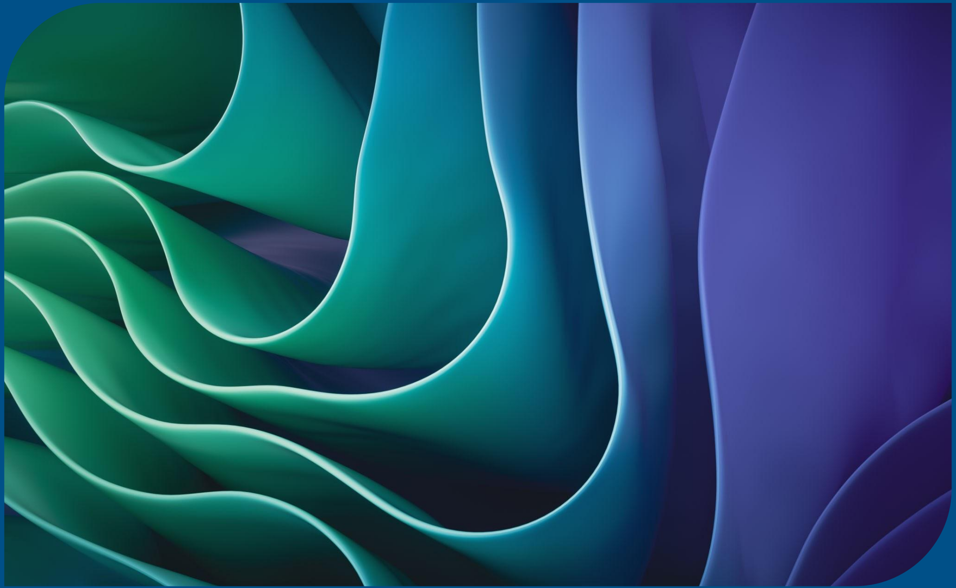
04-07-2025

FairNM: Fairness in Name Matching

Yuan Liu and Flavius Frasincar



NLDB 2025

1. **Motivation and Contribution**
2. Data and Data Insights
3. Methodology
4. Evaluation and Results
5. Conclusion
6. Future Work and References

Motivation

Motivation

- **Person names** are increasingly used in automated high-stakes decisions (e.g., hiring, fraud detection) through **Entity Matching (EM)** techniques

Motivation

- **Person names** are increasingly used in automated high-stakes decisions (e.g., hiring, fraud detection) through **Entity Matching (EM)** techniques
    - *Example:* can we find this applicant on our blacklist?

Motivation

- **Person names** are increasingly used in automated high-stakes decisions (e.g., hiring, fraud detection) through **Entity Matching (EM)** techniques
- A common method in EM, called **Fuzzy Name Matching,** focuses on linking different variations of the same name

Motivation

- **Person names** are increasingly used in automated high-stakes decisions (e.g., hiring, fraud detection) through **Entity Matching (EM)** techniques
- A common method in EM, called **Fuzzy Name Matching,** focuses on linking different variations of the same name
  - *Example:* Johnny Smith refers to the same person as John Smith

Motivation

- **Person names** are increasingly used in automated high-stakes decisions (e.g., hiring, fraud detection) through **Entity Matching (EM)** techniques
- A common method in EM, called **Fuzzy Name Matching,** focuses on linking different variations of the same name
- Previous work highlighted **algorithmic bias** in Fuzzy Name matching tasks between racial groups

Motivation

- **Person names** are increasingly used in automated high-stakes decisions (e.g., hiring, fraud detection) through **Entity Matching (EM)** techniques
- A common method in EM, called **Fuzzy Name Matching,** focuses on linking different variations of the same name
- Previous work highlighted **algorithmic bias** in Fuzzy Name matching tasks between racial groups
- However, underlying causes remain unexplored, and no effective mitigation strategies have been proposed

Motivation

Names can encode **sensitive information** about a person

Ethnicity:

- Jamal vs Hiroshi

Socioeconomic Status

- Alastor vs Destiny

Gender Identity

- Sarah vs Eric

Religious Preference

- Mohammed vs Ezekiel

Contributions

- **Investigate underlying factors** contributing to bias
- **Introduce a novel fairness measure** and **test bench** for evaluating algorithmic bias
- **Propose mitigation strategies** to improve fairness in name matching

Data Collection

- Group names by **linguistic origin** instead of race

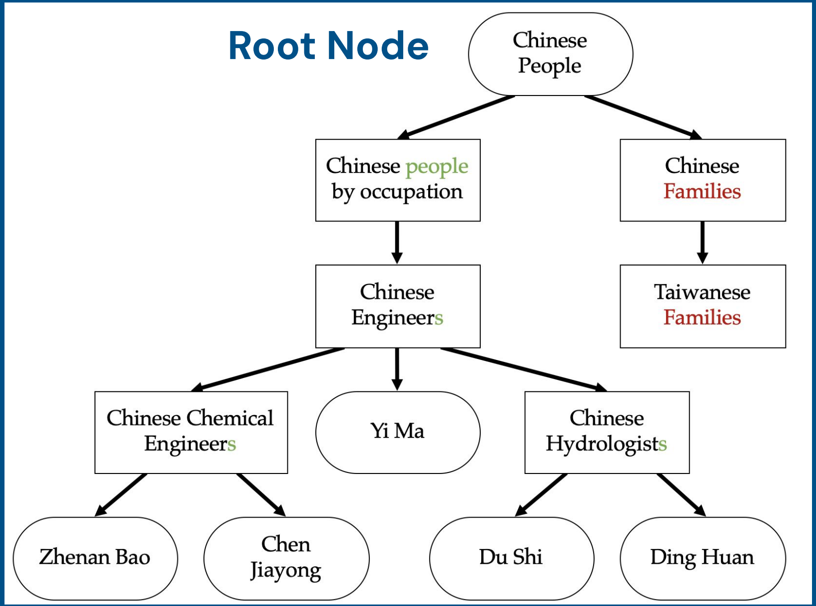
Data Collection

- Group names by **linguistic origin** instead of race
- No **public** labeled dataset for this task

Data Collection

- Group names by **linguistic origin** instead of race
- No **public** labeled dataset for this task
- Built a **custom dataset** via Wikipedia **scraping, filtering, and manual curation**

Schematized Overview



Leaf Node

Data Collection

- Group names by **linguistic origin** instead of race
- No **public** labeled dataset for this task
- Built a **custom dataset** via Wikipedia **scraping, filtering, and manual curation**

Root Node

Category:Chinese people

120 languages

Subcategories

This category has the following 42 subcategories, out of 42 total.

