Zubin Pahuja (zpahuja2@illinois.edu)

# PUBLICATION VENUE SUGGESTION IN HETEROGENOUS GRAPHS

## 1. SGD CLASSIFIER FOR PREDICTING VENUE FROM TITLE FEATURES

Macro-F1 score on Validation set: 0.19998591387183903

Micro-F1 score on Validation set: 0.30419682269376414

Classification Report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| aaai | 0.05 | 0.15 | 0.08 | 528 |
| aamas | 0.37 | 0.28 | 0.32 | 349 |
| acc | 0.00 | 0.00 | 0.00 | 109 |
| acm_multimedia | 0.28 | 0.16 | 0.20 | 379 |
| acm_trans._graph. | 0.00 | 0.00 | 0.00 | 1 |
| amcis | 0.31 | 0.28 | 0.29 | 665 |
| amia | 0.57 | 0.42 | 0.49 | 121 |
| asp-dac | 0.36 | 0.04 | 0.07 | 319 |
| bioinformatics | 0.00 | 0.00 | 0.00 | 17 |
| cdc | 0.48 | 0.58 | 0.53 | 843 |
| chi | 0.32 | 0.22 | 0.26 | 418 |
| chi_extended_abstracts | 0.31 | 0.42 | 0.35 | 568 |
| cikm | 0.20 | 0.08 | 0.11 | 377 |
| cogsci | 0.48 | 0.56 | 0.51 | 434 |
| coling | 0.40 | 0.25 | 0.31 | 367 |
| commun._acm | 0.33 | 0.03 | 0.05 | 68 |
| compsac | 0.39 | 0.05 | 0.08 | 296 |
| comput._graph._forum | 0.00 | 0.00 | 0.00 | 0 |
| comput._j. | 0.00 | 0.00 | 0.00 | 19 |

| | | | | |
|---|---|---|---|---|
| computer_communications | 0.00 | 0.00 | 0.00 | 0 |
| computer_networks | 0.25 | 0.50 | 0.33 | 2 |
| corr | 0.00 | 0.00 | 0.00 | 0 |
| cvpr | 0.18 | 0.17 | 0.17 | 593 |
| dac | 0.23 | 0.25 | 0.24 | 446 |
| date | 0.25 | 0.15 | 0.19 | 490 |
| ecai | 0.03 | 0.01 | 0.02 | 187 |
| ecis | 0.32 | 0.17 | 0.22 | 385 |
| embc | 0.41 | 0.28 | 0.33 | 342 |
| encyclopedia_of_database_systems | 0.29 | 0.82 | 0.42 | 301 |
| etfa | 0.36 | 0.23 | 0.28 | 255 |
| eurospeech | 0.28 | 0.04 | 0.07 | 318 |
| eusipco | 0.17 | 0.14 | 0.15 | 852 |
| expert_syst._appl. | 0.00 | 0.00 | 0.00 | 0 |
| focs | 0.36 | 0.16 | 0.22 | 259 |
| fskd | 0.28 | 0.04 | 0.07 | 289 |
| fundam._inform. | 0.00 | 0.00 | 0.00 | 1 |
| fusion | 0.42 | 0.40 | 0.41 | 286 |
| fuzz-ieee | 0.51 | 0.66 | 0.58 | 490 |
| gecco | 0.45 | 0.34 | 0.39 | 385 |
| globecom | 0.19 | 0.15 | 0.17 | 1323 |
| hicss | 0.48 | 0.27 | 0.35 | 981 |
| icalt | 0.56 | 0.47 | 0.51 | 338 |
| icarcv | 0.00 | 0.00 | 0.00 | 269 |
| icassp | 0.19 | 0.27 | 0.23 | 1662 |
| icc | 0.15 | 0.20 | 0.17 | 1380 |
| iccad | 0.23 | 0.10 | 0.14 | 302 |
| iccs | 0.32 | 0.20 | 0.25 | 211 |
| iccv | 0.12 | 0.09 | 0.10 | 272 |
| icdar | 0.56 | 0.68 | 0.62 | 284 |
| icde | 0.21 | 0.20 | 0.21 | 343 |

| | | | | |
|---|---|---|---|---|
| icecs | 0.17 | 0.03 | 0.05 | 293 |
| icip | 0.26 | 0.32 | 0.29 | 947 |
| icis | 0.27 | 0.04 | 0.07 | 407 |
| icmc | 0.54 | 0.73 | 0.62 | 435 |
| icme | 0.20 | 0.11 | 0.14 | 540 |
| icml | 0.16 | 0.12 | 0.14 | 264 |
| icnc | 0.26 | 0.05 | 0.08 | 210 |
| icpr | 0.12 | 0.02 | 0.04 | 456 |
| icra | 0.39 | 0.54 | 0.45 | 1793 |
| icse | 0.36 | 0.33 | 0.34 | 324 |
| icslp | 0.23 | 0.03 | 0.06 | 283 |
| ieee_computer | 0.00 | 0.00 | 0.00 | 18 |
| ieee_congress_on_evolutionary_computation | 0.41 | 0.37 | 0.39 | 396 |
| ieee_journal_on_selected_areas_in_communications | 0.00 | 0.00 | 0.00 | 1 |
| ieee_software | 0.00 | 0.00 | 0.00 | 4 |
| ieee_trans._computers | 0.00 | 0.00 | 0.00 | 19 |
| ieee_trans._information_theory | 0.00 | 0.00 | 0.00 | 0 |
| ieee_trans._knowl._data_eng. | 0.00 | 0.00 | 0.00 | 42 |
| ieee_trans._parallel_distrib._syst. | 0.00 | 0.00 | 0.00 | 1 |
| ieee_trans._pattern_anal._mach._intell. | 0.00 | 0.00 | 0.00 | 0 |
| ieee_trans._software_eng. | 0.05 | 0.02 | 0.03 | 46 |
| igarss | 0.72 | 0.87 | 0.79 | 1325 |
| ijcai | 0.09 | 0.26 | 0.13 | 752 |
| ijcnn | 0.35 | 0.26 | 0.30 | 543 |
| inf._process._lett. | 0.17 | 0.05 | 0.07 | 21 |
| inf._sci. | 0.00 | 0.00 | 0.00 | 0 |
| infocom | 0.27 | 0.26 | 0.26 | 690 |
| int._cmg_conference | 0.62 | 0.64 | 0.63 | 358 |
| interspeech | 0.46 | 0.71 | 0.56 | 1252 |
| ipdps | 0.32 | 0.30 | 0.31 | 449 |
| iros | 0.32 | 0.23 | 0.27 | 1399 |

