A Machine Learning Approach to User Profiling for Data Annotation of Online Behavior

**Moona Kanwal1\*, Najeed A. Khan1, Aftab A. Khan2**

1 Computer Science and Information Technology Department, NED University of Engineering and Technology, Karachi, Pakistan

2 College of Education, Health and Human Services, Longwood University, VA, USA

\*Corresponding Author: Moona Kanwal. Email: mkkhan221@cloud.neduet.edu.pk

**Supplementary Figures and Tables**

1. **Question Sample in the Questionnaire**

To depict Two Questions were explicitly put in the form of a case study to observe the actions of users when it comes to information dissemination and verification. Both the case studies presented fake news (at the time of this study Queen Elizabeth II was alive) propagating on the internet. The question in Figure 1a and Figure 1b are labeled as Case 1 and 2, respectively.

Fig SA- . Case 1

Graphical user interface, text, application, email

Description automatically generated

Fig SA- . Case 2

Graphical user interface, application

Description automatically generated

1. **Data Collection and Preprocessing**

Table S. Participants Demographic Information

| **Category** | **Options (Participation (%)** |
| --- | --- |
| Gender | Male (57), Female (40) |
| Age Group | 18-40 (63), > 40 (34) |
| Occupation Sector | Science and Technology (27), Education (20). Finance and Insurance (14). Unemployed (16), Others (20) |

**Table SII:**  Variable and Categories after Initial Preprocessing

|  |  |
| --- | --- |
| Main variables | Categories |
| Gender | Male, Female |
| Age | Adult (18 – 40 yrs), Older Adults (> 40 yes) |
| Occupation | Science and Technology, Finance and Insurance,  Education, Unemployed/Homemaker, Others. |
| Browsing Preference | News Channels-Internationals, News Channel-Local, Talk Shows, Financial websites, e-commerce websites, educational and technology websites, Entertainment websites, Nothing particular |
| Search Reasons | Any info on a particular topic, more information on a particular topic, Online Activity (shopping, downloading, gaming), Specific website |
| Information medium | Internet Search, YouTube, Newspaper, TV and Radio channels, Social Media, Word of Mouth |
| Share Info on Online Social Network | Educational Information, Disaster, Emergency, and other Breaking News, Entertainment Content, Finance information, General News, Gossip, Conspiracy, and History Information, Interviews and Videos, Job postings, Pray and Wellness Mantras, Medicinal remedies, Religious Content, Science and Technology info, None |
| Information Share Medium | Social media, blogs, private social groups, private messages, don’t share |
| Information credibility criteria | Article has high likes, Date provided, Article well written, Author profile is provided, Author is well known, Publisher is well known, references are provided, Agrees on Content, Website is considered credible |
| Medium for verifying Information | Ask an Expert, Ask on forums (e.g. Reddit, Quora), Ask a friend, Lookup on information repository, Lookup scholarly Articles, Search fact-checking websites, Search using Search Engines, Undisclosed. |
| Trusted online Source Criteria | Complete information provided, content has clarity, domain extension, Recommended by Friend, about and contact details provided, Website has good UX, good content, high likes or reviews, dated recent, in top search results, None. |
| Case1 | Confirm from Friend, forward to Friend, ignore post, post on social media, Verify. |
| Case 2 | Discard, Investigate, Share. |

