Chapter 5

Experiments and Results

This chapter gives the details of the experiments carried out and their results. Then, failure analysis and the final analysis of the system is done. The chapter concludes with the conclusion made from the result.

5.1 Experiments

5.1.1 Sentence Splitting

For splitting the sentences syntactic parsing and a rule-based approach has been used in following five different experiments:

1. **Syntactic Parsing only:** Splitting of the sentences is done only using the syntactic parsing at the occurrence of conjunctions.
2. **Syntactic Parsing and Word Classification:** Splitting of the sentences is done using the syntactic parsing and the word classification, and at the occurrence of conjunctions and the splitting terms between the two words that have been classified to entities.
3. **Syntactic Parsing followed by Word Classification:** Splitting of the sentences is done first using the syntactic parsing at the occurrence of conjunctions and then the word classification at the splitting terms between the two words that have been classified to entities.
4. **Word Classification only:** Splitting of the sentences is done only using the word classification at the splitting terms between the two words that have been classified to entities.
5. **Word Classification followed by Syntactic Parsing:** Splitting of the sentences is done first using the classification at the splitting terms between the two words that have been classified to entities followed by the syntactic parsing at the occurrence of conjunctions.

5.1.2 Feature Extraction

For feature extraction, following fourteen experiments have been performed:

1. **Unigrams:** For this experiment classifiers were trained only on the unigrams.

2. **Lowercase unigrams:** The classifiers used in this experiment were trained only on the lowercase unigrams.

3. **Preprocessed lowercase unigrams:** For this experiment classifiers were trained only on the preprocessed lowercase unigrams.

4. **Preprocessed lowercase unigrams with stopword removal:** In this experiment classifiers were trained only on the preprocessed lowercase unigrams with stopwords removed.

5. **Preprocessed lowercase lemmatised unigrams:** For this experiment classifiers were trained on the preprocessed lowercase unigrams and the lowercase lemmatised unigrams.

6. **Preprocessed lowercase lemmatised unigrams with stopword removal:** In this experiment classifiers were trained on the preprocessed lowercase unigrams with stopwords removed and the lowercase lemmatised unigrams.

7. **Preprocessed lowercase lemmatised unigrams with stopword removal and min. document frequency:** For this experiment classifiers were trained on the preprocessed lowercase unigrams with stopwords removed and the lowercase lemmatised unigrams whose minimum document frequency is 5.

8. **Preprocessed lowercase lemmatised unigrams and bigrams with stopword removal and min. document frequency syntactic parsing based splitting:** The classifiers used in this experiment were trained on the preprocessed lowercase unigrams and bigrams with stopwords removed and the lowercase lemmatised unigrams whose minimum document frequency is 5 after the sentences were split using only syntactic parsing based sentence splitting.

9. **Preprocessed lowercase lemmatised unigrams and bigrams with stopword removal and min. document frequency word classification based splitting:** For this experiment classifiers were trained on the preprocessed lowercase unigrams and bigrams with stopwords removed and the lowercase lemmatised unigrams whose minimum document frequency is 5 after the sentences were split using only word classification based sentence splitting.

10. **Preprocessed lowercase lemmatised unigrams with stopword removal and min. document frequency word classification and syntactic parsing based splitting and TFIDF:** For this experiment classifiers were trained on the preprocessed lowercase unigrams and their TFIDF with stopwords removed and the lowercase lemmatised unigrams whose minimum document frequency is 5 after the sentences were split using syntactic parsing based sentence splitting.

11. **Preprocessed lowercase lemmatised unigrams and bigrams with stopword removal and min. document frequency word classification and syntactic parsing based splitting and TFIDF:** For this experiment classifiers were trained on the preprocessed lowercase unigrams and bigrams with stopwords removed and the lowercase lemmatised unigrams whose minimum document frequency is 5 after the sentences were split using word classification and syntactic parsing based sentence splitting.

12. **Preprocessed lowercase lemmatised unigrams and bigrams with stopword removal and min. document frequency syntactic parsing followed by word classification based splitting and TFIDF:** For this experiment classifiers were trained on the preprocessed lowercase unigrams and bigrams and their TFIDF with stopwords removed and the lowercase lemmatised unigrams whose minimum document frequency is 5 after the sentences were split using syntactic parsing followed by word classification based sentence splitting.

13. **Preprocessed lowercase lemmatised unigrams and bigrams with stopword removal and min. document frequency word classification followed by syntactic parsing based splitting and TFIDF:** For this experiment classifiers were trained on the preprocessed lowercase unigrams and bigrams with stopwords removed and the lowercase lemmatised unigrams whose minimum document frequency is 5 after the sentences were split using word classification followed by syntactic parsing based sentence splitting.

14. **Preprocessed lowercase lemmatised unigrams and bigrams with stopword removal and min. document frequency word classification and syntactic parsing based splitting and TFIDF:** For this experiment classifiers were trained on the preprocessed lowercase unigrams and bigrams with stopwords removed and the lowercase lemmatised unigrams whose minimum document frequency is 5 for entity classifier and 6 for aspect classifier after the sentences were split using word classification and syntactic parsing based sentence splitting.

The use of the experiment described above have been documented in the sections 5.3.3 through section 5.3.6. The number shown in the figures in these sections are in tandem to the numbers mentioned in these experiments.

5.1.3 Classifier Development

For classifier development, three machine learning algorithms, Naive Bayes, SGD, and LinearSVC were experimented on similar features and for two algorithms, SGD and LinearSVC, two types of designs were experimented upon. Following are the five design of classifiers that have been experimented upon:

1. **Multinomail Naive Bayes Classifier:** A single Multinomail Naive Bayes classifier have been developed for the classification of ENTITY#ASPECT.

2. **Combined Stochastic Gradient Descent Classifier:** A single Stochastic Gradient Descent classifier have been developed for the classification of ENTITY#ASPECT.

3. **Separate Stochastic Gradient Descent Classifier:** Two separate Stochastic Gradient Descent classifier have been developed for the classification of ENTITY and ASPECT from the fifth experiment onward of feature selection. After the classification of ENTITY and ASPECT, decision function have been used to predict the ENTITY#ASPECT.