| | | | | |
|---|---|---|---|---|
| isbi | 0.46 | 0.63 | 0.53 | 417 |
| iscas | 0.38 | 0.42 | 0.40 | 889 |
| iscc | 0.09 | 0.03 | 0.05 | 323 |
| isit | 0.43 | 0.56 | 0.49 | 460 |
| itc | 0.55 | 0.72 | 0.63 | 452 |
| j._acm | 0.00 | 0.00 | 0.00 | 32 |
| j._parallel_distrib._comput. | 0.00 | 0.00 | 0.00 | 1 |
| j._symb._log. | 0.00 | 0.00 | 0.00 | 1 |
| journal_of_systems_and_software | 0.00 | 0.00 | 0.00 | 0 |
| kdd | 0.18 | 0.14 | 0.16 | 239 |
| lcn | 0.26 | 0.06 | 0.10 | 286 |
| lrec | 0.52 | 0.61 | 0.56 | 461 |
| multimedia_tools_appl. | 0.00 | 0.00 | 0.00 | 1 |
| neuroimage | 0.00 | 0.00 | 0.00 | 0 |
| nips | 0.17 | 0.31 | 0.22 | 588 |
| pacis | 0.42 | 0.04 | 0.07 | 292 |
| pattern_recognition | 0.00 | 0.00 | 0.00 | 9 |
| pdpta | 0.33 | 0.09 | 0.15 | 301 |
| pimrc | 0.24 | 0.14 | 0.18 | 898 |
| robio | 0.32 | 0.09 | 0.14 | 404 |
| sac | 0.10 | 0.12 | 0.11 | 577 |
| siam_j._comput. | 0.00 | 0.00 | 0.00 | 14 |
| sigcse | 0.66 | 0.70 | 0.68 | 392 |
| sigir | 0.37 | 0.26 | 0.31 | 351 |
| sigmod_conference | 0.18 | 0.14 | 0.16 | 299 |
| smc | 0.17 | 0.09 | 0.12 | 871 |
| soda | 0.34 | 0.29 | 0.31 | 292 |
| softw._pract._exper. | 0.00 | 0.00 | 0.00 | 11 |
| stoc | 0.36 | 0.21 | 0.26 | 311 |
| theor._comput._sci. | 0.11 | 0.07 | 0.08 | 15 |
| vlsi_design | 0.22 | 0.06 | 0.09 | 271 |

| | | | | |
|---|---|---|---|---|
| vtc_fall | 0.19 | 0.03 | 0.05 | 521 |
| vtc_spring | 0.21 | 0.09 | 0.13 | 604 |
| wcnc | 0.14 | 0.09 | 0.11 | 781 |
| winter_simulation_conference | 0.54 | 0.72 | 0.62 | 873 |
| **avg / total** | **0.31** | **0.30** | **0.29** | **44629** |

## 2. USING HIN FEATURES TO SUPPLEMENT TEXT FEATURES

Macro-F1 score on Validation set: 0.7620302763937719

Micro-F1 score on Validation set: 0.9824329471868067

Classification Report:

| | **precision** | **recall** | **f1-score** | **support** |
|---|---|---|---|---|
| aaai | 1.00 | 0.93 | 0.96 | 528 |
| aamas | 1.00 | 1.00 | 1.00 | 349 |
| acc | 1.00 | 1.00 | 1.00 | 109 |
| acm_multimedia | 1.00 | 1.00 | 1.00 | 379 |
| acm_trans._graph. | 0.00 | 0.00 | 0.00 | 1 |
| amcis | 1.00 | 1.00 | 1.00 | 665 |
| amia | 1.00 | 1.00 | 1.00 | 121 |
| asp-dac | 1.00 | 1.00 | 1.00 | 319 |
| bioinformatics | 1.00 | 1.00 | 1.00 | 17 |
| cdc | 1.00 | 1.00 | 1.00 | 843 |
| chi | 1.00 | 1.00 | 1.00 | 418 |
| chi_extended_abstracts | 1.00 | 1.00 | 1.00 | 568 |
| cikm | 0.92 | 0.86 | 0.89 | 377 |
| cogsci | 1.00 | 1.00 | 1.00 | 434 |
| coling | 1.00 | 1.00 | 1.00 | 367 |
| commun._acm | 0.43 | 0.09 | 0.15 | 68 |
| compsac | 1.00 | 1.00 | 1.00 | 296 |
| comput._graph._forum | 0.00 | 0.00 | 0.00 | 0 |
| comput._j. | 0.33 | 0.05 | 0.09 | 19 |