- People by organization in China (8 C)
- People by city in China (306 C)
- Chinese people by ethnic or national origin (21 C, 2 P)
- Chinese people by legal status (7 C)
- Chinese people by migration status (5 C)
- Chinese people by occupation (94 C)
- People's Republic of China people (2 C)
- Chinese people by period (9 C)
- Chinese people by political orientation (20 C)
- People by province in China (35 C)
- People by region in China (6 C)
- Chinese people by religion (14 C)
- Chinese people by role (4 C)

*

- People by non-city prefecture-level divisions in China (31 C)
- Lists of Chinese people (8 C, 42 P)

+

- Articles about multiple Chinese people (11 P)
- Chinese men (9 C)
- Chinese women (19 C, 6 P)

A

- Chinese award winners (4 C, 1 P)

B

- Chinese billionaires (10 C, 93 P)

C

- Chinese centenarians (3 C, 145 P)
- Chinese children (5 C, 15 P)
- Chinese imperial consorts (9 C, 14 P)
- Cultural depictions of Chinese people (3 C, 1 P)

D

- Deified Chinese people (11 C, 108 P)
- Chinese diaspora (23 C, 39 P)
- Chinese people with disabilities (6 C, 18 P)
- Chinese duellists (19 P)

E

- Chinese Esperantists (12 P)

F

- Chinese families (8 C, 4 P)
- Fictional Chinese people (8 C, 37 P)

H

- Chinese hermits (13 P)
- Hong Kong people (25 C, 7 P)

L

- Chinese LGBT people (6 C, 8 P)

M

- Macau people (12 C, 36 P)

N

- Chinese Nobel laureates (5 C, 1 P)
- Chinese nobility (14 C, 42 P)

T

- Chinese twins (1 C, 36 P)

W

- Chinese war casualties (3 C, 1 P)
- Chinese whistleblowers (9 P)

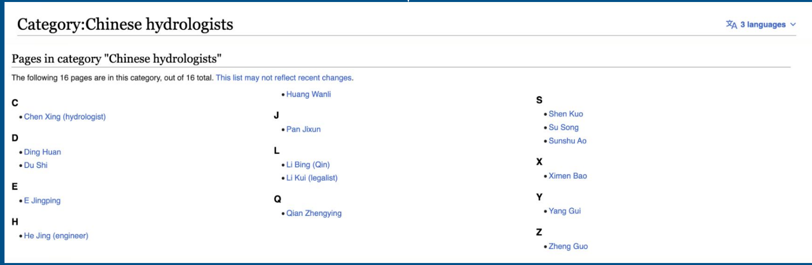
Σ

- Chinese people stubs (16 C, 561 P)

Ω

- Wikipedia categories named after Chinese people (11 C)

Leaf Node



Category:Chinese hydrologists

3 languages

Pages in category "Chinese hydrologists"

The following 16 pages are in this category, out of 16 total. This list may not reflect recent changes.

C

- Chen Xing (hydrologist)

D

- Ding Huan
- Du Shi

E

- E Jingping

H

- He Jing (engineer)
- Huang Wanli

J

- Pan Jixun

L

- Li Bing (Qin)
- Li Kui (legalist)

Q

- Qian Zhengying

S

- Shen Kuo
- Su Song
- Sunshu Ao

X

- Ximen Bao

Y

- Yang Gui

Z

- Zheng Guo

Data Collection

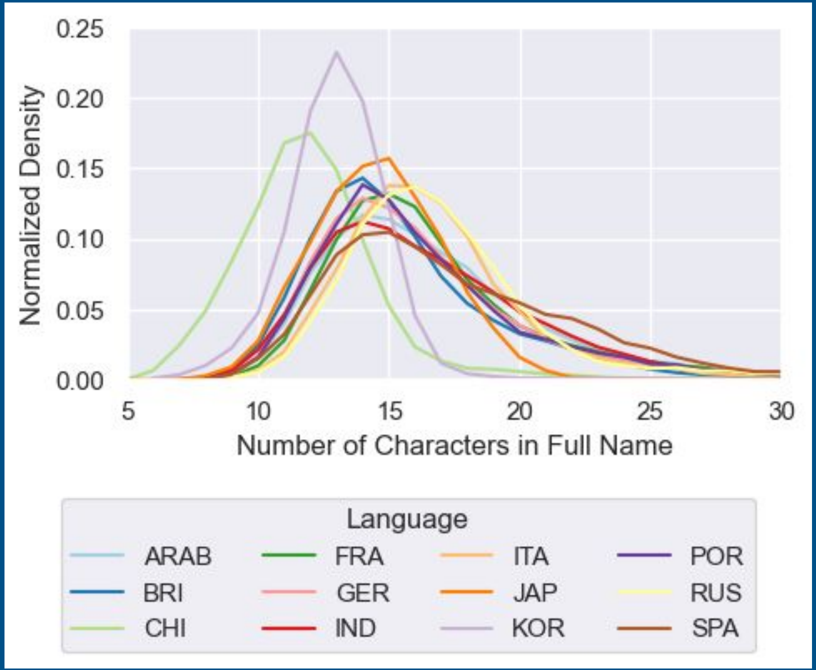
- Group names by **linguistic origin** instead of race
- No **public** labeled dataset for this task
- Built a **custom dataset** via Wikipedia **scraping, filtering, and manual curation**

Language	# Names	Example Name	Root Category
BRI	57,385	Sarah Wigglesworth	British
FRA	36,683	Hervé Berville	French
GER	29,286	Magdalena Neuner	German
IND	26,146	Sukhbir Sinha	Indian
ITA	21,769	Marcello Farabegoli	Italian
RUS	17,801	Alexander Abrosimov	Russian
JAP	16,612	Toshizō Nishio	Japanese
SPA	14,462	Begoña Sánchez	Spanish
POR	13,632	Candido Rondon	Portuguese, Brazilian
CHI	13,378	Baofang Jin	Chinese
ARAB	7,036	Mohamed Al-Fayed	Egyptian, Iraq, Saudi Arabian
KOR	5,960	Seung-hyun Choi	South Korean
Total	260,150		