4. **Combined Linear Support Vector Classifier:** A single Linear Support Vector classifier have been developed for the classification of ENTITY#ASPECT.

5. **Separate Linear Support Vector Classifier:** Two separate Linear Support Vector classifier have been developed for the classification of ENTITY and ASPECT from the fifth experiment onward of feature selection. After the classification of ENTITY and ASPECT, decision function have been used to predict the ENTITY#ASPECT.

5.2 Measurements

Evaluation of a supervised machine learning algorithm can be done using several measures as described in section 3.3. For this project, F1-score and accuracy will be used as an evaluation measurement.

5.3 Results

Various experiments were conducted for the implementation of various parts of the system. The results of those experiments are as follows:

5.3.1 Word classification

Classification of words to entities were done using LinearSVC classifier for the purpose of splitting the sentences. The system was trained on 664 words and was tested for 884 words. The system achieved a F1-score of 87.019.

5.3.2 Sentence Splitting

Sentences were split because many sentences have more than one entity attribute pairs. Sentences were split using syntactic parsing and rule-based approach. Accuracy was the evaluation measurement, which calculates the total number of splits of the sentences that matches the number of entity attribute pairs in the sentences by total number of splits of the sentences. For the sentence “There are many vegan options, they all are delicious but a little expensive.” there are three entity attribute pairs, viz., “FOOD#STYLE_OPTIONS”, “FOOD#QUALITY”, and “FOOD#PRICE”. If the system splits the aforementioned sentence into three sentences then it is counted as accurate and if it splits the sentence in more than three or less than three sentences then it is not counted as accurate.

Following are the result of five experiments that were performed for splitting the sentences:

1. **Syntactic Parsing only:** Following was the performance of splitting system which used only the syntactic parsing:

	Equal	Lesser	Greater	Total	Accuracy
Training	1052	233	423	1708	61.59%
Testing	363	83	141	587	61.84%
Average	707.5	158	282	1147.5	61.72%

2. **Syntactic Parsing and Word Classification:** Following was the performance of splitting system which used both, the syntactic parsing and word classification:

	Equal	Lesser	Greater	Total	Accuracy
Training	1037	145	526	1708	60.71%
Testing	372	52	163	587	63.37%
Average	704.5	98.5	344.5	1147.5	62.04%

3. **Syntactic Parsing followed by Word Classification:** Following was the performance of splitting system which used the syntactic parsing followed by the word classification:

	Equal	Lesser	Greater	Total	Accuracy
Training	1129	145	434	1708	66.10%
Testing	392	52	143	587	66.78%
Average	760.5	98.5	288.5	1147.5	66.44%

4. **Word Classification only:** Following was the performance of splitting system which used only the word classification:

	Equal	Lesser	Greater	Total	Accuracy
Training	1191	245	272	1708	69.73%
Testing	421	88	78	587	71.72%
Average	806	166.5	175	1147.5	70.73%

5. **Word Classification followed by Syntactic Parsing:** Following was the performance of splitting system which used the word classification followed by the syntactic parsing:

	Equal	Lesser	Greater	Total	Accuracy
Training	1274	146	288	1708	74.59%
Testing	448	52	87	587	76.32%
Average	861	99	187.5	1147.5	75.46%

It has been observed that the splitting system which splits the sentence based on the word classification followed by the splitting based on syntactic parsing achieved the highest accuracy of 75.46% and thus performs the best. The Figure 5.1 below shows the performance of the system for different datasets, viz., training and testing, and the average of it for different implementation.

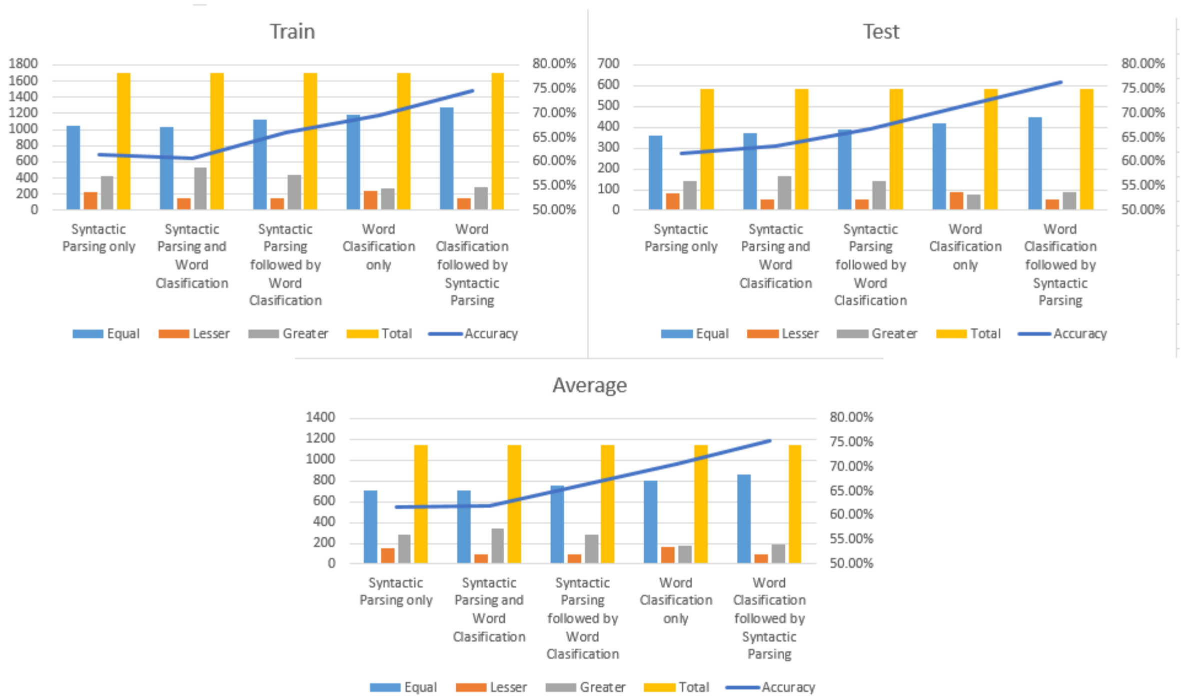
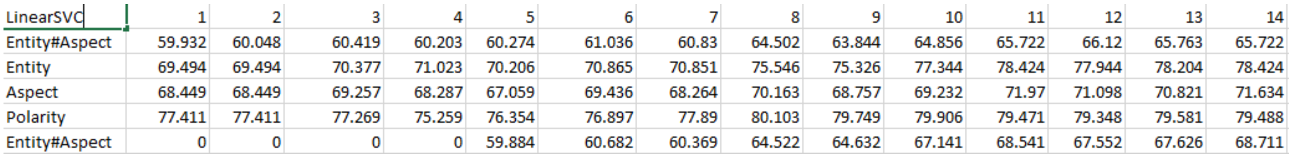