| | | | | |
|---|---|---|---|---|
| computer_communications | 0.00 | 0.00 | 0.00 | 0 |
| computer_networks | 0.33 | 0.50 | 0.40 | 2 |
| corr | 0.00 | 0.00 | 0.00 | 0 |
| cvpr | 0.99 | 0.89 | 0.94 | 593 |
| dac | 1.00 | 1.00 | 1.00 | 446 |
| date | 1.00 | 1.00 | 1.00 | 490 |
| ecai | 1.00 | 0.98 | 0.99 | 187 |
| ecis | 1.00 | 1.00 | 1.00 | 385 |
| embc | 1.00 | 1.00 | 1.00 | 342 |
| encyclopedia_of_database_systems | 0.97 | 1.00 | 0.99 | 301 |
| etfa | 1.00 | 1.00 | 1.00 | 255 |
| eurospeech | 1.00 | 1.00 | 1.00 | 318 |
| eusipco | 1.00 | 1.00 | 1.00 | 852 |
| expert_syst._appl. | 0.00 | 0.00 | 0.00 | 0 |
| focs | 1.00 | 1.00 | 1.00 | 259 |
| fskd | 1.00 | 1.00 | 1.00 | 289 |
| fundam._inform. | 0.00 | 0.00 | 0.00 | 1 |
| fusion | 0.99 | 1.00 | 1.00 | 286 |
| fuzz-ieee | 1.00 | 1.00 | 1.00 | 490 |
| gecco | 1.00 | 1.00 | 1.00 | 385 |
| globecom | 1.00 | 1.00 | 1.00 | 1323 |
| hicss | 1.00 | 1.00 | 1.00 | 981 |
| icalt | 1.00 | 1.00 | 1.00 | 338 |
| icarcv | 1.00 | 1.00 | 1.00 | 269 |
| icassp | 1.00 | 1.00 | 1.00 | 1662 |
| icc | 1.00 | 1.00 | 1.00 | 1380 |
| iccad | 1.00 | 1.00 | 1.00 | 302 |
| iccs | 1.00 | 1.00 | 1.00 | 211 |
| iccv | 0.85 | 0.79 | 0.82 | 272 |
| icdar | 1.00 | 1.00 | 1.00 | 284 |
| icde | 0.77 | 0.74 | 0.76 | 343 |

| | | | | |
|---|---|---|---|---|
| icecs | 1.00 | 1.00 | 1.00 | 293 |
| icip | 1.00 | 1.00 | 1.00 | 947 |
| icis | 1.00 | 1.00 | 1.00 | 407 |
| icmc | 1.00 | 1.00 | 1.00 | 435 |
| icme | 1.00 | 1.00 | 1.00 | 540 |
| icml | 1.00 | 0.94 | 0.97 | 264 |
| icnc | 1.00 | 1.00 | 1.00 | 210 |
| icpr | 1.00 | 1.00 | 1.00 | 456 |
| icra | 1.00 | 1.00 | 1.00 | 1793 |
| icse | 1.00 | 1.00 | 1.00 | 324 |
| icslp | 1.00 | 1.00 | 1.00 | 283 |
| ieee_computer | 0.00 | 0.00 | 0.00 | 18 |
| ieee_congress_on_evolutionary_computation | 1.00 | 1.00 | 1.00 | 396 |
| ieee_journal_on_selected_areas_in_communications | 0.00 | 0.00 | 0.00 | 1 |
| ieee_software | 0.00 | 0.00 | 0.00 | 4 |
| ieee_trans._computers | 0.00 | 0.00 | 0.00 | 19 |
| ieee_trans._information_theory | 0.00 | 0.00 | 0.00 | 0 |
| ieee_trans._knowl._data_eng. | 0.36 | 0.10 | 0.15 | 42 |
| ieee_trans._parallel_distrib._syst. | 0.00 | 0.00 | 0.00 | 1 |
| ieee_trans._pattern_anal._mach._intell. | 0.43 | 0.13 | 0.20 | 46 |
| ieee_trans._software_eng. | 1.00 | 1.00 | 1.00 | 1325 |
| igarss | 1.00 | 1.00 | 1.00 | 752 |
| ijcai | 1.00 | 1.00 | 1.00 | 543 |
| ijcnn | 0.14 | 0.05 | 0.07 | 21 |
| inf._process._lett. | 0.00 | 0.00 | 0.00 | 0 |
| inf._sci. | 1.00 | 1.00 | 1.00 | 690 |
| infocom | 1.00 | 1.00 | 1.00 | 358 |
| int._cmg_conference | 1.00 | 1.00 | 1.00 | 1252 |
| interspeech | 1.00 | 1.00 | 1.00 | 449 |
| ipdps | 1.00 | 1.00 | 1.00 | 1399 |
| iros | 1.00 | 1.00 | 1.00 | 417 |

| | | | | |
|---|---|---|---|---|
| isbi | 0.78 | 0.99 | 0.87 | 889 |
| iscas | 1.00 | 1.00 | 1.00 | 323 |
| iscc | 1.00 | 1.00 | 1.00 | 460 |
| isit | 0.96 | 0.92 | 0.94 | 452 |
| itc | 0.11 | 0.12 | 0.11 | 32 |
| j._acm | 0.00 | 0.00 | 0.00 | 1 |
| j._parallel_distrib._comput. | 0.00 | 0.00 | 0.00 | 1 |
| j._symb._log. | 0.00 | 0.00 | 0.00 | 0 |
| journal_of_systems_and_software | 1.00 | 1.00 | 1.00 | 239 |
| kdd | 1.00 | 1.00 | 1.00 | 286 |
| lcn | 1.00 | 1.00 | 1.00 | 461 |
| lrec | 0.03 | 1.00 | 0.05 | 1 |
| multimedia_tools_appl. | 1.00 | 1.00 | 1.00 | 588 |
| neuroimage | 1.00 | 1.00 | 1.00 | 292 |
| nips | 0.56 | 1.00 | 0.72 | 9 |
| pacis | 1.00 | 1.00 | 1.00 | 301 |
| pattern_recognition | 1.00 | 1.00 | 1.00 | 898 |
| pdpta | 1.00 | 1.00 | 1.00 | 404 |
| pimrc | 1.00 | 1.00 | 1.00 | 577 |
| robio | 0.00 | 0.00 | 0.00 | 14 |
| sac | 1.00 | 1.00 | 1.00 | 392 |
| siam_j._comput. | 0.92 | 0.89 | 0.91 | 351 |
| sigcse | 0.79 | 0.78 | 0.79 | 299 |
| sigir | 1.00 | 1.00 | 1.00 | 871 |
| sigmod_conference | 1.00 | 1.00 | 1.00 | 292 |
| smc | 0.33 | 0.09 | 0.14 | 11 |
| soda | 1.00 | 0.99 | 1.00 | 311 |
| softw._pract._exper. | 0.00 | 0.00 | 0.00 | 15 |
| stoc | 1.00 | 1.00 | 1.00 | 271 |
| theor._comput._sci. | 1.00 | 1.00 | 1.00 | 521 |
| vlsi_design | 1.00 | 1.00 | 1.00 | 604 |