Data Insights

Data Insights

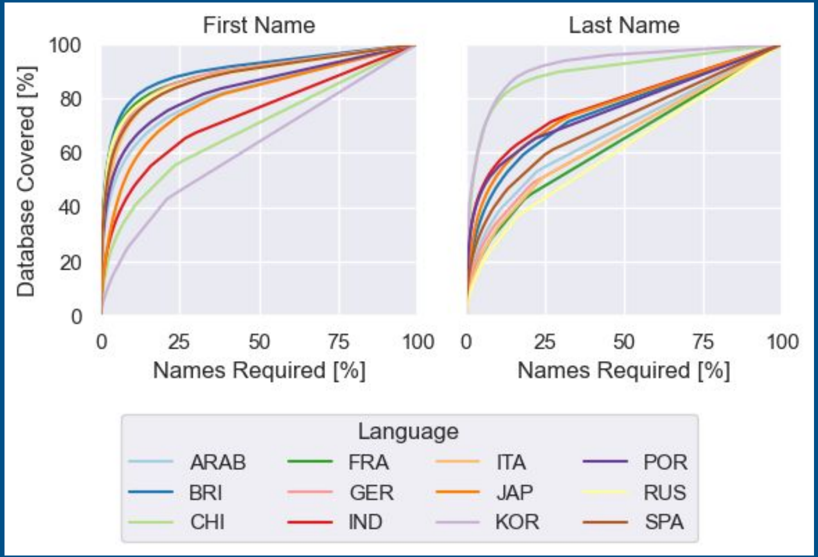
- **Chinese** and **Korean** full names are typically shorter and have less variation in name length



Normalized Name Length Density

Data Insights

- **Chinese** and **Korean** full names are typically shorter and have less variation in name length
- **Chinese** and **Korean** names have lower variation in last name, but higher variation in first name



Percentage of Unique Names Required to Cover Percentage of Database

1. Motivation and Contributions
2. Data Collection and Insights
3. **Methodology**
4. Evaluation
5. Conclusion
6. Future Work and References

Methodology

Methodology

- **Simulate real-world Fuzzy Name Matching** using a screening list scenario

Methodology

- **Simulate real-world Fuzzy Name Matching** using a screening list scenario
- **Evaluate performance across different linguistic groups** by using single origin screening name lists

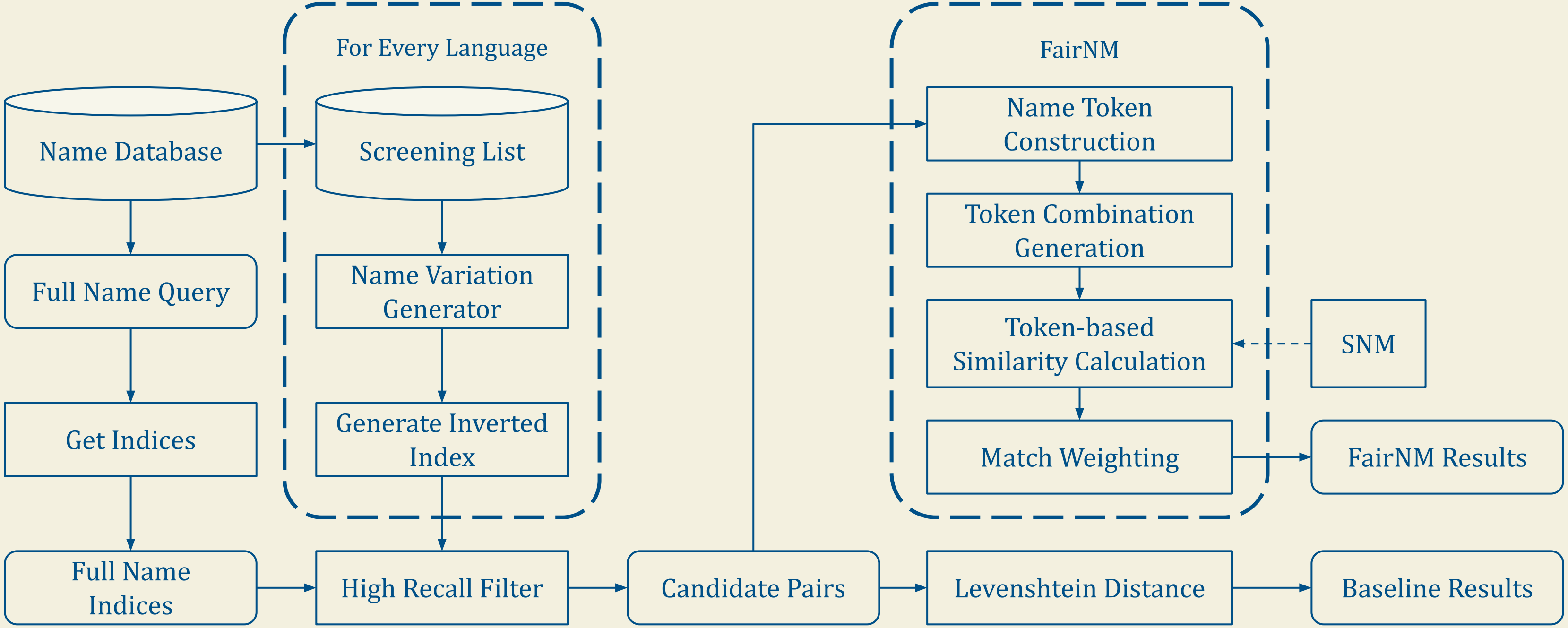
Methodology

- **Simulate real-world Fuzzy Name Matching** using a screening list scenario
- **Evaluate performance across different linguistic groups** by using single origin screening name lists
- **Introduce name variations** to create realistic fuzziness in the input
  - **Original Name:** John Smith Doe
  - **Swapped Tokens:** Doe John Smith
  - **Random Deletion:** Jon Smith Doe
  - **Fat-Finger Error:** Jonh Smith Doe