Figure 5.1: Splitting System Performance

5.3.3 Feature Extraction

A total of 14 combinations of feature extraction were experimented. For for Slot 1 (aspect classification) on linearSVC classifier a F1-score of 68.711 was achieved. Figure 5.2 and 5.3 shows the performance of different feature extraction combinations.



LinearSVC	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Entity#Aspect	59.932	60.048	60.419	60.203	60.274	61.036	60.83	64.502	63.844	64.856	65.722	66.12	65.763	65.722
Entity	69.494	69.494	70.377	71.023	70.206	70.865	70.851	75.546	75.326	77.344	78.424	77.944	78.204	78.424
Aspect	68.449	68.449	69.257	68.287	67.059	69.436	68.264	70.163	68.757	69.232	71.97	71.098	70.821	71.634
Polarity	77.411	77.411	77.269	75.259	76.354	76.897	77.89	80.103	79.749	79.906	79.471	79.348	79.581	79.488
Entity#Aspect	0	0	0	0	59.884	60.682	60.369	64.522	64.632	67.141	68.541	67.552	67.626	68.711

Figure 5.2: Feature Extraction Performance on LinearSVC classifier

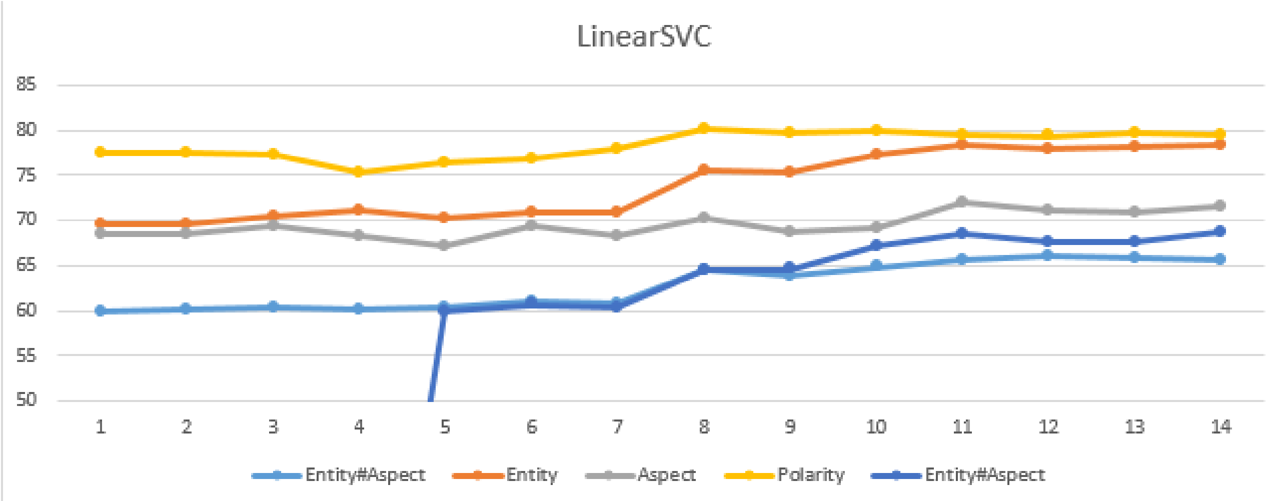


Figure 5.3: Feature Extraction Performance on LinearSVC classifier - Graph

5.3.4 Classifier Development

A total of 5 designs of classifiers were made. It was observed that among the three algorithms that were chosen, Multinomial Naive Bayes classifier performed the worst. Apart from three feature extraction, Stochastic Descent Gradient (SGD) classifier perform worse than the LinearSVC classifier, and apart from those three feature extraction, LinearSVC classifier performed the best. For SGD and LinearSVC classifier, two designs were made, one in which they were suppose to predict the ENTITY#ASPECT together and another in which they were supposed to predict ENTITY and ASPECT separately and then make an ENTITY#ASPECT pair. SGD classifier performed almost the same for both the designs, whereas the LinearSVC classifier better in the design in which it was supposed to predict ENTITY and ASPECT separately and then make an ENTITY#ASPECT pair. This is because for the separate prediction, it got more data to train on for lesser target label. Figure 5.4 below show the performance of different classifiers for different classifications and the comparison on the best design of the three algorithms on the ENTITY#ASPECT classification.