| | | | | |
|---|---|---|---|---|
| vtc_fall | 1.00 | 1.00 | 1.00 | 781 |
| vtc_spring | 1.00 | 1.00 | 1.00 | 873 |
| **avg / total** | **0.98** | **0.98** | **0.98** | **44629** |

## 3. USING WORD2VEC EMBEDDINGS AS TEXT FEATURES

Macro-F1 score on Validation set: 0.11890360989735102

Micro-F1 score on Validation set: 0.22043962445943222

Classification Report:

| | **precision** | **recall** | **f1-score** | **support** |
|---|---|---|---|---|
| aaai | 0.11 | 0.13 | 0.12 | 528 |
| aamas | 0.24 | 0.06 | 0.09 | 349 |
| acc | 0.00 | 0.00 | 0.00 | 109 |
| acm_multimedia | 0.14 | 0.08 | 0.10 | 379 |
| acm_trans._graph. | 0.00 | 0.00 | 0.00 | 1 |
| amcis | 0.10 | 0.03 | 0.05 | 665 |
| amia | 0.33 | 0.03 | 0.06 | 121 |
| asp-dac | 0.20 | 0.01 | 0.01 | 319 |
| bioinformatics | 0.00 | 0.00 | 0.00 | 17 |
| cdc | 0.34 | 0.56 | 0.43 | 843 |
| chi | 0.13 | 0.09 | 0.10 | 418 |
| chi_extended_abstracts | 0.15 | 0.17 | 0.16 | 568 |
| cikm | 0.09 | 0.20 | 0.12 | 377 |
| cogsci | 0.49 | 0.10 | 0.16 | 434 |
| coling | 0.16 | 0.16 | 0.16 | 367 |
| commun._acm | 0.08 | 0.01 | 0.02 | 68 |
| compsac | 0.10 | 0.01 | 0.02 | 296 |
| comput._graph._forum | 0.00 | 0.00 | 0.00 | 19 |
| comput._j. | 0.00 | 0.00 | 0.00 | 2 |
| computer_communications | 0.15 | 0.05 | 0.08 | 593 |
| computer_networks | 0.15 | 0.18 | 0.17 | 446 |

| | | | | |
|---|---|---|---|---|
| corr | 0.17 | 0.01 | 0.02 | 490 |
| cvpr | 0.01 | 0.01 | 0.01 | 187 |
| dac | 0.10 | 0.56 | 0.17 | 385 |
| date | 0.24 | 0.05 | 0.08 | 342 |
| ecai | 0.38 | 0.45 | 0.41 | 301 |
| ecis | 0.00 | 0.00 | 0.00 | 255 |
| embc | 0.00 | 0.00 | 0.00 | 318 |
| encyclopedia_of_database_systems | 0.15 | 0.03 | 0.06 | 852 |
| etfa | 0.40 | 0.01 | 0.02 | 259 |
| eurospeech | 0.04 | 0.03 | 0.03 | 289 |
| eusipco | 0.00 | 0.00 | 0.00 | 1 |
| expert_syst._appl. | 0.59 | 0.07 | 0.13 | 286 |
| focs | 0.51 | 0.60 | 0.55 | 490 |
| fskd | 0.23 | 0.35 | 0.28 | 385 |
| fundam._inform. | 0.15 | 0.28 | 0.20 | 1323 |
| fusion | 0.33 | 0.14 | 0.20 | 981 |
| fuzz-ieee | 0.15 | 0.44 | 0.22 | 338 |
| gecco | 0.00 | 0.00 | 0.00 | 269 |
| globecom | 0.10 | 0.29 | 0.15 | 1662 |
| hicss | 0.12 | 0.01 | 0.02 | 1380 |
| icalt | 0.20 | 0.05 | 0.08 | 302 |
| icarcv | 0.03 | 0.11 | 0.04 | 211 |
| icassp | 0.01 | 0.00 | 0.01 | 272 |
| icc | 0.60 | 0.25 | 0.35 | 284 |
| iccad | 0.12 | 0.01 | 0.02 | 343 |
| iccs | 0.03 | 0.01 | 0.01 | 293 |
| iccv | 0.19 | 0.24 | 0.21 | 947 |
| icdar | 0.09 | 0.04 | 0.06 | 407 |
| icde | 0.55 | 0.52 | 0.54 | 435 |
| icecs | 0.03 | 0.00 | 0.01 | 540 |
| icip | 0.00 | 0.00 | 0.00 | 264 |

