Chapter 4

Results

4.1 Performance

4.1.1 Credibility Classification

The performance of all 91 models were evaluated via the credibility classification experiment. Divided into the three main groups of models (NB, SVM and QRNN), the NB-based models (Table 4.2) outperformed the SVM-based (Table 4.4) and QRNN-based models (Table 4.6) overall with an aggregated average performance measure of 0.795, followed by SVM-based models with a measure of 0.780 and finally QRNN-based models with a measure of 0.397. Despite the NB-based models having performed better on average, the best performing model for labelling was the SVM-based, SVM + TF-IDF model with the stopwords removed from the article's text.

Performance of the QRNN-based models was lower than what was initially expected, however this can be theorised to be due to the relatively low number of training samples used for training the fine-tuned LM. When compared to the work conducted by Howard et al. [23], the number of training samples used to construct the discriminative fine-tuned LM for question classification was 5,500 which only improved the model's performance by 0.3%

Note #1 for Adam: Im thinking of investigating the relationship between the resulting f1-scores for a criteria and the label imbalance of that criteria (See the average f1 scores for Criteria 3, the most imbalanced label also has the highest f1 scores). Do you think this is worthwhile?

Naive Bayes

Model	Criteria 1	Criteria 2	Criteria 3	Criteria 4	Criteria 5	Criteria 6	Criteria 7	Average Performance
NB + BoW (all words)	0.79	0.70	0.86	**0.84**	0.79	**0.75**	**0.82**	0.792
NB + BoW (stopwords removed)	**0.85**	0.79	**0.90**	**0.84**	**0.80**	0.74	0.80	**0.820**
NB + TF-IDF (all words)	0.79	0.74	0.87	0.80	0.79	0.62	0.80	0.773
NB + TF-IDF (stopwords removed)	0.79	**0.80**	0.89	0.81	0.77	0.69	0.81	0.794
NB + GloVe	0.82	0.78	0.89	0.82	0.78	0.68	0.83	0.800

Table 4.1: Micro averaged f1-Scores of the NB models

Model	NB + BoW (all words)	NB + BoW (stopwords removed)	NB + TF-IDF (all words)	NB + TF-IDF (stopwords removed)	NB + GloVe
Criteria 1	0.79	0.85	0.79	0.79	0.82
Criteria 2	0.70	0.79	0.74	0.80	0.78
Criteria 3	0.86	0.90	0.87	0.89	0.89
Criteria 4	0.84	0.84	0.80	0.81	0.82
Criteria 5	0.79	0.80	0.79	0.77	0.78
Criteria 6	0.75	0.74	0.62	0.69	0.68
Criteria 7	0.82	0.80	0.80	0.81	0.83
Average Performance	0.792	**0.820**	0.773	0.794	0.800

Table 4.2: Micro averaged f1-Scores of the NB models

Support Vector Machine

Model	**Criteria 1**	**Criteria 2**	**Criteria 3**	**Criteria 4**	**Criteria 5**	**Criteria 6**	**Criteria 7**	**Average Performance**
SVM + BoW (all words)	0.82	0.70	0.86	0.80	0.68	0.75	**0.86**	0.799
SVM + BoW (stopwords removed)	0.78	0.80	**0.90**	**0.87**	0.78	0.72	0.81	0.809
SVM + TF-IDF (all words)	0.80	0.70	0.84	0.84	**0.80**	0.68	0.82	0.783
SVM + TF-IDF (stopwords removed)	**0.86**	**0.85**	0.89	0.86	**0.80**	**0.77**	0.82	**0.836**
SVM + GloVe	0.60	0.77	0.63	0.61	0.59	0.67	0.81	0.669

Table 4.3: Micro averaged f1-Scores of the SVM models

Model	**SVM + BoW (all words)**	**SVM + BoW (stopwords removed)**	**SVM + TF-IDF (all words)**	**SVM + TF-IDF (stopwords removed)**	**SVM + GloVe**
Criteria 1	0.82	0.78	0.80	**0.86**	0.60
Criteria 2	0.70	0.80	0.70	**0.85**	0.77
Criteria 3	0.86	**0.90**	0.84	0.89	0.63
Criteria 4	0.80	**0.87**	0.84	0.86	0.61
Criteria 5	0.68	0.78	**0.80**	**0.80**	0.59
Criteria 6	0.75	0.72	0.68	**0.77**	0.67
Criteria 7	**0.86**	0.81	0.82	0.82	0.81
Average Performance	0.799	0.809	0.783	**0.836**	0.669

