We can combine both scores to determine whether the focus of the project is on the identified rare disease. One simple approach is to consider the product of the normalized confidence score and the semantic similarity score as a function to check. Using the manually checked datasets, we evaluate the performance metrics for various thresholds, as shown in the following table:

Table 1- Performance Metrics for Combined Normalized Confidence and Semantic Similarity Scores in Identifying Rare Diseases

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Threshold** | **TP** | **FP** | **FN** | **TN** | **FP percentage** | **Accuracy** |
| **0.01** | 1210 | 479 | 43 | 90 | 26.29 | 0.714 |
| **0.025** | 1209 | 469 | 44 | 100 | 25.741 | 0.718 |
| **0.05** | 1205 | 437 | 48 | 132 | 23.985 | 0.734 |
| **0.1** | 1186 | 371 | 67 | 198 | 20.362 | 0.76 |
| **0.2** | 1074 | 245 | 179 | 324 | 13.447 | 0.767 |
| **0.3** | 988 | 173 | 265 | 396 | 9.495 | 0.76 |
| **0.4** | 905 | 118 | 348 | 451 | 6.476 | 0.744 |
| **0.5** | 824 | 88 | 429 | 481 | 4.83 | 0.716 |
| **0.55** | 748 | 70 | 505 | 499 | 3.842 | 0.684 |
| **0.6** | 686 | 55 | 567 | 514 | 3.019 | 0.659 |
| **0.7** | 584 | 40 | 669 | 529 | 2.195 | 0.611 |
| **0.9** | 231 | 7 | 1022 | 562 | 0.384 | 0.435 |

Another approach is to consider the average of the normalized confidence score and the semantic similarity score as a metric. Using the manually checked datasets, we evaluate performance metrics for various thresholds, as shown in the following table:

Table 2- Performance Metrics for Average Normalized Confidence and Semantic Similarity Scores in Identifying Rare Diseases

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Threshold** | **TP** | **FP** | **FN** | **TN** | **FP percentage** | **Accuracy** |
| **0.01** | 1253 | 563 | 0 | 6 | 30.9 | 0.691 |
| **0.025** | 1253 | 559 | 0 | 10 | 30.681 | 0.693 |
| **0.05** | 1253 | 553 | 0 | 16 | 30.351 | 0.696 |
| **0.1** | 1251 | 543 | 2 | 26 | 29.802 | 0.701 |
| **0.2** | 1243 | 496 | 10 | 73 | 27.223 | 0.722 |
| **0.3** | 1225 | 423 | 28 | 146 | 23.216 | 0.752 |
| **0.4** | 1175 | 341 | 78 | 228 | 18.716 | 0.77 |
| **0.5** | 1105 | 252 | 148 | 317 | 13.831 | 0.78 |
| **0.55** | 1038 | 198 | 215 | 371 | 10.867 | 0.773 |
| **0.6** | 978 | 144 | 275 | 425 | 7.903 | 0.77 |
| **0.7** | 841 | 93 | 412 | 476 | 5.104 | 0.723 |
| **0.9** | 440 | 24 | 813 | 545 | 1.317 | 0.541 |

In general, we can explore various functions using both scores, different normalization factors for the confidence score, and various thresholds to determine whether the focus of the project is on the identified rare disease. We conducted a grid search to evaluate their performance.

1. Normalization factors for the confidence score are chosen from [3, 5, 10, 15, 20, 30].
2. Thresholds can be selected from [0.01, 0.025, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.55, 0.6, 0.65, 0.7, 0.8, 0.9, 0.95].
3. Functions can include , Ꞩ, ( + 4 Ꞩ *)/5, ( + 3* Ꞩ)/4, ( + 2Ꞩ*)/3, ( +* Ꞩ*)/2,* Ꞩ, \*(), \*(), where and Ꞩ represent the normalized confidence score and semantic similarity score, respectively.

We can define our goal as minimizing FP, maximizing accuracy, or optimizing both. For each goal, we performed a grid search to identify the top 3 choices for threshold, normalization factor, and function, as shown in the table below.

Table 8- Grid Search Results for Optimizing Rare Disease Identification Using Combined Scores

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Goal** | **Accuracy** | **FP percentage (count)** | **Threshold** | **Normalization coefficient** | **Function** |
| **FP <1** | 0.318 | %0.055 (1) | 0.9 | 30 |  |
| **FP <1** | 0.314 | %0.055 (1) | 0.95 | 30 |  |
| **FP <1** | 0.35 | %0.055 (1) | 0.95 | 30 | Ꞩ |
| **Accuracy>80** | 0.801 | %12.075 (220) | 0.6 | 10 | *( + 3* Ꞩ)/4 |
| **Accuracy>80** | 0.806 | %12.514 (228) | 0.55 | 30 | *( + 3* Ꞩ)/4 |
| **Accuracy>80** | 0.802 | %12.294 (224) | 0.6 | 10 | ( + 4 Ꞩ *)/5* |
| **Accuracy>0.75 and FP<100** | 0.755 | %5.379 (98) | 0.2 | 30 | \*() |
| **Accuracy>0.75 and FP<100** | 0.76 | %5.324 (97) | 0.7 | 20 | ( + 2Ꞩ*)/3* |
| **Accuracy>0.75 and FP<100** | 0.767 | %5.43 (99) | 0.7 | 30 | *( + 3* Ꞩ)/4 |