| | | | | |
|---|---|---|---|---|
| icis | 0.19 | 0.02 | 0.04 | 210 |
| icmc | 0.07 | 0.02 | 0.03 | 456 |
| icme | 0.35 | 0.62 | 0.45 | 1793 |
| icml | 0.32 | 0.04 | 0.08 | 324 |
| icnc | 0.04 | 0.00 | 0.01 | 283 |
| icpr | 0.00 | 0.00 | 0.00 | 18 |
| icra | 0.40 | 0.06 | 0.10 | 396 |
| icse | 0.00 | 0.00 | 0.00 | 1 |
| icslp | 0.00 | 0.00 | 0.00 | 4 |
| ieee_computer | 0.00 | 0.00 | 0.00 | 19 |
| ieee_congress_on_evolutionary_computation | 0.09 | 0.12 | 0.10 | 42 |
| ieee_journal_on_selected_areas_in_communications | 0.00 | 0.00 | 0.00 | 1 |
| ieee_software | 0.00 | 0.00 | 0.00 | 46 |
| ieee_trans._computers | 0.56 | 0.82 | 0.66 | 1325 |
| ieee_trans._information_theory | 0.06 | 0.03 | 0.04 | 752 |
| ieee_trans._knowl._data_eng. | 0.20 | 0.06 | 0.10 | 543 |
| ieee_trans._parallel_distrib._syst. | 0.00 | 0.00 | 0.00 | 21 |
| ieee_trans._pattern_anal._mach._intell. | 0.19 | 0.07 | 0.10 | 690 |
| ieee_trans._software_eng. | 0.28 | 0.57 | 0.37 | 358 |
| igarss | 0.34 | 0.72 | 0.46 | 1252 |
| ijcai | 0.10 | 0.16 | 0.12 | 449 |
| ijcnn | 0.18 | 0.07 | 0.10 | 1399 |
| inf._process._lett. | 0.22 | 0.59 | 0.32 | 417 |
| inf._sci. | 0.33 | 0.12 | 0.17 | 889 |
| infocom | 0.00 | 0.00 | 0.00 | 323 |
| int._cmg_conference | 0.36 | 0.40 | 0.38 | 460 |
| interspeech | 0.59 | 0.44 | 0.50 | 452 |
| ipdps | 0.00 | 0.00 | 0.00 | 32 |
| iros | 0.00 | 0.00 | 0.00 | 1 |
| isbi | 0.00 | 0.00 | 0.00 | 1 |
| iscas | 0.04 | 0.01 | 0.02 | 239 |

| | | | | |
|---|---|---|---|---|
| iscc | 0.00 | 0.00 | 0.00 | 286 |
| isit | 0.44 | 0.50 | 0.47 | 461 |
| itc | 0.00 | 0.00 | 0.00 | 1 |
| j._acm | 0.20 | 0.02 | 0.04 | 588 |
| j._parallel_distrib._comput. | 0.33 | 0.00 | 0.01 | 292 |
| j._symb._log. | 0.00 | 0.00 | 0.00 | 9 |
| journal_of_systems_and_software | 0.05 | 0.29 | 0.08 | 301 |
| kdd | 0.14 | 0.32 | 0.20 | 898 |
| lcn | 0.06 | 0.01 | 0.02 | 404 |
| lrec | 0.06 | 0.07 | 0.06 | 577 |
| multimedia_tools_appl. | 0.00 | 0.00 | 0.00 | 14 |
| neuroimage | 0.60 | 0.58 | 0.59 | 392 |
| nips | 0.20 | 0.27 | 0.23 | 351 |
| pacis | 0.36 | 0.02 | 0.03 | 299 |
| pattern_recognition | 0.05 | 0.00 | 0.00 | 871 |
| pdpta | 0.11 | 0.46 | 0.17 | 292 |
| pimrc | 0.00 | 0.00 | 0.00 | 11 |
| robio | 0.23 | 0.21 | 0.22 | 311 |
| sac | 0.00 | 0.00 | 0.00 | 15 |
| siam_j._comput. | 0.14 | 0.04 | 0.06 | 271 |
| sigcse | 0.06 | 0.01 | 0.02 | 521 |
| sigir | 0.11 | 0.00 | 0.01 | 604 |
| sigmod_conference | 0.12 | 0.08 | 0.10 | 781 |
| smc | 0.51 | 0.63 | 0.56 | 873 |
| | | | | |
| **avg / total** | **0.21** | **0.22** | **0.18** | **44629** |

## 4. USING WORD2VEC EMBEDDINGS WITH HIN FEATURES

Macro-F1 score on Validation set: 0.7887026160652078

Micro-F1 score on Validation set: 0.9783772883102915

Classification Report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| aaai | 0.83 | 0.94 | 0.88 | 528 |
| aamas | 1.00 | 1.00 | 1.00 | 349 |
| acc | 1.00 | 1.00 | 1.00 | 109 |
| acm_multimedia | 1.00 | 1.00 | 1.00 | 379 |
| acm_trans._graph. | 0.00 | 0.00 | 0.00 | 1 |
| amcis | 1.00 | 1.00 | 1.00 | 665 |
| amia | 1.00 | 1.00 | 1.00 | 121 |
| asp-dac | 1.00 | 1.00 | 1.00 | 319 |
| bioinformatics | 1.00 | 1.00 | 1.00 | 17 |
| cdc | 1.00 | 1.00 | 1.00 | 843 |
| chi | 1.00 | 1.00 | 1.00 | 418 |
| chi_extended_abstracts | 1.00 | 1.00 | 1.00 | 568 |
| cikm | 0.72 | 0.86 | 0.79 | 377 |
| cogsci | 1.00 | 1.00 | 1.00 | 434 |
| coling | 0.99 | 0.96 | 0.98 | 367 |
| commun._acm | 0.00 | 0.00 | 0.00 | 68 |
| compsac | 1.00 | 1.00 | 1.00 | 296 |
| comput._graph._forum | 0.00 | 0.00 | 0.00 | 19 |
| comput._j. | 0.00 | 0.00 | 0.00 | 2 |
| computer_communications | 0.99 | 0.89 | 0.93 | 593 |
| computer_networks | 1.00 | 1.00 | 1.00 | 446 |
| corr | 1.00 | 1.00 | 1.00 | 490 |
| cvpr | 1.00 | 0.97 | 0.99 | 187 |
| dac | 1.00 | 1.00 | 1.00 | 385 |
| date | 1.00 | 1.00 | 1.00 | 342 |
| ecai | 1.00 | 1.00 | 1.00 | 301 |
| ecis | 1.00 | 1.00 | 1.00 | 255 |

