# Exploring Supporting-Phenomena to Improve Official Statistics by Using Natural Language Processing (NLP): A Case Study in East Java, Indonesia

UNECE Expert Meeting on Statistical Data Collection

12 – 14 Juni 2023

Joko Ade Nursiyono, joko.ade@bps.go.id
Ima Sartika Dewi, imasartika@bps.go.id

# Supporting-Phenomena as Evidence Base of Official Statistics

- Public awareness of official statistics has been raised over time
- Beside maintaining data collection process, data quality can also be improved by obtaining evidence base
- Supporting-phenomena can be very useful to complement survey-based Official Statistics products
- Text-and news-based measures as source of information to provide supporting-phenomena

# Data Source and Data Collection

## Economic Phenomena in East Java Province, Indonesia

News scraping from several news sites

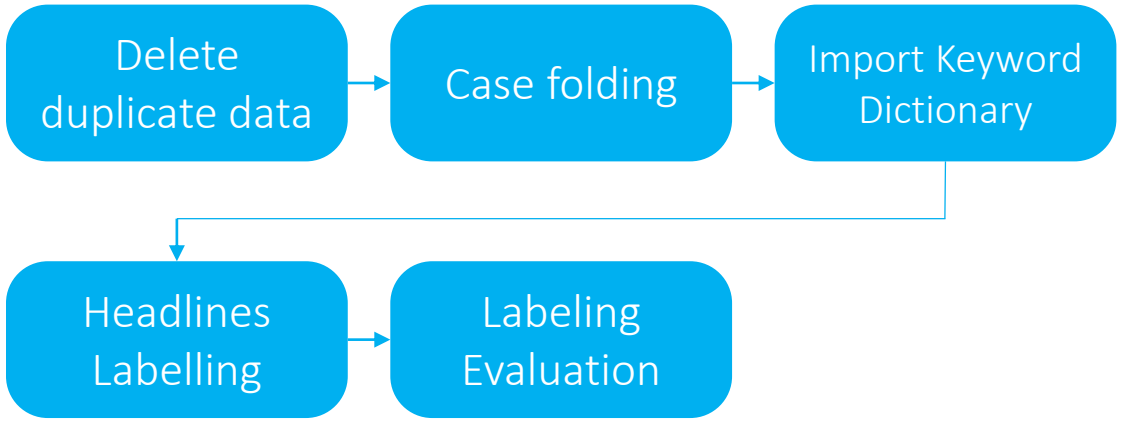
(surabayatoday.id, wartaekonomi.co.id, investor.id, cnbcindonesia.com, beritajatim.com, jatim.antaranews.com.)

521 news text data were obtained

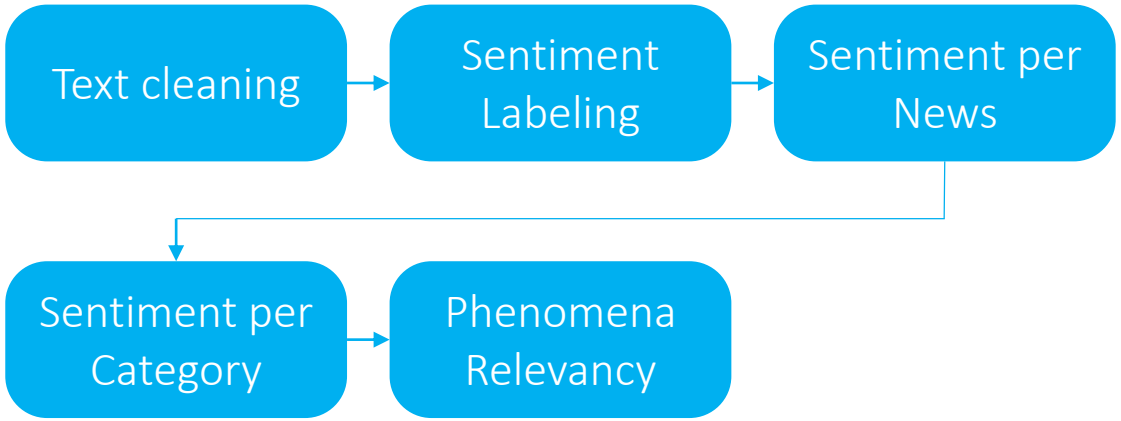
Economic growth data of Quarter IV – 2022 (q to q) in East Java Province