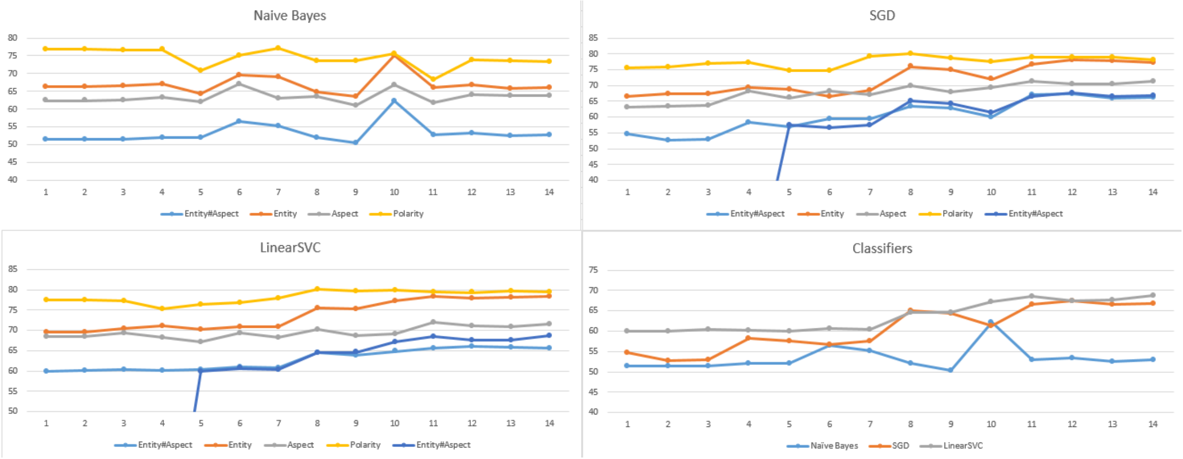


Figure 5.4: Classifier Comparison

5.3.5 Aspect Classification

The aspect classification was able the achieve the F1-score of 68.711. The Figure 5.5 shows the performance of aspect classification for different feature extraction for LinearSVC classifier.

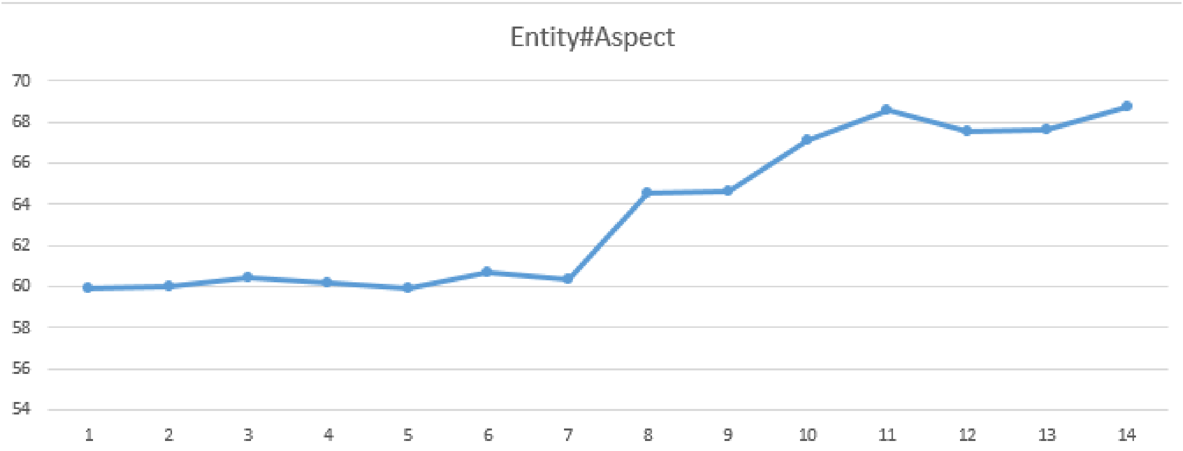


Figure 5.5: Entity#Aspect Classification Performance

5.3.6 Sentiment Classification

The sentiment classification was able to achieve the accuracy of 80.908%. The Figure 5.6 shows the performance of sentiment classification for different feature extraction for LinearSVC classifier.

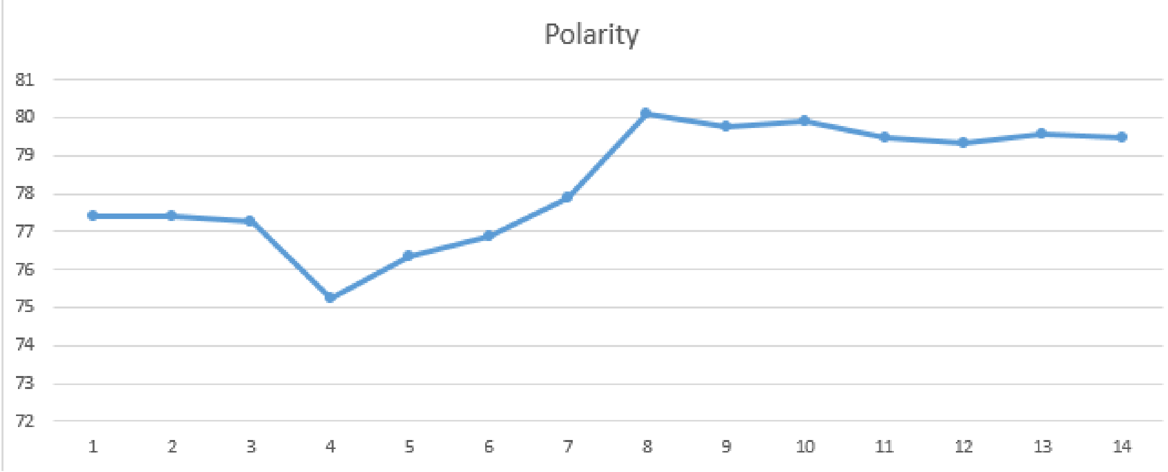


Figure 5.6: Polarity Classification Performance

5.3.7 Overall System Performance

The goal of the project was to develop a domain independent system, so although the system was developed keeping the restaurant domain in focus, the same unmodified system was tested for the laptop domain. The Aspect classification of the laptop domain was a much more fine grained task as compared to that of restaurant domain. This is because there are 71 ENTITY#ASPECT pairs in laptop domain as compared to restaurant domain which only have 12.

For slot 1 of the laptop domain, the system was able to pass the baseline F1-score of 37.481 and was able to achieve the F1-score of 42.176 in the constrained mode. The best system in constrained mode had the F1-score of 47.89 while the best system in unconstrained had the F1-score of 51.973 which the system was unable to achieve.

For slot 1 of the restaurant domain, the system performed well, and was able to achieve the F1-score of 68.711 which ranked it second among thirteen systems in the constrained mode and seventh among thirty one systems in the unconstrained mode. The best system in the constrained mode was able to achieve a F1-score of 71.494 and that in the unconstrained mode was able to achieve a F1-score of 73.031. Figure 5.7, 5.8, and 5.9 below shows the final performance of different classification, the standing of this system as compared to other systems in SemEval 2016 ABSA, and the performance of every system in SemEval 2016 ABSA and that of this system if this system would have taken part in it.