| | | | | |
|---|---|---|---|---|
| embc | 1.00 | 1.00 | 1.00 | 318 |
| encyclopedia_of_database_systems | 1.00 | 0.94 | 0.97 | 852 |
| etfa | 1.00 | 1.00 | 1.00 | 259 |
| eurospeech | 1.00 | 1.00 | 1.00 | 289 |
| eusipco | 0.00 | 0.00 | 0.00 | 1 |
| expert_syst._appl. | 1.00 | 1.00 | 1.00 | 286 |
| focs | 1.00 | 1.00 | 1.00 | 490 |
| fskd | 1.00 | 1.00 | 1.00 | 385 |
| fundam._inform. | 1.00 | 1.00 | 1.00 | 1323 |
| fusion | 1.00 | 1.00 | 1.00 | 981 |
| fuzz-ieee | 1.00 | 1.00 | 1.00 | 338 |
| gecco | 1.00 | 1.00 | 1.00 | 269 |
| globecom | 1.00 | 1.00 | 1.00 | 1662 |
| hicss | 1.00 | 1.00 | 1.00 | 1380 |
| icalt | 1.00 | 1.00 | 1.00 | 302 |
| icarcv | 1.00 | 0.99 | 1.00 | 211 |
| icassp | 0.82 | 0.79 | 0.80 | 272 |
| icc | 0.99 | 0.93 | 0.95 | 284 |
| iccad | 0.65 | 0.73 | 0.69 | 343 |
| iccs | 1.00 | 1.00 | 1.00 | 293 |
| iccv | 1.00 | 1.00 | 1.00 | 947 |
| icdar | 1.00 | 1.00 | 1.00 | 407 |
| icde | 1.00 | 1.00 | 1.00 | 435 |
| icecs | 1.00 | 1.00 | 1.00 | 540 |
| icip | 1.00 | 0.91 | 0.95 | 264 |
| icis | 1.00 | 1.00 | 1.00 | 210 |
| icmc | 1.00 | 1.00 | 1.00 | 456 |
| icme | 1.00 | 1.00 | 1.00 | 1793 |
| icml | 1.00 | 1.00 | 1.00 | 324 |
| icnc | 1.00 | 1.00 | 1.00 | 283 |
| icpr | 0.00 | 0.00 | 0.00 | 18 |

| | | | | |
|---|---|---|---|---|
| icra | 1.00 | 1.00 | 1.00 | 396 |
| icse | 0.00 | 0.00 | 0.00 | 1 |
| icslp | 0.00 | 0.00 | 0.00 | 4 |
| ieee_computer | 0.00 | 0.00 | 0.00 | 19 |
| ieee_congress_on_evolutionary_computation | 0.57 | 0.10 | 0.16 | 42 |
| ieee_journal_on_selected_areas_in_communications | 0.00 | 0.00 | 0.00 | 1 |
| ieee_software | 0.00 | 0.00 | 0.00 | 46 |
| ieee_trans._computers | 1.00 | 1.00 | 1.00 | 1325 |
| ieee_trans._information_theory | 1.00 | 1.00 | 1.00 | 752 |
| ieee_trans._knowl._data_eng. | 1.00 | 1.00 | 1.00 | 543 |
| ieee_trans._parallel_distrib._syst. | 0.00 | 0.00 | 0.00 | 21 |
| ieee_trans._pattern_anal._mach._intell. | 1.00 | 1.00 | 1.00 | 690 |
| ieee_trans._software_eng. | 1.00 | 1.00 | 1.00 | 358 |
| igarss | 1.00 | 1.00 | 1.00 | 1252 |
| ijcai | 1.00 | 1.00 | 1.00 | 449 |
| ijcnn | 1.00 | 1.00 | 1.00 | 1399 |
| inf._process._lett. | 1.00 | 1.00 | 1.00 | 417 |
| inf._sci. | 0.74 | 0.97 | 0.84 | 889 |
| infocom | 1.00 | 1.00 | 1.00 | 323 |
| int._cmg_conference | 1.00 | 1.00 | 1.00 | 460 |
| interspeech | 0.92 | 0.87 | 0.89 | 452 |
| ipdps | 0.00 | 0.00 | 0.00 | 32 |
| iros | 0.00 | 0.00 | 0.00 | 1 |
| isbi | 0.00 | 0.00 | 0.00 | 1 |
| iscas | 1.00 | 0.99 | 1.00 | 239 |
| iscc | 1.00 | 0.99 | 1.00 | 286 |
| isit | 1.00 | 1.00 | 1.00 | 461 |
| itc | 0.00 | 0.00 | 0.00 | 1 |
| j._acm | 1.00 | 1.00 | 1.00 | 588 |
| j._parallel_distrib._comput. | 1.00 | 1.00 | 1.00 | 292 |
| j._symb._log. | 0.00 | 0.00 | 0.00 | 9 |

| | | | | |
|---|---|---|---|---|
| journal_of_systems_and_software | 1.00 | 1.00 | 1.00 | 301 |
| kdd | 1.00 | 1.00 | 1.00 | 898 |
| lcn | 1.00 | 1.00 | 1.00 | 404 |
| lrec | 1.00 | 1.00 | 1.00 | 577 |
| multimedia_tools_appl. | 0.00 | 0.00 | 0.00 | 14 |
| neuroimage | 1.00 | 1.00 | 1.00 | 392 |
| nips | 0.98 | 0.82 | 0.89 | 351 |
| pacis | 0.62 | 0.83 | 0.71 | 299 |
| pattern_recognition | 1.00 | 1.00 | 1.00 | 871 |
| pdpta | 1.00 | 1.00 | 1.00 | 292 |
| pimrc | 0.00 | 0.00 | 0.00 | 11 |
| robio | 1.00 | 1.00 | 1.00 | 311 |
| sac | 0.00 | 0.00 | 0.00 | 15 |
| siam_j._comput. | 1.00 | 0.97 | 0.99 | 271 |
| sigcse | 1.00 | 1.00 | 1.00 | 521 |
| sigir | 1.00 | 1.00 | 1.00 | 604 |
| sigmod_conference | 1.00 | 1.00 | 1.00 | 781 |
| smc | 1.00 | 1.00 | 1.00 | 873 |
| **avg / total** | **0.98** | **0.98** | **0.98** | **44629** |