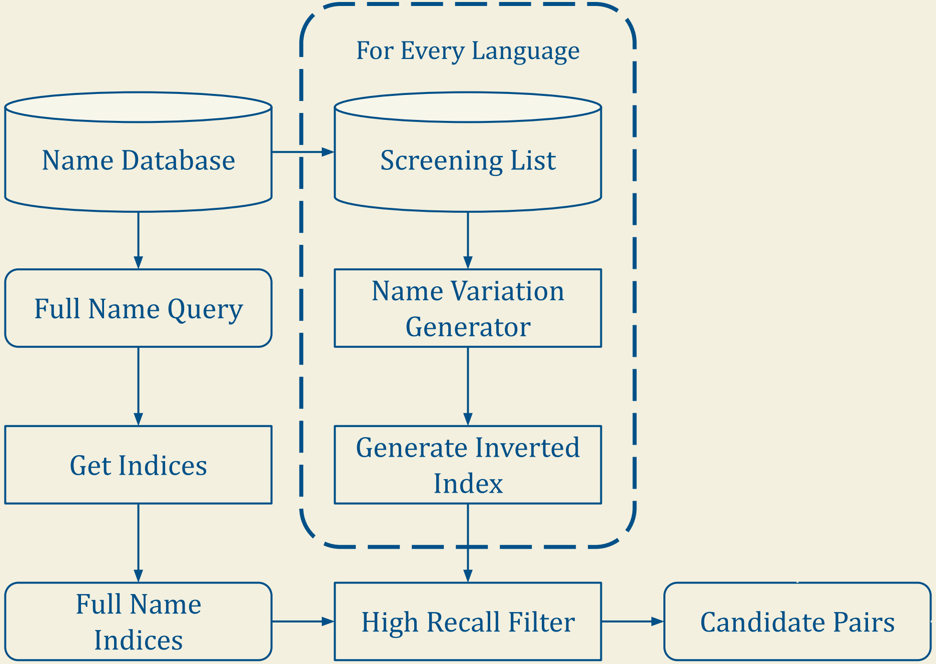
Methodology

- **Simulate real-world Fuzzy Name Matching** using a screening list scenario
- **Evaluate performance across different linguistic groups** by using single origin screening name lists
- **Introduce name variations** to create realistic fuzziness in the input
  - **Original Name:** John Smith Doe
  - **Swapped Tokens:** Doe John Smith
  - **Random Deletion:** Jon Smith Doe
  - **Fat-Finger Error:** Jonh Smith Doe
- **Objective:** Accurately link each name variation back to its original canonical form

