**Experiment Paper**

Table 1: Each document is tested in system and by the expert per sentence

|  |  |  |
| --- | --- | --- |
| Sentences | Inappropriate | |
| System | Expert |
| 1 |  |  |
| 2 |  |  |
| 3 |  |  |
| 4 |  |  |
| 5 |  |  |
| 6 |  |  |
| 7 |  |  |
| 8 |  |  |
| 9 |  |  |
| 10 |  |  |

**\***Sampling Methodology is Purposive Random Sampling

Table 1 is where the answers for the sentence analysis is recorded for each input. The expert labeled each input as yes, if the input is inappropriate, or no if it is not. The researchers also did the same for the system. Based from the results, each sentences was classified into the following:

\* TP (True Positive) – expert and system both determined the input is inappropriate

\* FP (False Positive) – System determined the input is inappropriate present, the expert indicated it’s not

\* TN (True Negative) – both the expert and the system indicated that the input is not inappropriate

\* FN (False Negative) – system indicated that the input is not offensive, the expert indicated it is inappropriate

Table 2: Input Scoring

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sentence | TP | FP | TN | FN |
| 1 |  |  |  |  |
| 2 |  |  |  |  |
| 3 |  |  |  |  |
| 4 |  |  |  |  |
| 5 |  |  |  |  |
| 6 |  |  |  |  |
| 7 |  |  |  |  |
| 8 |  |  |  |  |
| 9 |  |  |  |  |
| 10 |  |  |  |  |
| Total |  |  |  |  |
| Average |  |  |  |  |

Table 2 is where each input will be labeled as true positive, false positive, true negative, and false positive. This scoring will be used for the evaluation of data per document in which the following metrics is used: Precision: the percent of identified inputs that is inappropriate. Recall: the percent of inappropriate expressions correctly identified. Specificity: rate of the results without the condition, which have a negative test result. F- measure: the weighted harmonic mean of precision and recall. We used F- measure which gives equal weight to precision and recall.

Table 3: Averaging of Results per Document

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Document | Ave. TP | Ave. FP | Ave. TN | Ave. FN |
| 1 |  |  |  |  |
| 2 |  |  |  |  |
| 3 |  |  |  |  |
| 4 |  |  |  |  |
| 5 |  |  |  |  |
| 6 |  |  |  |  |
| 7 |  |  |  |  |
| 8 |  |  |  |  |
| 9 |  |  |  |  |
| 10 |  |  |  |  |
| Total |  |  |  |  |
| Average |  |  |  |  |

Table 3 is where each document’s average true positive, average false positive, true negative, and false positive. This scoring will be used for the evaluation of data as overall system performance in which the following metrics is used: Precision: the percent of identified inputs that is inappropriate. Recall: the percent of inappropriate sentences correctly identified. Specificity: rate of the results without the condition, which have a negative test result. F- measure: the weighted harmonic mean of precision and recall. We used F- measure which gives equal weight to precision and recall.