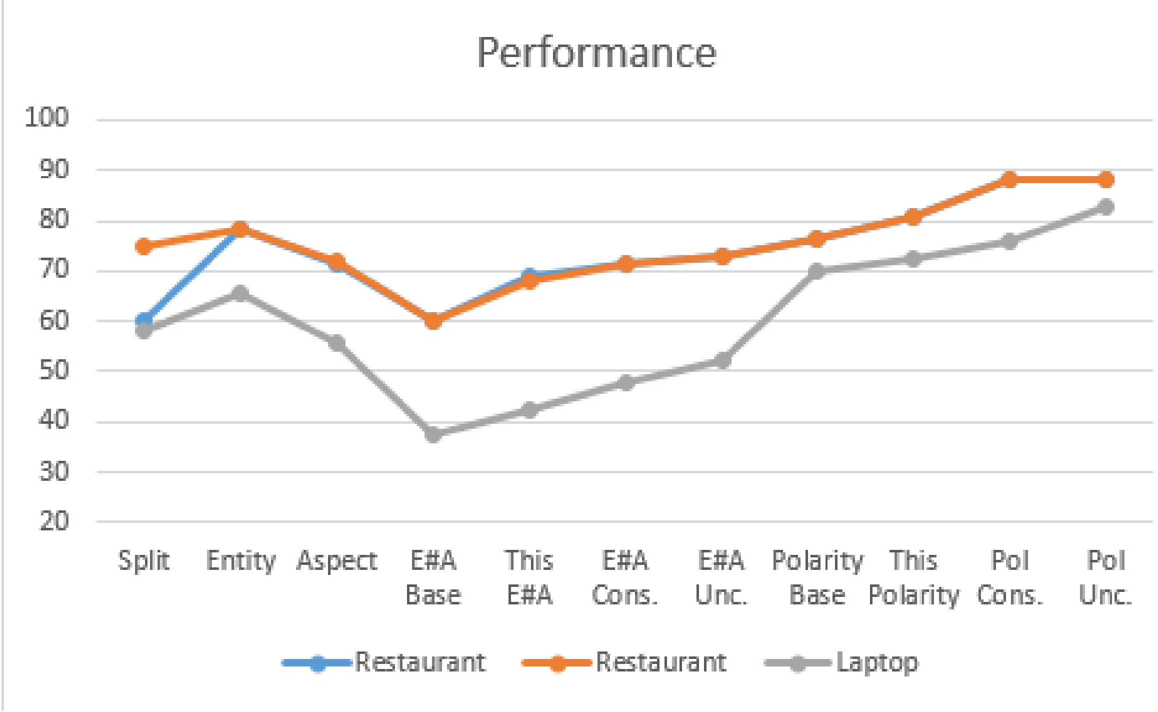


Figure 5.7: Overall System Performance

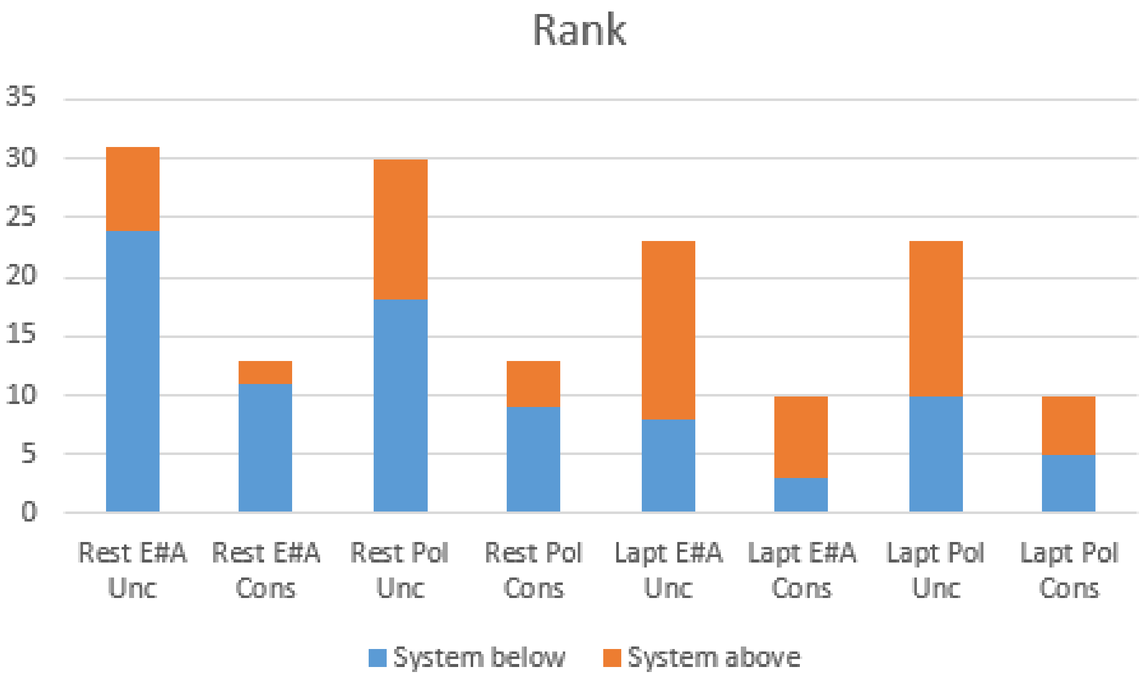
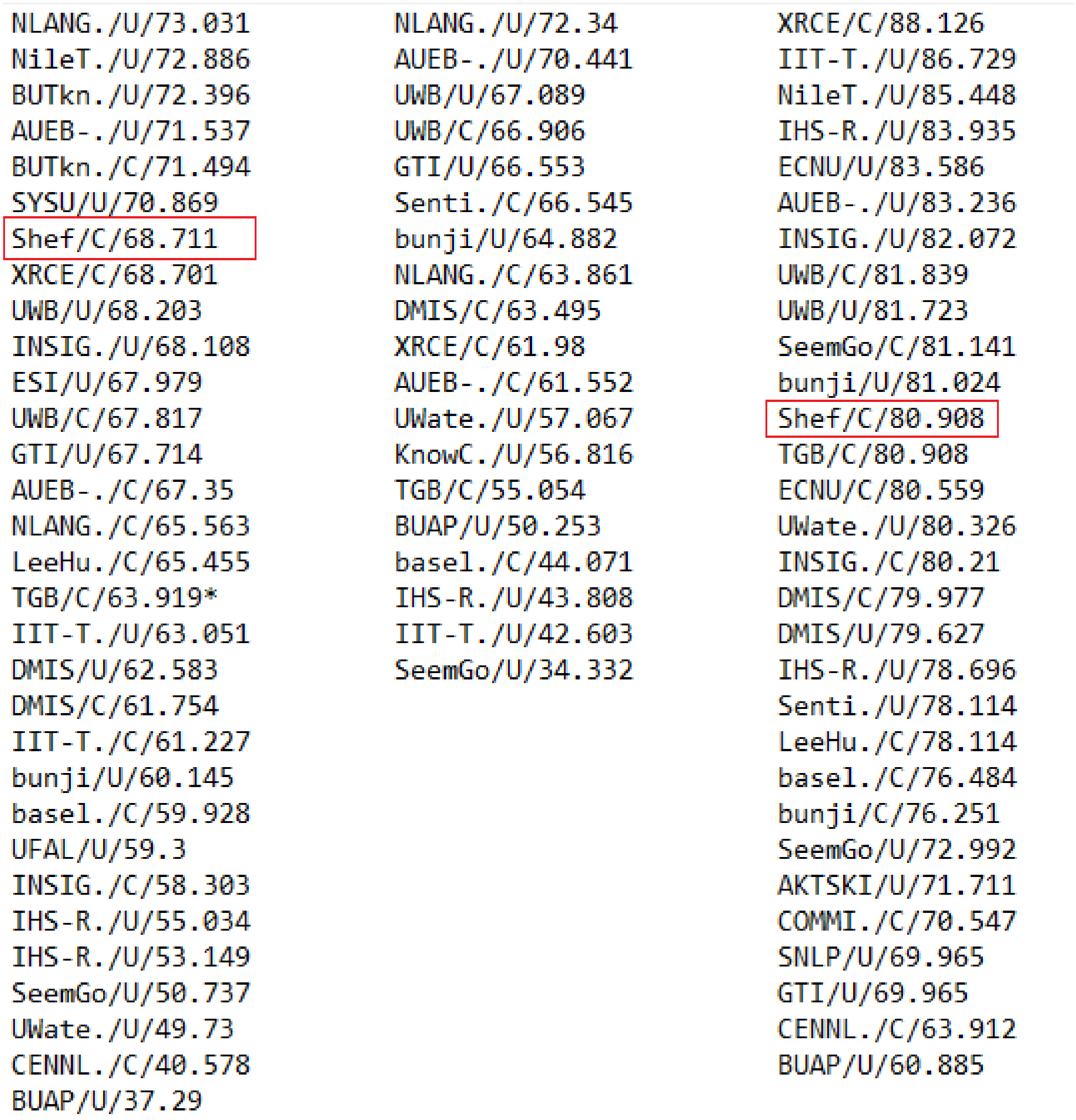