1. **Data Preprocessing and Feature selection**

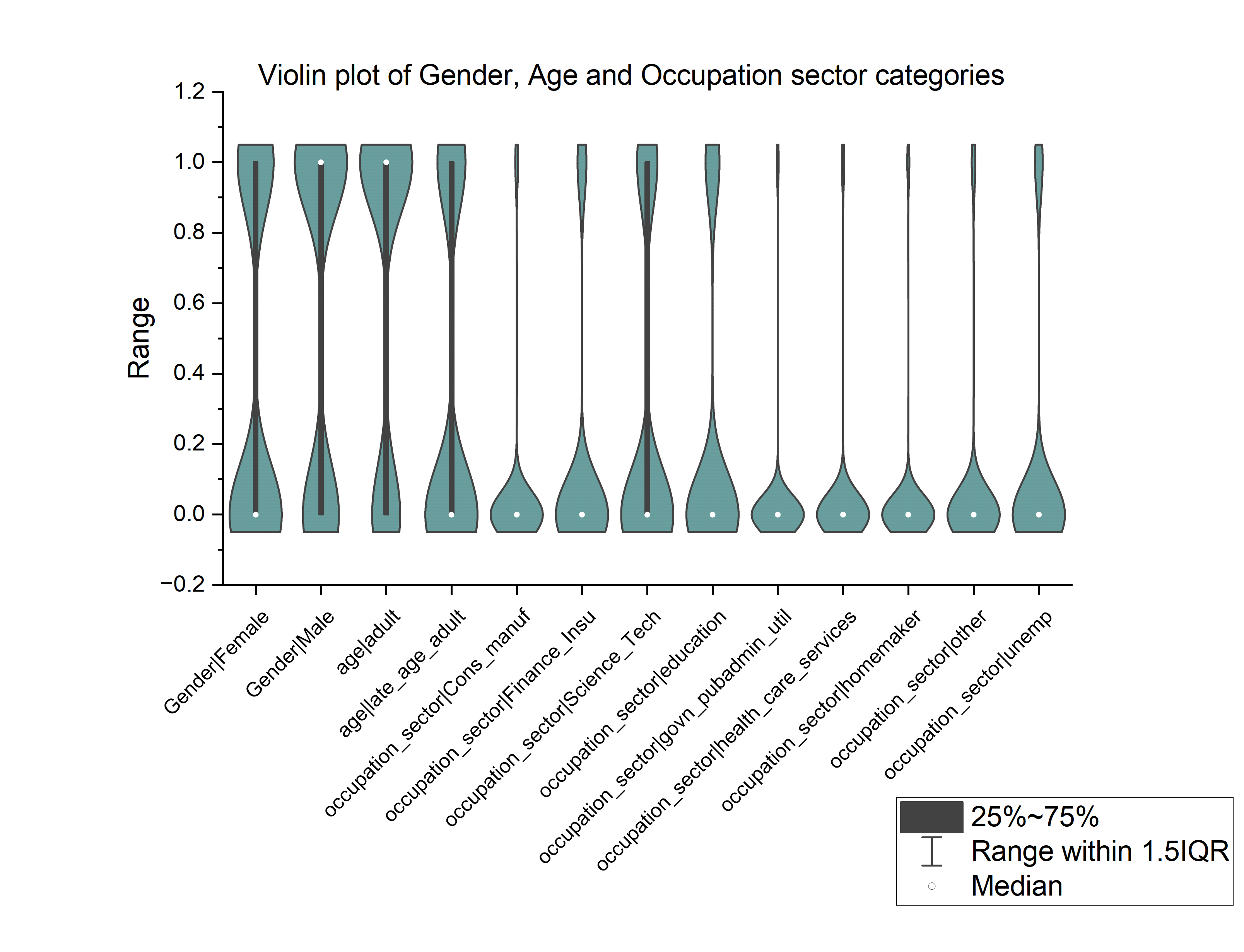