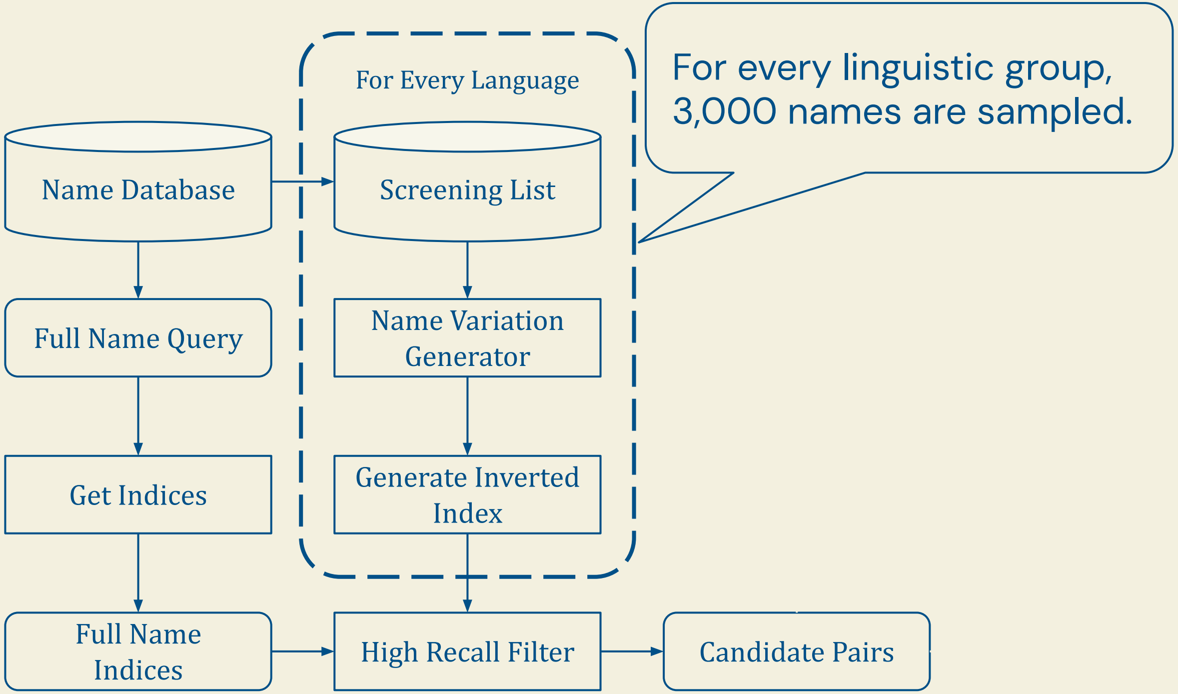
Test Bench Overview



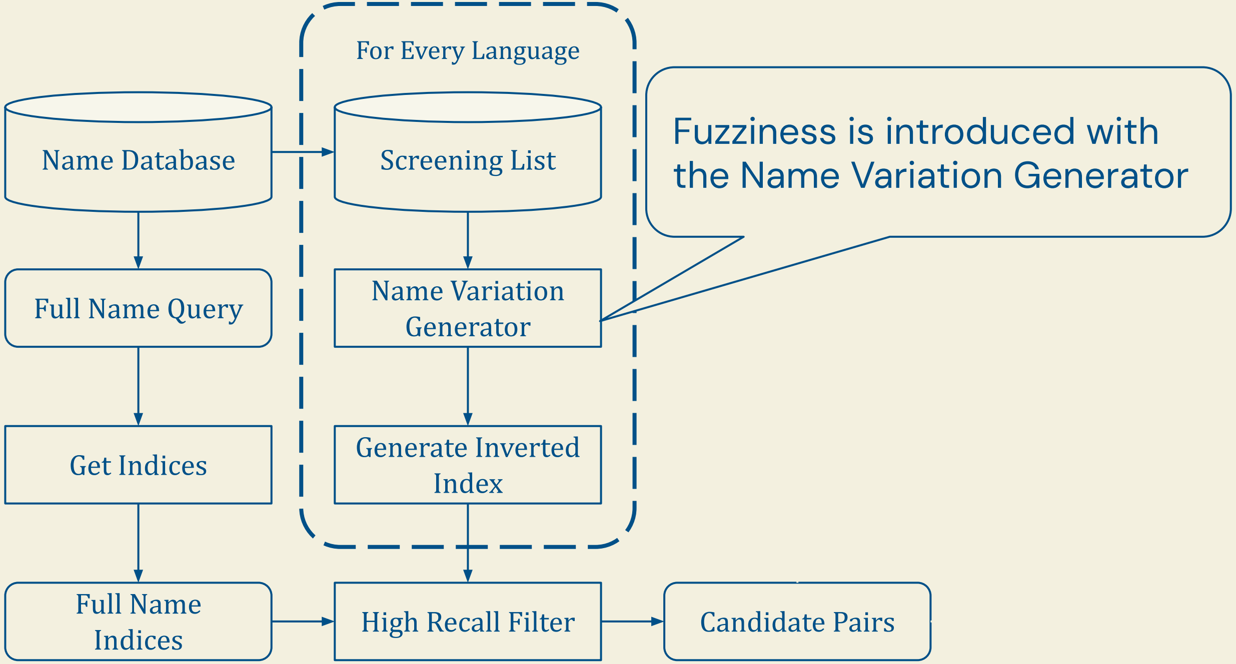
Step 1
High Recall Filter



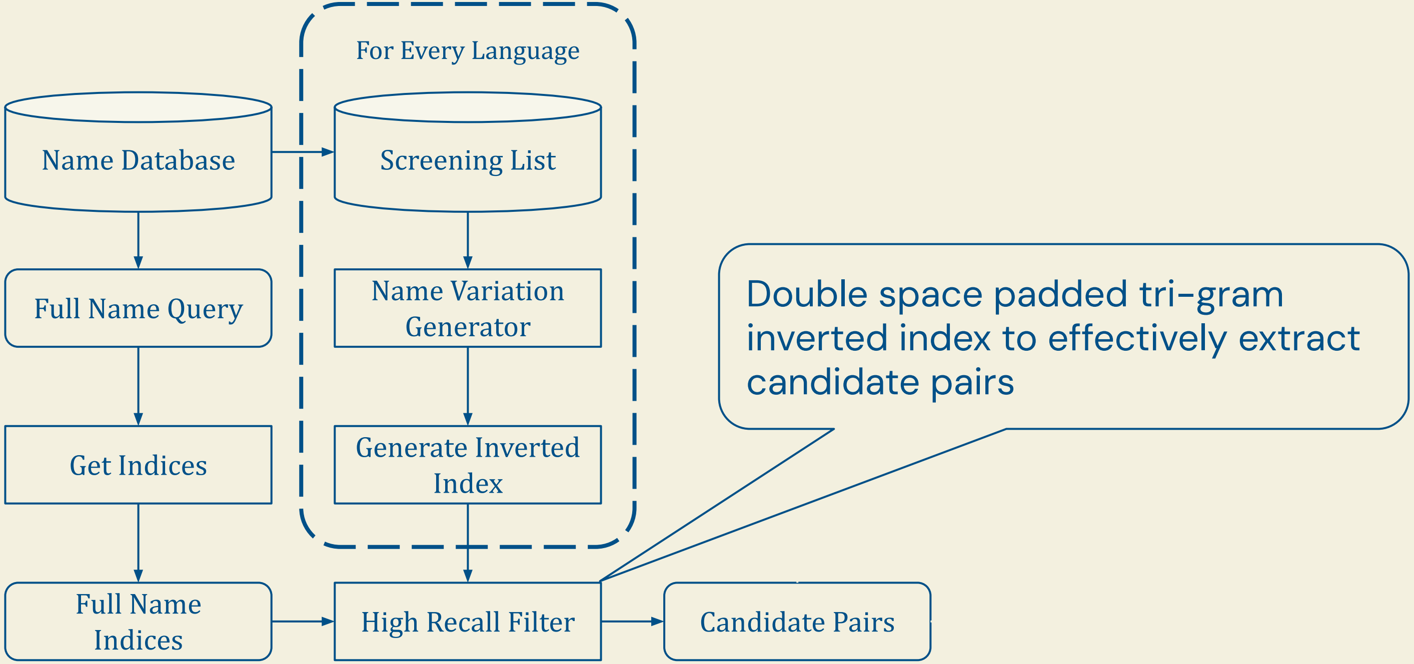
Step 1
High Recall Filter



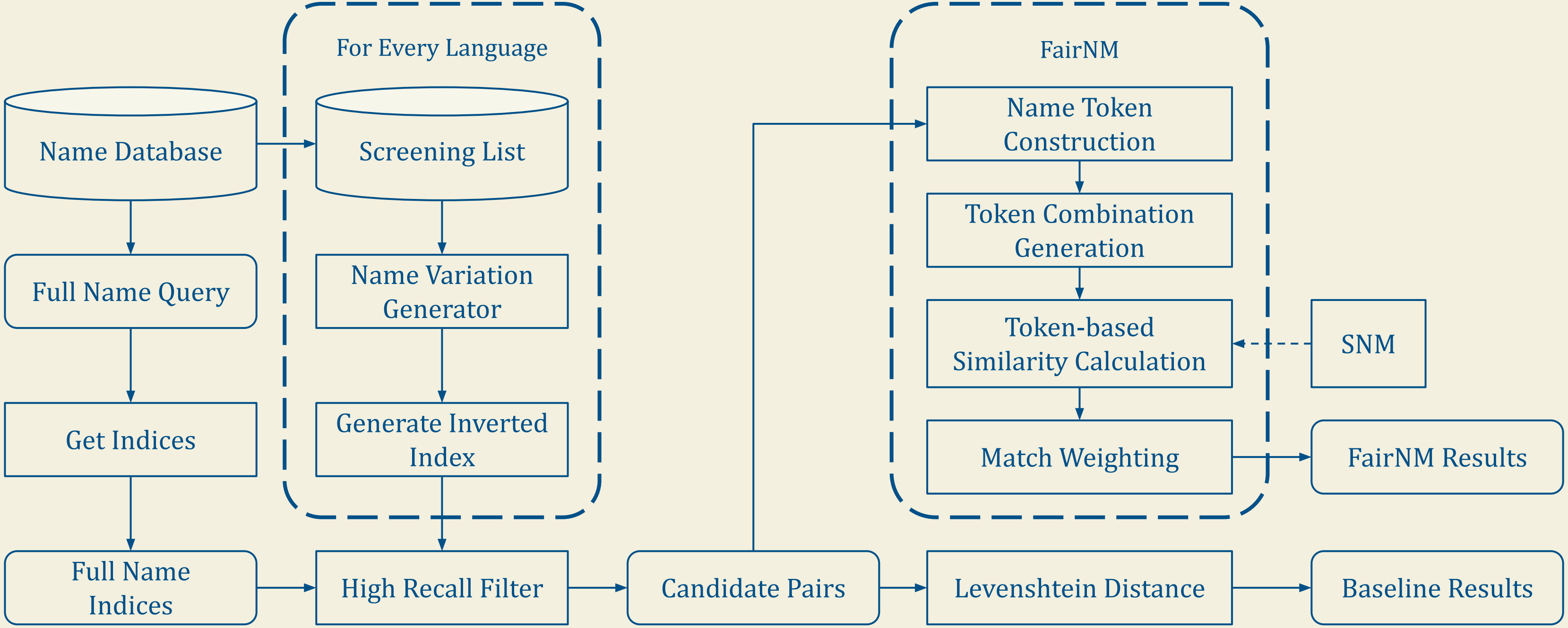
Step 1
High Recall Filter



Step 1
High Recall Filter



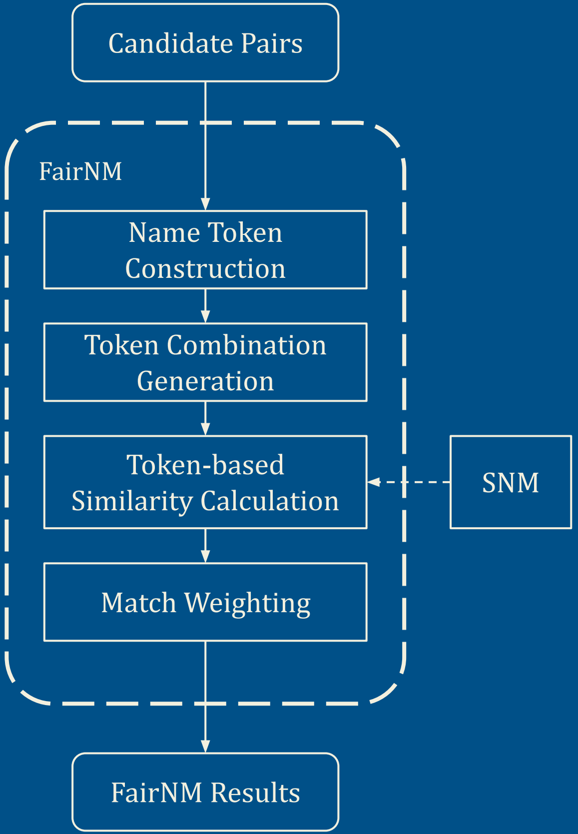
Step 2
Similarity Scoring



FairNM Overview

Modules

1. Token-based Normalized Levenshtein Similarity
2. Short Name Module (SNM)
3. Match Weighting



Token-based Normalized Levenshtein Similarity

- Uses normalized Levenshtein similarity per token combination
- Handles swapped tokens and finds optimal alignment

Example:
Yuan Lin vs Enzo Liu, Yuan

	Yuan	Lin
Enzo	0.00	0.00
Liu	0.00	0.67
Yuan	1.00	0.25

Token-based Normalized Levenshtein Similarity

- Uses normalized Levenshtein similarity per token combination
- Handles swapped tokens and finds optimal alignment

Example:
Yuan Lin vs Enzo Liu, Yuan

	Yuan	Lin
Enzo	X	X
Liu	X	0.67
Yuan	1.00	X

Average score of **0.84**

Token-based Normalized Levenshtein Similarity