## 5. TF-IDF WEIGHTED BAG OF WORDS MODEL

Macro-F1 score on Validation set: 0.22451030875267966

Micro-F1 score on Validation set: 0.31983687736673466

Classification Report:

| | **precision** | **recall** | **f1-score** | **support** |
|---|---|---|---|---|
| aaai | 0.09 | 0.07 | 0.08 | 528 |
| aamas | 0.32 | 0.40 | 0.36 | 349 |
| acc | 0.05 | 0.03 | 0.04 | 109 |
| acm_multimedia | 0.23 | 0.22 | 0.23 | 379 |
| acm_trans._graph. | 0.00 | 0.00 | 0.00 | 1 |

| | | | | |
|---|---|---|---|---|
| amcis | 0.31 | 0.24 | 0.27 | 665 |
| amia | 0.44 | 0.60 | 0.51 | 121 |
| asp-dac | 0.16 | 0.13 | 0.14 | 319 |
| bioinformatics | 0.06 | 0.06 | 0.06 | 17 |
| cdc | 0.43 | 0.67 | 0.53 | 843 |
| chi | 0.26 | 0.24 | 0.25 | 418 |
| chi_extended_abstracts | 0.39 | 0.36 | 0.38 | 568 |
| cikm | 0.12 | 0.11 | 0.12 | 377 |
| cogsci | 0.47 | 0.61 | 0.53 | 434 |
| coling | 0.34 | 0.33 | 0.33 | 367 |
| commun._acm | 0.09 | 0.10 | 0.10 | 68 |
| compsac | 0.16 | 0.12 | 0.14 | 296 |
| comput._graph._forum | 0.12 | 0.05 | 0.07 | 19 |
| comput._j. | 0.00 | 0.00 | 0.00 | 2 |
| computer_communications | 0.26 | 0.18 | 0.22 | 593 |
| computer_networks | 0.29 | 0.22 | 0.25 | 446 |
| corr | 0.22 | 0.16 | 0.19 | 490 |
| cvpr | 0.05 | 0.03 | 0.04 | 187 |
| dac | 0.29 | 0.18 | 0.22 | 385 |
| date | 0.37 | 0.36 | 0.37 | 342 |
| ecai | 0.41 | 0.56 | 0.48 | 301 |
| ecis | 0.30 | 0.24 | 0.27 | 255 |
| embc | 0.21 | 0.13 | 0.16 | 318 |
| encyclopedia_of_database_systems | 0.19 | 0.14 | 0.16 | 852 |
| etfa | 0.23 | 0.16 | 0.19 | 259 |
| eurospeech | 0.13 | 0.04 | 0.06 | 289 |
| eusipco | 0.00 | 0.00 | 0.00 | 1 |
| expert_syst._appl. | 0.40 | 0.47 | 0.43 | 286 |
| focs | 0.47 | 0.72 | 0.57 | 490 |
| fskd | 0.44 | 0.37 | 0.40 | 385 |
| fundam._inform. | 0.20 | 0.17 | 0.18 | 1323 |

| | | | | |
|---|---|---|---|---|
| fusion | 0.38 | 0.29 | 0.33 | 981 |
| fuzz-ieee | 0.49 | 0.54 | 0.52 | 338 |
| gecco | 0.02 | 0.01 | 0.01 | 269 |
| globecom | 0.26 | 0.24 | 0.25 | 1662 |
| hicss | 0.19 | 0.14 | 0.16 | 1380 |
| icalt | 0.22 | 0.15 | 0.17 | 302 |
| icarcv | 0.23 | 0.23 | 0.23 | 211 |
| icassp | 0.15 | 0.10 | 0.12 | 272 |
| icc | 0.44 | 0.76 | 0.56 | 284 |
| iccad | 0.21 | 0.17 | 0.19 | 343 |
| iccs | 0.13 | 0.05 | 0.08 | 293 |
| iccv | 0.28 | 0.27 | 0.28 | 947 |
| icdar | 0.17 | 0.14 | 0.15 | 407 |
| icde | 0.44 | 0.80 | 0.57 | 435 |
| icecs | 0.16 | 0.09 | 0.12 | 540 |
| icip | 0.17 | 0.19 | 0.18 | 264 |
| icis | 0.12 | 0.12 | 0.12 | 210 |
| icmc | 0.07 | 0.03 | 0.04 | 456 |
| icme | 0.41 | 0.52 | 0.46 | 1793 |
| icml | 0.31 | 0.39 | 0.34 | 324 |
| icnc | 0.20 | 0.07 | 0.11 | 283 |
| icpr | 0.00 | 0.00 | 0.00 | 18 |
| icra | 0.37 | 0.43 | 0.40 | 396 |
| icse | 0.00 | 0.00 | 0.00 | 1 |
| icslp | 0.00 | 0.00 | 0.00 | 4 |
| ieee_computer | 0.00 | 0.00 | 0.00 | 19 |
| ieee_congress_on_evolutionary_computation | 0.15 | 0.05 | 0.07 | 42 |
| ieee_journal_on_selected_areas_in_communications | 0.00 | 0.00 | 0.00 | 1 |
| ieee_software | 0.04 | 0.02 | 0.03 | 46 |
| ieee_trans._computers | 0.51 | 0.91 | 0.65 | 1325 |
| ieee_trans._information_theory | 0.20 | 0.17 | 0.18 | 752 |