Fig SB- . Demographic Features

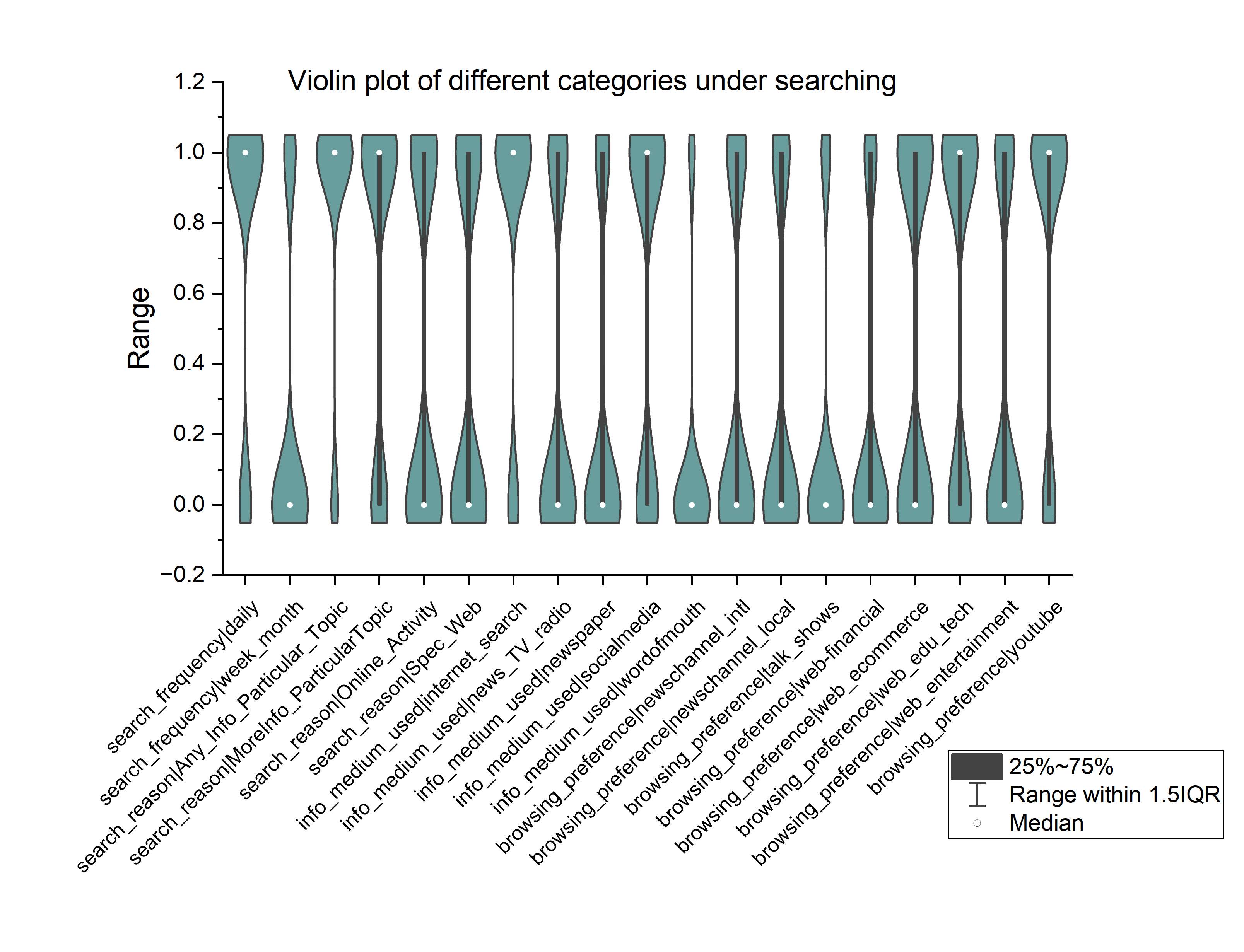


Fig SB- . Information Searching Features.