Figure 5.8: Rank of the system



```
NLANG./U/73.031         NLANG./U/72.34          XRCE/C/88.126
NileT./U/72.886         AUEB-./U/70.441         IIT-T./U/86.729
BUTkn./U/72.396         UWB/U/67.089            NileT./U/85.448
AUEB-./U/71.537         UWB/C/66.906            IHS-R./U/83.935
BUTkn./C/71.494         GTI/U/66.553            ECNU/U/83.586
SYSU/U/70.869           Senti./C/66.545         AUEB-./U/83.236
Shef/C/68.711           bunji/U/64.882          INSIG./U/82.072
XRCE/C/68.701           NLANG./C/63.861         UWB/C/81.839
UWB/U/68.203            DMIS/C/63.495           UWB/U/81.723
INSIG./U/68.108         XRCE/C/61.98            SeemGo/C/81.141
ESI/U/67.979            AUEB-./C/61.552         bunji/U/81.024
UWB/C/67.817            UWate./U/57.067         Shef/C/80.908
GTI/U/67.714            KnowC./U/56.816         TGB/C/80.908
AUEB-./C/67.35          TGB/C/55.054            ECNU/C/80.559
NLANG./C/65.563         BUAP/U/50.253           UWate./U/80.326
LeeHu./C/65.455         basel./C/44.071         INSIG./C/80.21
TGB/C/63.919*           IHS-R./U/43.808         DMIS/C/79.977
IIT-T./U/63.051         IIT-T./U/42.603         DMIS/U/79.627
DMIS/U/62.583           SeemGo/U/34.332         IHS-R./U/78.696
DMIS/C/61.754                                   Senti./U/78.114
IIT-T./C/61.227                                 LeeHu./C/78.114
bunji/U/60.145                                  basel./C/76.484
basel./C/59.928                                 bunji/C/76.251
UFAL/U/59.3                                     SeemGo/U/72.992
INSIG./C/58.303                                 AKTSKI/U/71.711
IHS-R./U/55.034                                 COMMI./C/70.547
IHS-R./U/53.149                                 SNLP/U/69.965
SeemGo/U/50.737                                 GTI/U/69.965
UWate./U/49.73                                  CENNL./C/63.912
CENNL./C/40.578                                 BUAP/U/60.885
BUAP/U/37.29
```

Figure 5.9: Final performance of each system

5.4 Failure Analysis

The accuracy of the system is dependent on various factors. This includes, the vastness of training data available, proper splitting of the sentences, and the training of the system on the relevant features. If the system has not been trained in a particular feature, the system will be unable to predict new data pertaining to said feature thereby negatively impacting the accuracy of the system. An example of the effect of the system not being trained on the relevant feature can be found in the instance that suitable amount of training data was not available for the ENTITY#ASPECT DRINKS#PRICES label. This led to the under-performance of the system corresponding to this label. The results of the same can be seen in the confusion matrix shown in the Figure 5.11 below. There are sentences which the system is unable to split because either it does not contain any conjuction or it does not contain any previously defined spliting terms, but have more than one clauses. If it were possible to develop a feature that would help the system to split these types of sentences, the accuracy of the system would be enhanced further. For example, the system was unable to split the sentence “The food was delicious for such a cheap price” which contained FOOD#QUALITY and FOOD#PRICES ENTITY#ASPECT label.