- Uses normalized Levenshtein similarity per token combination
- Handles swapped tokens and finds optimal alignment
- Also checks for merged tokens

	YuanLiu
Yuan	X
Liu	X
YuanLiu	1.00

Short Name Module (SNM)

- Siamese Neural Network (SNN) trained on short name tokens (<4 char)
- Boosts performance for languages with typical short name tokens

Example:
Yuan Lin vs Enzo Liu, Yuan

	Yuan	Lin
Enzo	X	X
Liu	X	
Yuan	1.00	X

Average score of **0.84**

Short Name Module (SNM)

- Siamese Neural Network (SNN) trained on short name tokens (<4 char)
- Boosts performance for languages with typical short name tokens

Example:
Yuan Lin vs Enzo Liu, Yuan

	Yuan	Lin
Enzo	X	X
Liu	X	0.23
Yuan	1.00	X

Average score of **0.62**

Match Weighting

- Weights name tokens using Inverse Document Frequency (IDF)
- Reduces emphasize on matches in common names

$$simScore = \sum_{m=1}^{M} \frac{w_{m1} + w_{m2}}{w_{tot}} \times sim_m$$

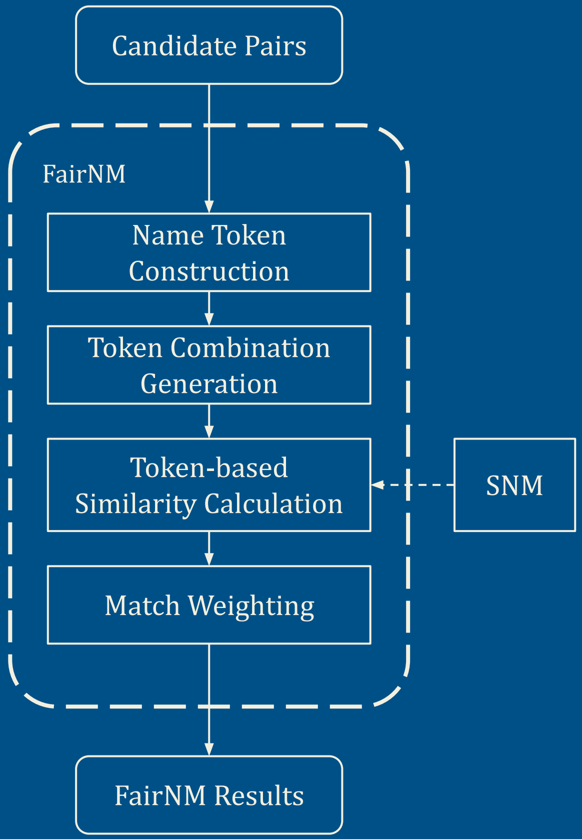
	Ho-young (74.8%)	Lee (25.2%)
Hyun-Jun (72%)	0.36	X
Lee (28.0%)	X	1.00