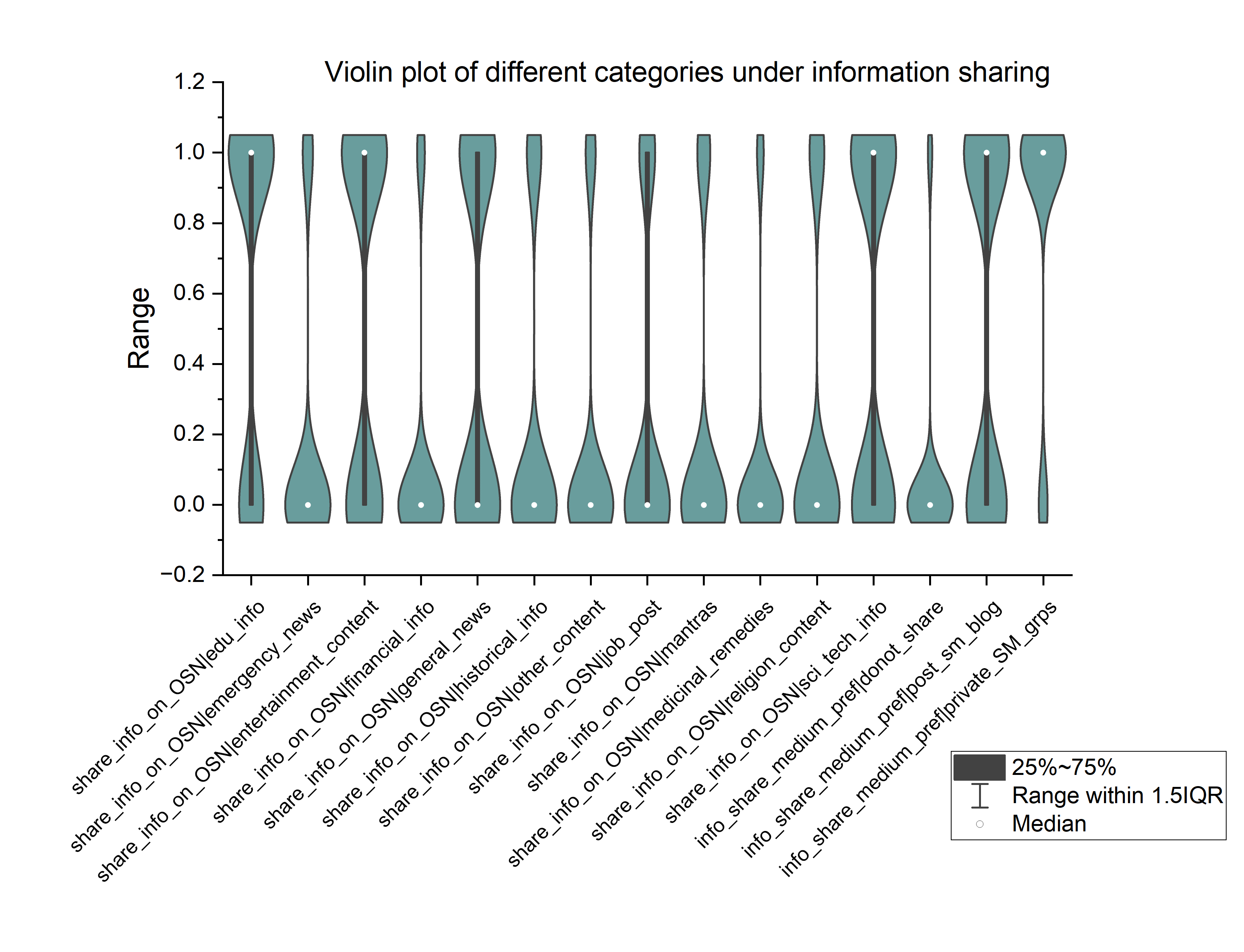


Fig SB- . Information Share Features.