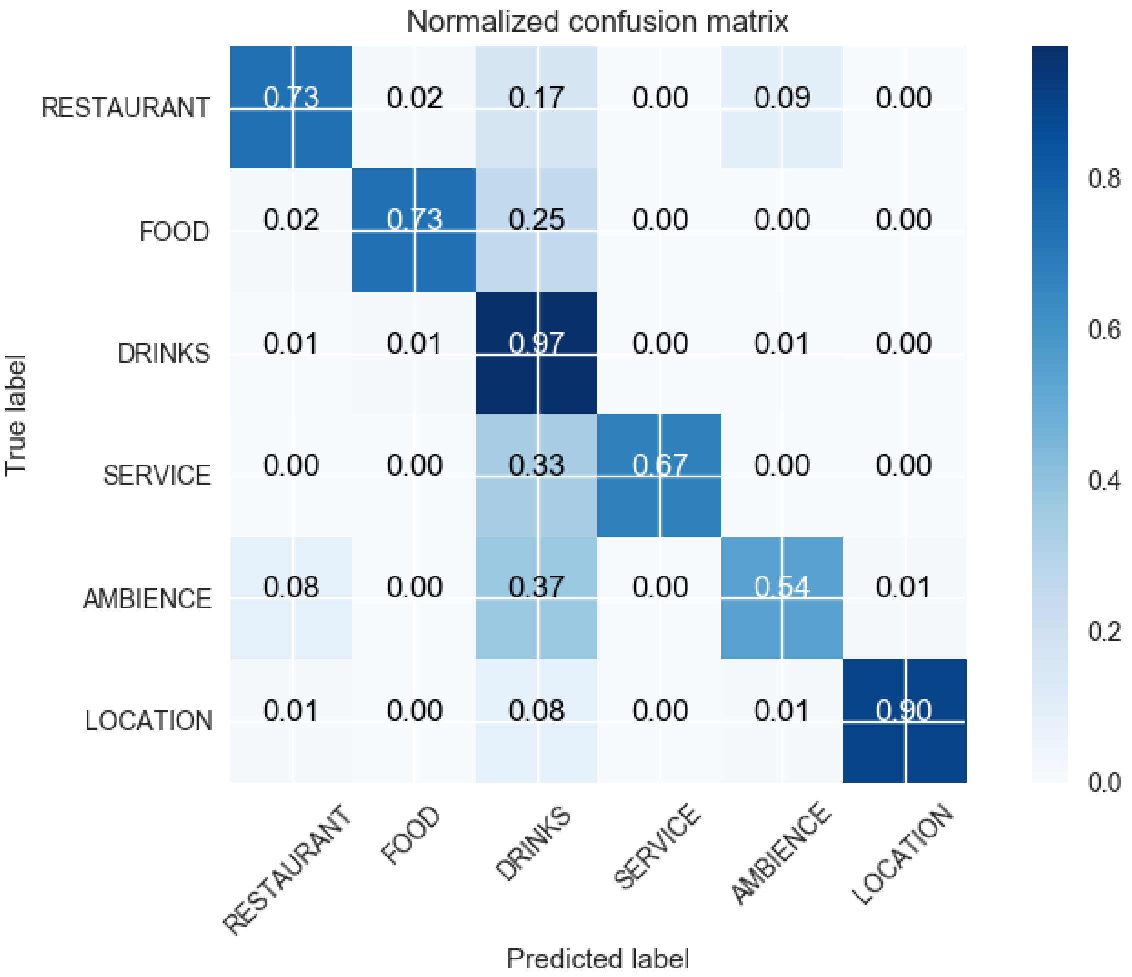


Figure 5.10: Confusion matrix for word classification

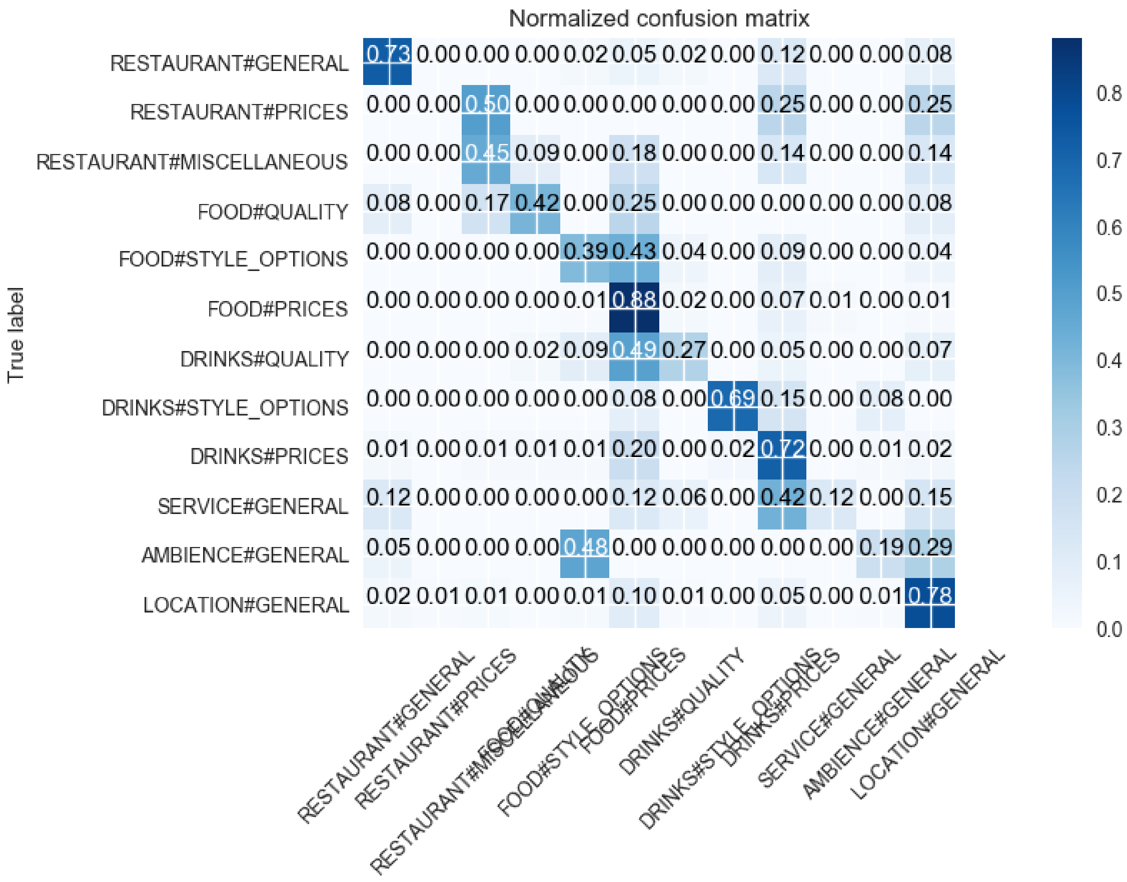


Figure 5.11: Confusion matrix for Entity#Aspect

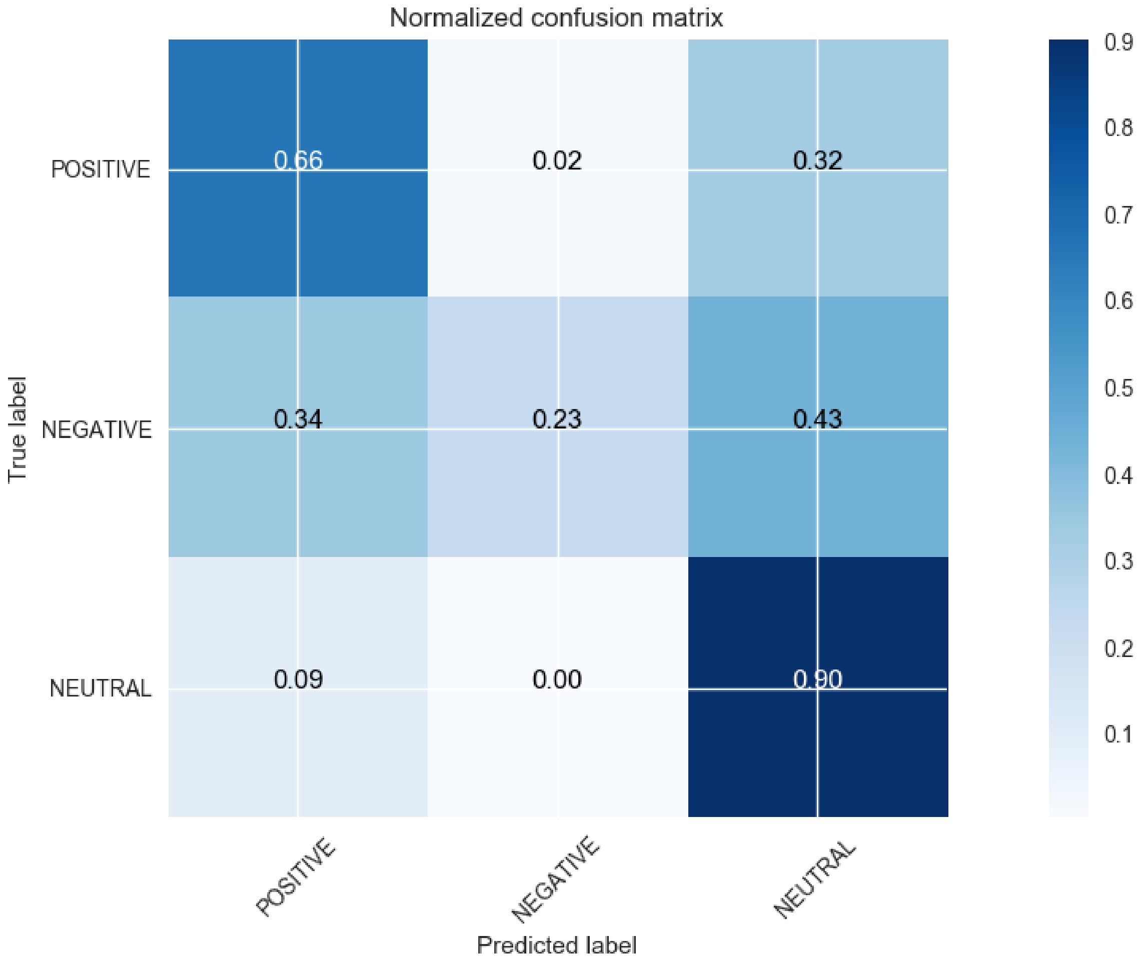


Figure 5.12: Confusion matrix for polarity clarification

5.5 Analysis

Other systems in SemEval 2016 ABSA task developed a separate binary classifier for each entities and aspects and calculated the probability for each label on the whole sentence and then used only the labels whose probability passed a threshold value. The drawback of such systems is that for a sentence with multiple entity-aspect pairs, each entity and aspect will read the signals for other entities and aspects and the probability of each entity and aspect will be less and the classifier won't be able to learn that efficiently. In the sentence "There are many vegan options, they all are delicious but a little expensive." the aforementioned systems will read the signal for "STYLE_OPTIONS", "QUALITY", and "PRICE" aspects, and each aspect will cancel the signal of every other aspect therefore reducing the probability of each aspect and thus the learning will not be accurate. But in this system in which sentences are split with an accuracy of 75%, the probability of one entity and aspect being in the split sentences is higher, which will further increase when the text simplification and splitting system will improve as discussed in the Future works section of next chapter. Thus the signal of the entity and aspect will not get canceled out by the signal of other entities and aspects, and thus the learning will be more accurate.

5.6 Conclusion

The experiments mentioned in this chapter were performed with the intention of creating a better and domain independent system for the subtasks of SemEval 2016 ABSA. From the results of the experiments, it was observed that the system was able to outperform other systems of SemEval 2016 ABSA with higher number of features when the sentences were split by this system. It might also be noted that the aforementioned system is also domain independent as it performed admirably in the restaurant domain that it was developed for while also performing competently in the laptop domain.