Unweighted Score: **0.68**
Weighted Score: **0.53** (↓0.15)

FairNM Overview

Modules

1. Token-based Normalized Levenshtein Similarity
2. Short Name Module (SNM)
3. Match Weighting



Evaluation

- **Performance:**
- **Fairness:**

- **Goal:** Evaluate and balance the trade-off between Performance and Fairness

Evaluation

- **Performance: F1 score**
  - Measured in recall-restricted and unrestricted setups
- **Fairness:**

- **Goal:** Evaluate and balance the trade-off between Performance and Fairness

Evaluation

- **Performance: F1 score**
  - Measured in recall-restricted and unrestricted setups
- **Fairness:**
  - Existing metrics (e.g., Equalized Odds) focus only on recall or precision parity

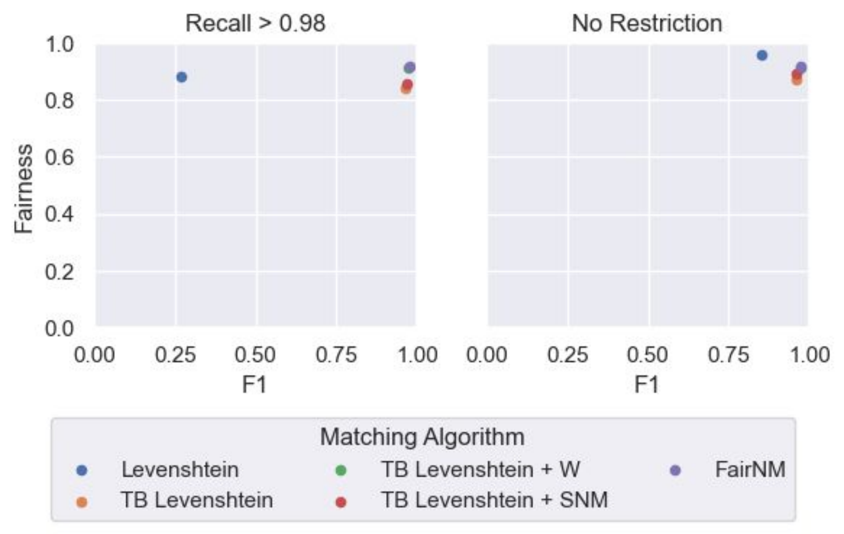
- **Goal:** Evaluate and balance the trade-off between Performance and Fairness

Evaluation

- **Performance: F1 score**
  - Measured in recall-restricted and unrestricted setups
- **Fairness: Novel Metric**
  - Existing metrics (e.g., Equalized Odds) focus only on recall or precision parity
  - We propose a new fairness metric combining recall + precision parity. In this metric a score of 1.0 means perfect Fairness between subgroups.
- **Goal:** Evaluate and balance the trade-off between Performance and Fairness

$$1 - \frac{\Delta p + \Delta r}{2} = 1 - \frac{1}{2}\left(\left(\max_{\forall x \in A} p_x - \min_{\forall x \in A} p_x\right) + \left(\max_{\forall x \in A} r_x - \min_{\forall x \in A} r_x\right)\right)$$

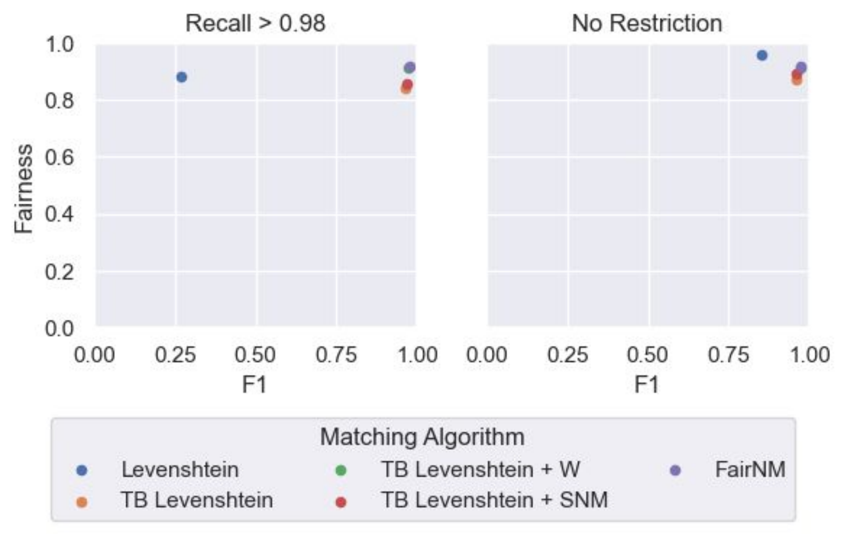
Results



- TB = Token-Based
- W = Match Weighting
- SNM = Short Name Module

Results

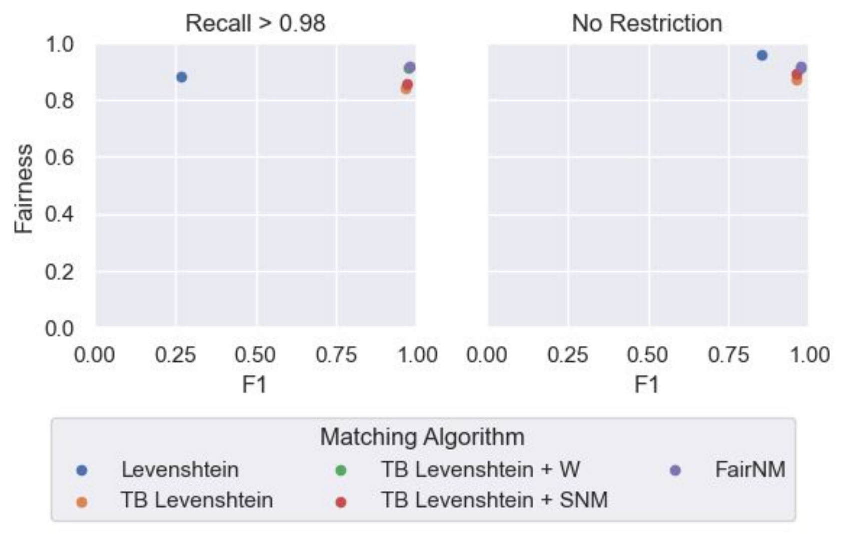
- Both F1 and Fairness increase, **especially under recall constraints**