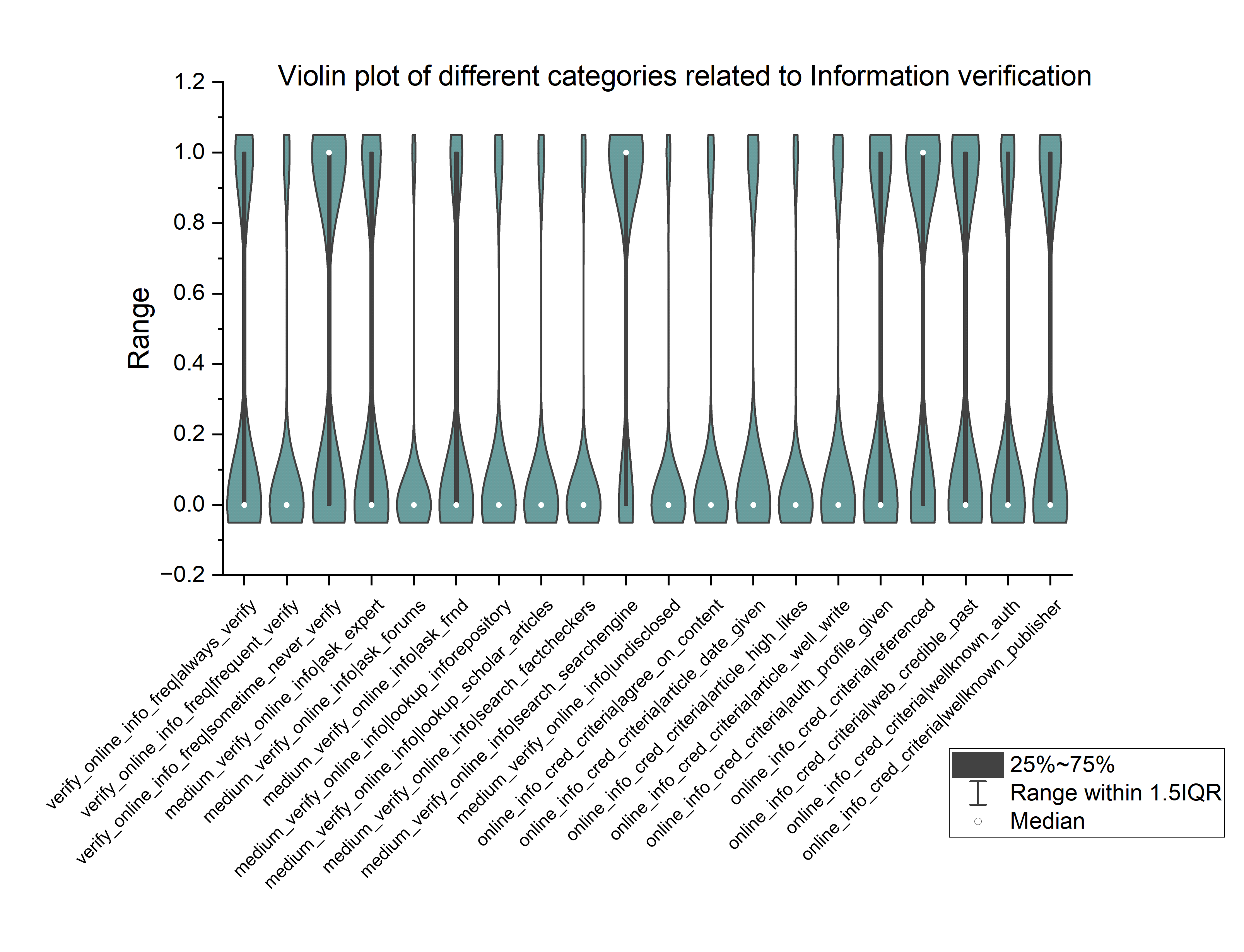


Fig SB- . Information Verification Features.