Table 4.4: Micro averaged f1-Scores of the SVM models

Quasi-Recurrent Neural Network

Model	**Criteria 1**	**Criteria 2**	**Criteria 3**	**Criteria 4**	**Criteria 5**	**Criteria 6**	**Criteria 7**	**Average Performance**
QRNN + General LM	0.38	0.36	0.41	0.37	0.36	0.38	0.39	0.379
QRNN + Fine-tuned LM	**0.42**	**0.39**	**0.44**	**0.40**	**0.41**	**0.41**	**0.43**	**0.414**

Table 4.5: Micro averaged f1-Scores of the QRNN models

Model	QRNN + General LM	QRNN + Fine-tuned LM
Criteria 1	0.38	**0.42**
Criteria 2	0.36	**0.39**
Criteria 3	0.41	**0.44**
Criteria 4	0.37	**0.40**
Criteria 5	0.36	**0.41**
Criteria 6	0.38	**0.41**
Criteria 7	0.39	**0.43**
Average Performance	0.379	**0.414**

Table 4.6: Micro averaged f1-Scores of the QRNN models

4.1.2 Low Credibility Identification

Single Model Approach

–This model was chosen because it had the highest average performance value (average micro f1-score) from all of the models evaluated.–

Model	Precision	Recall	Micro Averaged f1-Score	Average Predicted Score Differential (Correctly Labelled)	Average Predicted Score Differential (Incorrectly Labelled)
SVM + TF-IDF (stopwords removed)	0.89	0.89	0.89	0.52	2.14

Table 4.7: Performance of low credibility identification task via an ensemble approach.

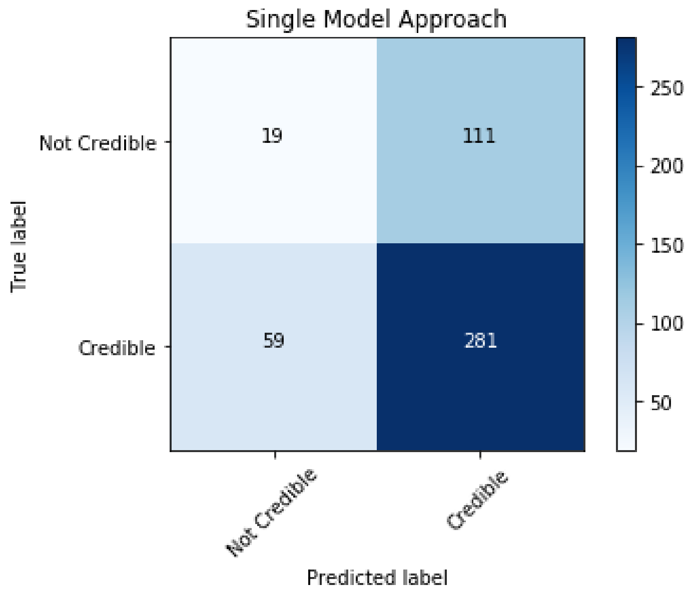

Figure 4.1: Performance of single model approach after 10-fold cross validation.

Ensemble Approach

–This will be composed of the classifiers that had the highest f1 score for each criteria–

Model	**Precision**	**Recall**	**Micro Averaged f1-Score**	**Average Predicted Score Differential (Correctly Labelled)**	**Average Predicted Score Differential (Incorrectly Labelled)**
Ensemble	0.91	0.91	0.91	0.55	2.0

Table 4.8: Performance of low credibility identification task via an ensemble approach.