- TB = Token-Based
- W = Match Weighting
- SNM = Short Name Module

Results

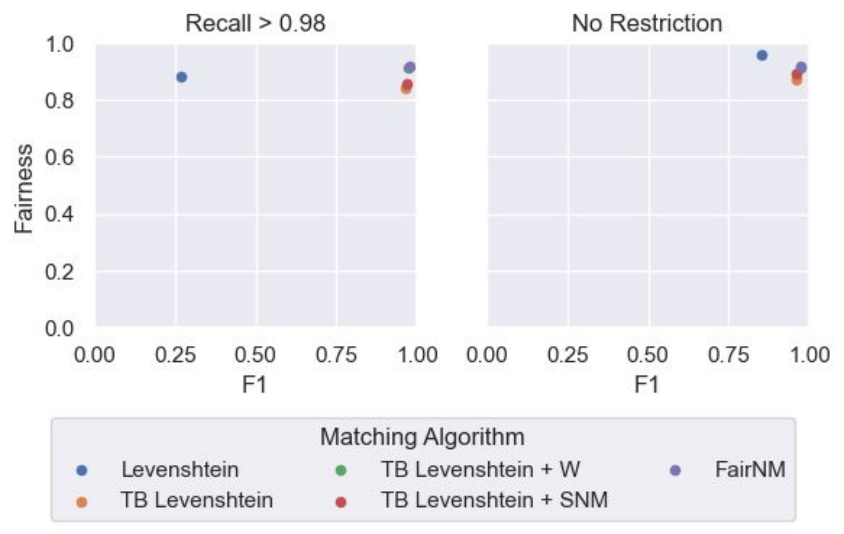
- Both F1 and Fairness increase, **especially under recall constraints**
- Each module contributes **independently** and **complementarily**



TB	=	Token-Based
W	=	Match Weighting
SNM	=	Short Name Module

Results

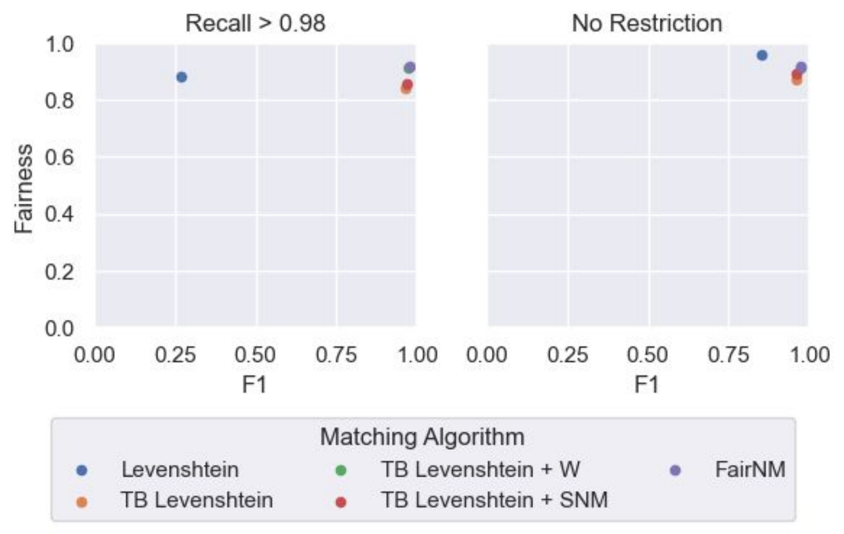
- Both F1 and Fairness increase, **especially under recall constraints**
- Each module contributes **independently** and **complementarily**
- Match Weighting provides the **largest fairness gain**, especially in recall-restricted scenarios



- TB = Token-Based
- W = Match Weighting
- SNM = Short Name Module

Results

- Both F1 and Fairness increase, **especially under recall constraints**
- Each module contributes **independently** and **complementarily**
- Match Weighting provides the **largest fairness gain**, especially in recall-restricted scenarios
- **No Fairness–performance trade-off**: FairNM shows both can be improved simultaneously



- TB = Token-Based
- W = Match Weighting
- SNM = Short Name Module

Conclusion

- There are **significant** differences in name characteristics across linguistic groups
- The proposed **FairNM modules** effectively reduce algorithmic bias, each addressing **different fairness dimensions**
- FairNM performs especially well in **high-recall settings**, where fairness is paramount

1. Motivation and Contributions
2. Data Collection and Insights
3. Methodology
4. Evaluation
5. Conclusion
6. **Future Work and References**

Future Work

- Explore additional linguistic groups;
- Investigate different “sub-categorization strategies” beyond linguistic origin;
- Validate performance on real-life dataset;
- Benchmark against recent LLM techniques.

References

- Karakasidis, A., Pitoura, E.: Identifying bias in name matching tasks. In: Proceedings of the 22nd International Conference on Extending Database Technology (2019)
- Ambekar, A., Ward, C.B., Mohammed, J., Male, S., Skiena, S.: Name-ethnicity classification from open sources. In: Proceedings of the 15th SIGKDD '09, ACM (2009) In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 49–58. SIGKDD '09, ACM (2009)
- Hosseini, K., Nanni, F., Ardanuy, M.C.: Deezymatch: A flexible deep learning approach to fuzzy string matching. In: Proceedings of the 25th Conference on Empirical Methods in Natural Language Processing. pp. 62–69. EMNLP '20, ACL (2020)

Further Information

- Code is written in Python
- Code is available under: https://github.com/yuanliueur/FairNM
- Contact: yuanliuenzo@gmail.com if you have further questions