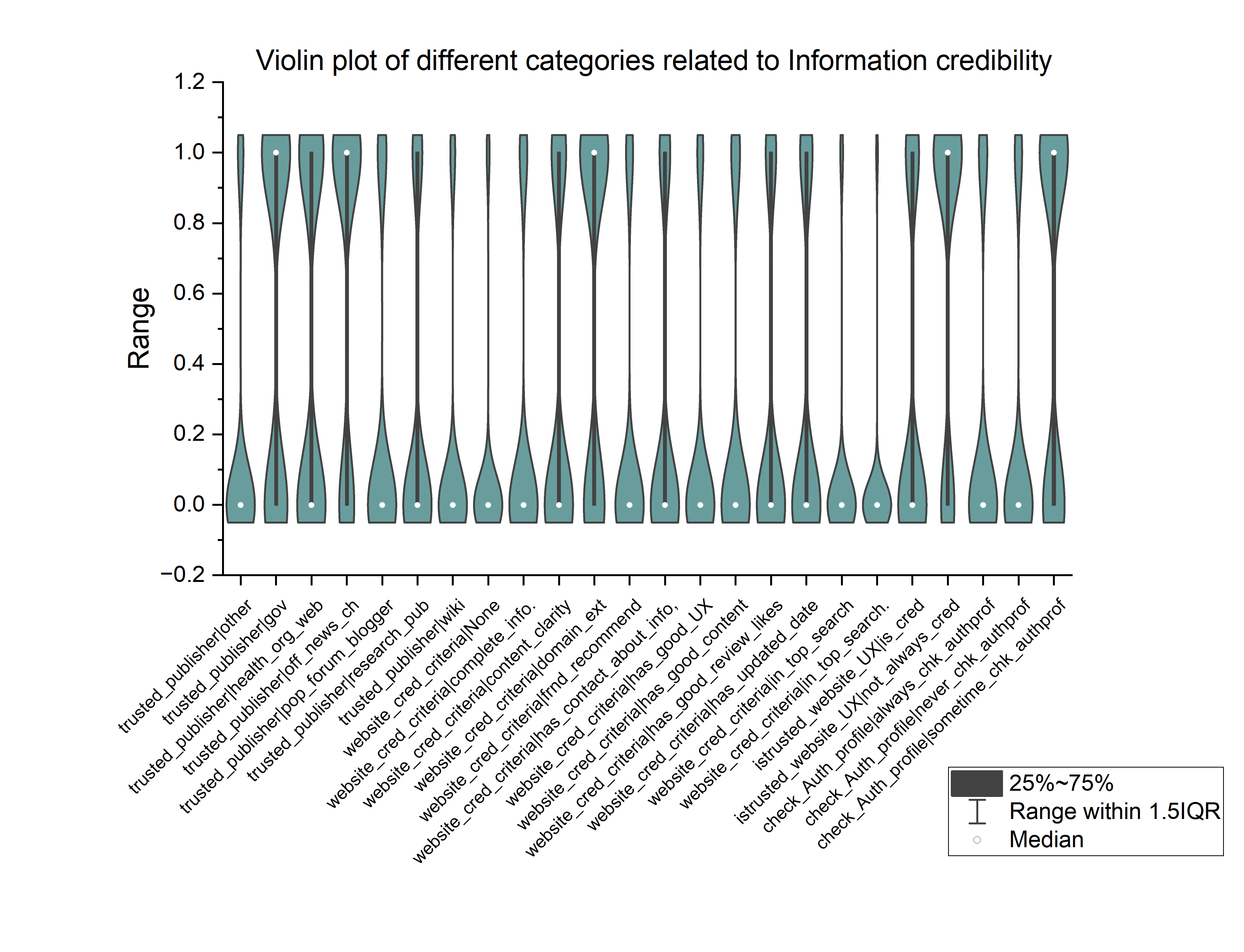


Fig SB- V. Information Credibility Criteria Features

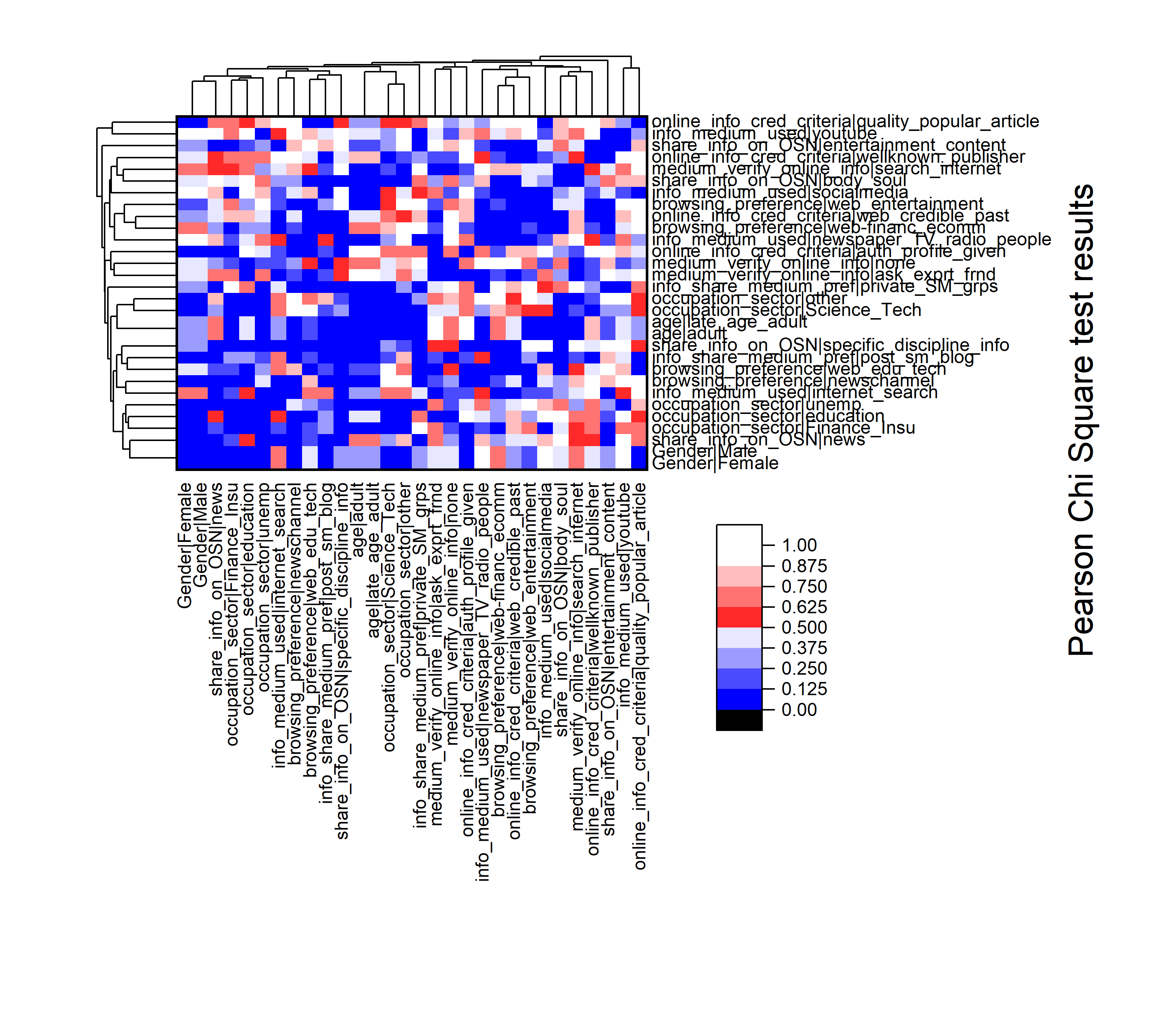


Fig SB- . Chi-Square Test results on adjusted α

Table S: p-value Adjusted for Features under Main Variables.

| Main Variables | αc | αt |
| --- | --- | --- |
| Gender | 0.025 | 0.15 |
| Age | 0.025 | 0.15 |
| Occupation | 0.01 | 0.06 |
| Information Medium Used | 0.0125 | 0.025 |
| Browsing Preference | 0.0125 | 0.025 |
| Share Information on OSN | 0.0125 | 0.025 |
| Information sharing medium | 0.025 | 0.15 |
| Information verification medium | 0.025 | 0.15 |
| Information verification Criteria | 0.016 | 0.033 |

1. **Clustering Unsupervised Learning Results**

**Table SIV:** Evaluation of Clustering Models

|  |  |  |
| --- | --- | --- |
| **Clustering Techniques** | **Silhouette Coefficient** | **No of Clusters (k)** |
| K-Mean | 0.34 | 5 |
|  | 0.25 | 4 |
| Hierarchal Clustering | 0.30 | 5 |
|  | 0.21 | 4 |
| Spectral Clustering | 0.32 | 5 |
|  | 0.24 | 4 |

1. **Average Participation of User Intent Class Features**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table S: Mean Participation of Demographic Features in User Intent Class (UIC)   |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | **Average of Demographic Features** | | | | | | | | | | | **Cluster/UIC** | **Female** | **Male** | **Adult** | **Late