# Data Processing

## News Headlines Labeling



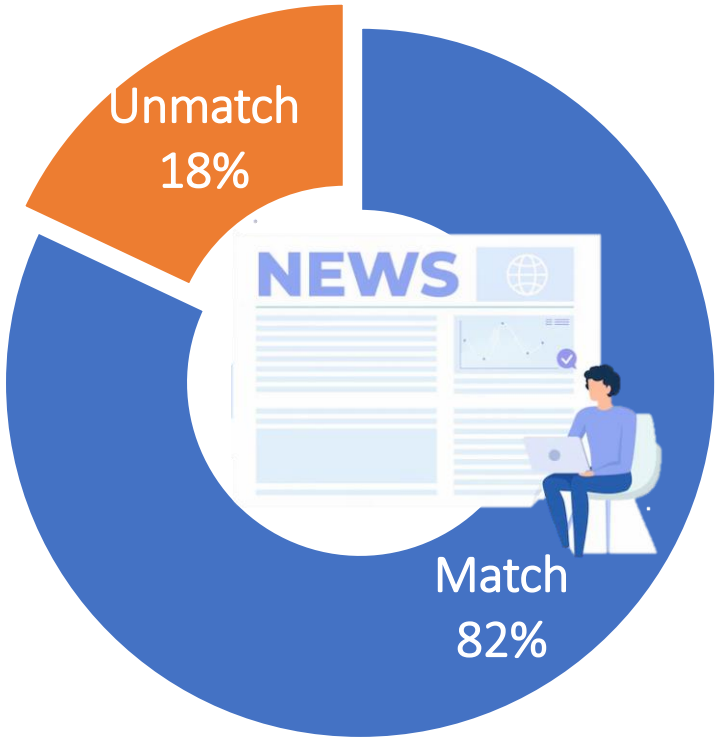

## Phenomena Relevance




News Headlines Labeling using Natural Language Processing (NLP) with Latent Dirichlet Algorithm (LDA).

# What we can learn

The accuracy of news headlines labeling




To improve the accuracy of news headlines labeling by increasing the effectiveness of the title labeling dictionary

# What we can learn (2)

"The correlation of sentiment news phenomena with economic growth is weak (namely 0.22), but the **majority sentiment is in line with economic growth per category of business fields"**

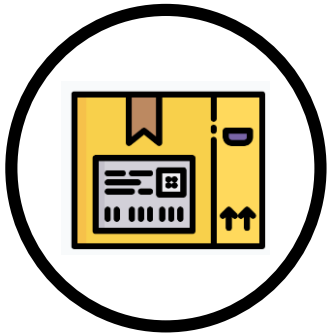
## Phenomena data relevancy to economic growth

| Category | Mean Sentiment | q to q | Sentiment status | q to q status | GRDP | Accuracy | Total Accuracy |
|---|---|---|---|---|---|---|---|
| Accomodations | -2.86 | 2.89 | negative | positive | Production Account Component | 73.33 | 72.22 |
| Manufacturing | 4.56 | 1.44 | positive | positive | | | |
| Forestry | 5.00 | 14.07 | positive | positive | | | |
| Financial and Insurance Activities | 5.66 | 1.54 | positive | positive | | | |
| Construction | 0.43 | 4.06 | positive | positive | | | |
| Education | 8.00 | 5.59 | positive | positive | | | |
| Water Supply Activities | -9.56 | -0.66 | negative | negative | | | |
| Electricity | 6.00 | 1.98 | positive | positive | | | |
| Gas | 4.03 | 6.19 | positive | positive | | | |
| Food Service Activities | -5.60 | 4.56 | negative | positive | | | |
| Wholesale and Retail Trade | 8.00 | 1.37 | positive | positive | | | |
| Agriculture | -2.79 | -24.76 | negative | negative | | | |
| Land Transport | -2.86 | 3.73 | negative | positive | | | |
| Sea Transport | 1.50 | 1.34 | positive | positive | | | |
| Air Transport | -1.78 | 6.37 | negative | positive | | | |
| Export | 6.82 | -6.60 | positive | negative | Good and Services Account Component | 66.67 | |
| Import | -2.50 | -3.45 | Negative | Negative | | | |
| Gross Fixed Capital Formation | 7.26 | 2.70 | positive | positive | | | |

# Concluding Remark

- NLP is able to categorize news item text precisely as much as 82.34%
- The accuracy of supporting-phenomena to official statistics on economic growth by industries is 73.33% and by expenditure is 66.67%
- The accuracy of news phenomena to economic growth in total reaches 72.22%
- **Thus, NLP are very helpful in producing supporting phenomena data as an evidence base of official statistics.**

# In the future…



Add more label dictionary in Natural Language Processing

Improve the quality of the sentiment label dictionary


Build a database of news phenomena

# Thank You!

joko.ade@bps.go.id
imasartika@bps.go.id