| | | | | |
|---|---|---|---|---|
| ieee_trans._knowl._data_eng. | 0.30 | 0.25 | 0.27 | 543 |
| ieee_trans._parallel_distrib._syst. | 0.00 | 0.00 | 0.00 | 21 |
| ieee_trans._pattern_anal._mach._intell. | 0.27 | 0.28 | 0.28 | 690 |
| ieee_trans._software_eng. | 0.53 | 0.70 | 0.60 | 358 |
| igarss | 0.46 | 0.72 | 0.56 | 1252 |
| ijcai | 0.33 | 0.30 | 0.31 | 449 |
| ijcnn | 0.31 | 0.20 | 0.24 | 1399 |
| inf._process._lett. | 0.44 | 0.68 | 0.54 | 417 |
| inf._sci. | 0.36 | 0.42 | 0.39 | 889 |
| infocom | 0.06 | 0.05 | 0.05 | 323 |
| int._cmg_conference | 0.39 | 0.61 | 0.48 | 460 |
| interspeech | 0.50 | 0.74 | 0.60 | 452 |
| ipdps | 0.07 | 0.03 | 0.04 | 32 |
| iros | 0.00 | 0.00 | 0.00 | 1 |
| isbi | 0.00 | 0.00 | 0.00 | 1 |
| iscas | 0.18 | 0.24 | 0.21 | 239 |
| iscc | 0.15 | 0.16 | 0.16 | 286 |
| isit | 0.44 | 0.66 | 0.53 | 461 |
| itc | 0.00 | 0.00 | 0.00 | 1 |
| j._acm | 0.24 | 0.26 | 0.25 | 588 |
| j._parallel_distrib._comput. | 0.19 | 0.13 | 0.15 | 292 |
| j._symb._log. | 0.00 | 0.00 | 0.00 | 9 |
| journal_of_systems_and_software | 0.19 | 0.21 | 0.20 | 301 |
| kdd | 0.22 | 0.14 | 0.17 | 898 |
| lcn | 0.19 | 0.13 | 0.16 | 404 |
| lrec | 0.14 | 0.10 | 0.12 | 577 |
| multimedia_tools_appl. | 0.00 | 0.00 | 0.00 | 14 |
| neuroimage | 0.52 | 0.78 | 0.62 | 392 |
| nips | 0.33 | 0.34 | 0.34 | 351 |
| pacis | 0.20 | 0.18 | 0.19 | 299 |
| pattern_recognition | 0.14 | 0.11 | 0.12 | 871 |

| | | | | |
|---|---|---|---|---|
| pdpta | 0.29 | 0.35 | 0.32 | 292 |
| pimrc | 0.33 | 0.09 | 0.14 | 11 |
| robio | 0.33 | 0.20 | 0.25 | 311 |
| sac | 0.08 | 0.07 | 0.07 | 15 |
| siam_j._comput. | 0.10 | 0.07 | 0.08 | 271 |
| sigcse | 0.12 | 0.13 | 0.12 | 521 |
| sigir | 0.12 | 0.15 | 0.14 | 604 |
| sigmod_conference | 0.13 | 0.09 | 0.10 | 781 |
| smc | 0.48 | 0.78 | 0.59 | 873 |
| **avg / total** | **0.29** | **0.32** | **0.29** | **44629** |

## 6. ANALYSIS

Statistical experiments on validation set yield that 96.164% percent of papers are published in one of their cited venues, which shows how important cited venues features are because papers are usually published in the same venue as the papers they cite.

We observe that combining HIN-based features with text-based features greatly outperforms model that uses text-based features alone at both precision, recall and hence F-1 score. This is because textual content of title hardly gives us any information about the venue it may be suitable for, as well as there are several venues for the same research domain. It is plausible that author cites his own paper, reads papers of a venue and submits to the venue of his/her preference. Cited venue also demonstrates the research domain of paper better than textual content of its title. This is because titles are sometimes intended to be catchy but are not very informative.

Therefore, heterogenous information networks contain richer information than simple text-based model. Concatenating cited venue strings in title essentially adds them to tokens in bag-of-words vector. Since venue names are unique and distinct from words found in vocabulary of paper titles, therefore their occurence in title name essentially serves as concatenating title bag-of-words vector with one-hot-vector-encoding of cited paper's venues. Thus, our model incorporates the meta-path *Paper1 → cites → Paper2 → published_in → Venue1*. This illustrates that more than content, context matters as well. I believe including more metapaths will enrich this model even more.

We also observe in later section, even with word embeddings, text is much less important than cited venues in determining venue of a publication. However, embeddings achieve similar results as bag of words model with as few as 200 length feature vectors.

**The downside of bag-of-words unigram model is:**

1. Features vectors are rather large and sparse, which results in slow training

2. Our model only uses word counts and for example, cannot differentiate between a statement and a question that use the same words, or a random permutation of the same words/ phrases which could mean something entirely else. Therefore, a better idea would be to use bi-grams or tri-grams, but this would explode the feature vector length.

3. Text features used are a simple bag-of words, as such similar words are not considered when doing classification. Contextual similarity can be incorporated using word embeddings.

**Ways to improve:**

1. Add more heterogeneity such as author nodes which will encode author's preferences to publish at a certain venue.

2. Add more metapaths such as *Paper-> Keywords -> Venue*

3. Longer metapaths that include cited venues of cited papers as well.

4. Better data cleaning such as removing stopwords.

5. Reduce feature vector length by using word embeddings such as word2vec.

6. Count Vectors should be tf-idf weighted since not all words carry same information.

7. Rather than unigram bag of words model, use bi-grams or tri-grams.