Age Adult** | **Finance\_Insu** | **Science\_Tech** | **Edu** | **Other** | **Uunemp** | | UIC0 | 36% | 64% | 87% | 13% | 0% | 92% | 7% | 2% | 0% | | UIC1 | 87% | 13% | 68% | 32% | 0% | 4% | 91% | 0% | 4% | | UIC2 | 0% | 100% | 50% | 50% | 100% | 0% | 0% | 0% | 0% | | UIC3 | 31% | 69% | 46% | 54% | 0% | 5% | 0% | 92% | 3% | | UIC4 | 73% | 27% | 100% | 0% | 0% | 7% | 0% | 3% | 90% | | **Grand Total** | **48%** | **52%** | **72%** | **28%** | **12%** | **31%** | **23%** | **19%** | **15%** | |

Table S: Mean Participation of Search Openness Features in User Intent Class (UIC)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Average of Search Mediums and Browsing Preference Features** | | | | | | | | |
| **Cluster/UIC** | **Internet\_search medium** | **Newspaper\_TV\_radio\_people** | **Social Media** | **YouTube** | **News**  **Channel** | **Web-financ\_ecomm** | **Web\_edu\_tech** | **Web\_Entertainment** |
| UIC0 | 81.97% | 44% | 67% | 79% | 46% | 57% | 59% | 33% |
| UIC1 | 82.98% | 60% | 70% | 83% | 38% | 57% | 72% | 43% |
| UIC2 | 87.50% | 71% | 54% | 83% | 79% | 83% | 54% | 33% |
| UIC3 | 84.62% | 44% | 72% | 79% | 46% | 51% | 64% | 38% |
| UIC4 | 66.67% | 37% | 67% | 63% | 37% | 53% | 53% | 30% |
| **Grand Total** | **81.09%** | **50%** | **67%** | **78%** | **47%** | **59%** | **62%** | **36%** |

