versions to `data/processed/`.

[1]:

```python
import sys
sys.path.append('../src')

from dataset_preprocessor import (
    preprocess_cnhpsx,
    preprocess_pakistan_news,
    preprocess_psx_stocks,
    get_recent_movers
)
```

[4]:

```python
#   CNH-PSX (main recommendation corpus)
df_cnhpsx = preprocess_cnhpsx('../data/raw/CNH-PSX_Ver2.csv')
print(df_cnhpsx.head())
df_cnhpsx.to_csv('../data/processed/cnhpsx_clean.csv', index=False)
```

```
[CNH-PSX] Loaded 8858 clean headlines from ../data/raw/CNH-PSX_Ver2.csv
        date                                           headline category  \
0 2006-04-01                 ['KSE index plunges by 83 points']   Market
1 2006-04-06  ['Karachi stocks record mixed trend,,,,,By our…   Market
2 2006-04-08        ['KSE index touches another all-time high']   Market
3 2006-04-09      ['KSE performs well, index gains 451 points']   Market
4 2006-04-11            ['KSE breaches 12,000 barrier briefly']   Market

      hierarchy                                    headline_clean
0  Stock Market                         kse index plunge 83 point
1  Stock Market  karachi stock record mixed trendby correspondent
2  Stock Market              kse index touch another alltime high
3  Stock Market            kse performs well index gain 451 point
4  Stock Market                  kse breach 12000 barrier briefly
```

[5]:

```python
#   Pakistan News (Word2Vec training corpus)
df_news = preprocess_pakistan_news('../data/raw/pakistan_news.csv')
print(df_news.head())
df_news.to_csv('../data/processed/pakistan_news_clean.csv', index=False)
```

```
[Pakistan News] Loaded 25912 clean articles from ../data/raw/pakistan_news.csv
                                             heading  \
0  Federation of Pakistan v Gen Pervez Musharraf:…
1  Chinese national held for beating traffic poli…
2  Iraqi paramilitaries call for withdrawal from …
3  Sarmad Khoosat reveals why Zindagi Tamasha's t…
4                PSL 2020 set to begin on February 20

                                             excerpt   section  \
0  After the special court's verdict in the high …       NaN
1  The suspect assaulted the constable after he p…  Pakistan
2  Overnight, demonstrators pitched tents and cam…     World
3                       We also have a release date.  Pakistan
4  Of the total 34 matches, 14 will take place in…     Sport

                                       text_combined  \
0  Federation of Pakistan v Gen Pervez Musharraf:…
1  Chinese national held for beating traffic poli…
2  Iraqi paramilitaries call for withdrawal from …
3  Sarmad Khoosat reveals why Zindagi Tamasha's t…
4  PSL 2020 set to begin on February 20 Of the to…

                                          text_clean
0  federation of pakistan v gen pervez musharraf …
1  chinese national held for beating traffic poli…
2  iraqi paramilitaries call for withdrawal from …
3  sarmad khoosat reveals why zindagi tamashas tr…
4  psl 2020 set to begin on february 20 of the to…
```

[6]:

```python
#   PSX Stocks (recency weighting)
df_stocks = preprocess_psx_stocks('../data/raw/psx_stocks.csv')
print(df_stocks.head())
df_stocks.to_csv('../data/processed/psx_stocks_clean.csv', index=False)
```

```
[PSX Stocks] Loaded 813588 trading rows from ../data/raw/psx_stocks.csv
  Tickers: 891 | Date range: 2017-01-02 → 2025-10-24
        date symbol   ldcp  open  high   low  close  change  change_pct  \
0 2018-09-04     786  2.999  3.99  3.99  3.99   3.99   0.991   33.044348
1 2018-09-05     786  3.990  4.99  4.99  4.99   4.99   1.000   25.062657
2 2018-09-06     786  4.990  5.99  5.99  5.99   5.99   1.000   20.040080
3 2018-09-07     786  5.990  6.99  6.99  6.61   6.99   1.000   16.694491
4 2018-09-10     786  6.990  7.98  7.98  6.99   7.39   0.400    5.722461

   volume
0   21000
1  150500
2   10500
3  214500
```

```
4   130000
```

[7]:

```
#   Top movers on latest date
top_movers = get_recent_movers(df_stocks, top_n=20)
print('Top movers on latest date:')
print(top_movers)
```

```
Top movers on latest date:
     symbol   close  change_pct    volume
0       CTM    7.82   14.662757   4990550
1      PINL    9.72   11.467890   1533314
2      ARUJ   10.68   10.330579    449895
3     FCIBL   17.87  -10.020141    217787
4       ZAL   40.11   10.010971   7035804
5      FECM   51.54   10.010672     39895
6      TSMF   17.44  -10.010320    157515
7      UDLI   18.36   10.005992    138350
8      BECO   56.51   10.005840   9861563
9      SERT   64.06  -10.002810    365892
10     UDPL  133.31   10.000825    126875
11  GEMNETS   39.60   10.000000      3800
12     TSBL   13.75   10.000000     79521
13     FPJM   11.33   10.000000   1442515
14     TATM  142.47    9.998456    604191
15     FNEL   14.75    9.992543  29300149
16     BIPL   33.87   -9.992028   2717044
17     PKGP   72.98   -9.956817     58964
18     KOHP   40.97   -9.956044   7965733
19    LOADS   18.00   -9.774436  15084723
```

[8]:

```
#   Quick sanity check
print('CNH-PSX clean shape:', df_cnhpsx.shape)
print('Pakistan News clean shape:', df_news.shape)
print('PSX Stocks clean shape:', df_stocks.shape)
print('\nCNH-PSX nulls:')
print(df_cnhpsx.isnull().sum())
print('\nPakistan News nulls:')
print(df_news.isnull().sum())
```

```
CNH-PSX clean shape: (8858, 5)
Pakistan News clean shape: (25912, 5)
PSX Stocks clean shape: (813588, 10)

CNH-PSX nulls:
date              0
headline          0
category          0
hierarchy         0
headline_clean    0
```

```
dtype: int64

Pakistan News nulls:
heading           0
excerpt           0
section          53
text_combined     0
text_clean        0
dtype: int64
```

1.1 Preprocessing Conclusions

- □ CNH-PSX final shape: ...
- □ Pakistan News final shape: ...
- □ PSX Stocks final shape: ...
- □ Ready for embeddings: yes/no