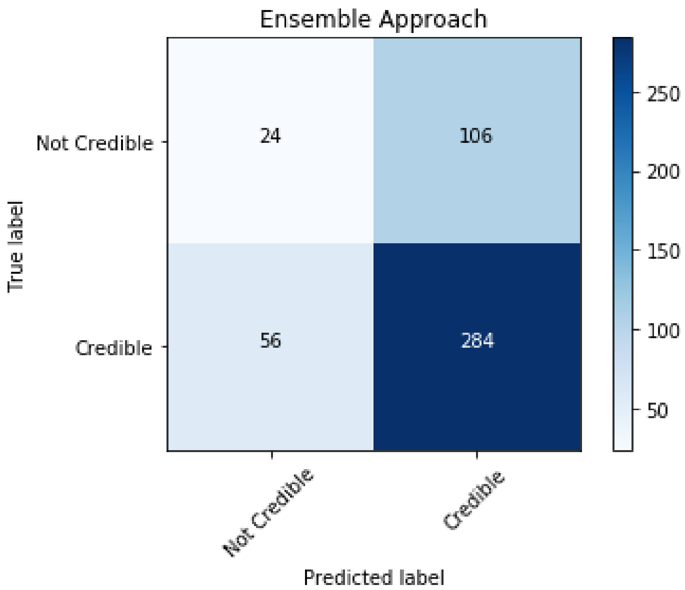

Figure 4.2: Performance of ensemble classifier after 10-fold cross validation.

Binary Classification Approach

Model	**Precision**	**Recall**	**Micro Averaged f1-Score**
SVM + TF-IDF (stopwords removed)	0.61	0.61	0.61

Table 4.9: Performance of low credibility identification task via a binary classification approach.

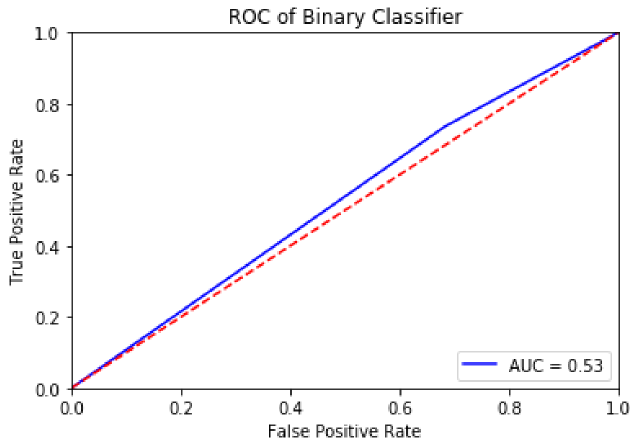

Figure 4.3: Performance of binary classification approach after 10-fold cross validation.

4.2 Training Time

The training times reported for each model includes the time required to perform the feature-specific preprocessing requirements in addition to the training time of the classifier itself. This helps explain the substantial differences of the training times between models.

Since the size and complexity in constructing count-based features such as BoW and TF-IDF (Table 4.10), especially when limited to an upper bound, is relatively lower when compared to the construction of statistical-based features such as a language model (Table 4.11) and relatively smaller than a pre-trained feature such as GloVe (Table 4.12), it is expected that the training times for models that utilise the count-based features are lower.

Model	**Training Time (seconds)**
NB + BoW (all words)	5.20
NB + BoW (stopwords removed)	**4.86**
NB + TF-IDF (all words)	5.05
NB + TF-IDF (stopwords removed)	5.05
SVM + BoW (all words)	5.91
SVM + BoW (stopwords removed)	4.87
SVM + TF-IDF (all words)	5.35
SVM + TF-IDF (stopwords removed)	5.58

Table 4.10: Average training time of models using count-based features.

Model	**Epoch Completion (minutes)**
QRNN + General LM	**271**
QRNN + Fine-tuned LM	292

Table 4.11: Average training time of models using statistical LMs.

Model	**Training Time (seconds)**
NB + GloVe	3,628.82
SVM + GloVe	**3,558.26**

Table 4.12: Average training time of models using pre-trained GloVe feature.

4.3 Model Storage Requirements

NB Model	**Size (MB)**
NB + BoW (all words)	11.10
NB + BoW (stopwords removed)	**11.00**
NB + TF-IDF (all words)	13.20
NB + TF-IDF (stopwords removed)	13.20
NB + GloVe	25,410

Table 4.13: Aggregated size of NB models.

NB Model	**Size (MB)**
SVM + BoW (all words)	7.91
SVM + BoW (stopwords removed)	**7.87**
SVM + TF-IDF (all words)	10.00
SVM + TF-IDF (stopwords removed)	7.88
SVM + GloVe	25,340

Table 4.14: Aggregated size of SVM models.

NB Model	**Size (MB)**
QRNN + General LM	**3,612.70**
QRNN + Fine-tuned LM	4,193.35

Table 4.15: Aggregated size of QRNN models.