Table S: Mean Participation of Online Extravert Features in User Intent Class (UIC)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Average of Share Info on OSN, Share Medium Features and Cases** | | | | | | | |
| **Cluster/UIC** | **Specific\_discipline\_info** | **News** | **Entertainment\_content** | **Body\_soul** | **Post\_sm\_blog** | **Private\_SM\_grps** | **se\_case** |
| UIC0 | 85% | 48% | 62% | 30% | 49% | 93% | 10% |
| UIC1 | 70% | 40% | 68% | 40% | 70% | 85% | 10% |
| UIC2 | 100% | 63% | 42% | 58% | 46% | 88% | 11% |
| UIC3 | 90% | 54% | 56% | 49% | 56% | 97% | 8% |
| UIC4 | 80% | 43% | 47% | 37% | 63% | 83% | 13% |
| **Grand Total** | **84%** | **48%** | **58%** | **40%** | **57%** | **90%** | **10%** |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table S: Mean Participation of Information Consciousness Features in User Intent Class (UIC)   |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | **Average of Verification Mediums, Criteria and Cases Features** | | | | | | | | **Cluster/ UIC** | **ic\_case** | **web\_publisher\_credible** | **auth\_profile\_given** | **quality\_popular\_article** | **search\_internet\_sm** | **ask\_exprt\_frnd** | | UIC0 | 106% | 62% | 49% | 80% | 90% | 48% | | UIC1 | 96% | 72% | 47% | 87% | 85% | 66% | | UIC2 | 96% | 79% | 79% | 79% | 92% | 63% | | UIC3 | 109% | 67% | 56% | 74% | 79% | 56% | | UIC4 | 98% | 73% | 47% | 73% | 80% | 57% | | **Grand Total** | **102%** | **69%** | **53%** | **80%** | **86%** | **57%** | |

1. **User Intent Class Template**

Table S: Generalization Rules for UIC Template.

|  |  |
| --- | --- |
| **UIC Categories** | **Generalization** |
| SO Information Medium | Uses **General search (G) or Social Media (S)** as an information medium.  Data suggests Social Media alone is not used**. GS/S** generalized to **GS** |
| Uses **General Search (G) or Traditional (T)** as an information medium.  Data suggest that Traditional or General alone doesn’t exist. **GT/G/T** is genderized to **GT.** |
| SO Browsing Preference | Browsing Preference is **Navigational (N), Informational (I), and Transactional (T)**.  Data suggests that the transactional content alone cannot be considered. **NIT/NT** generalized to **NIT** |
| Browse websites for Transactional or Informational purposes. The data revealed that informational content related to finance or commerce could include transactional activities **IT/I/T** generalized to **IT**. |
| OE Share Content Type (SCT) | Shares on social platforms, **Particular (P), Causal (C), News (N) or Wellbeing (W)**. Data suggest, that Well-being content as Casual or a particular topic. **PCNW/PCN/CN** is generalized as **PCNW** |
| Shares on social platforms Particular (P), Causal (C), or Wellbeing (W).  Data suggests that well-being content as Casual or a particular topic. **CW/C/W/PCW/PC/PW** generalized as **PCW**. |
| Shares on social platforms, particularly (P) or News (N). The data shows that a News Article can be considered as a particular topic, **P/N/0/PN** generalized to **PN**. |
| IC Medium (ICM) | Uses **General search mediums (G)** to verify the content.  The users whose ICM cannot be decided can be treated as general. **G/0** generalized as **G**. |
| IC Criteria (ICC) | Uses **Content Quality (C)**or **Author (A)** as credibility criteria. The data revealed that most users have used Content with Author. **CA/A** generalized to **CA**. |
| Considers **Content Quality (C), Author Profile (A), or Publisher (P)** for verification. The data reveals that content quality is considered by many users. **CAP/AP** generalized to **CAP**. |