Aspect Based Sentiment Analysis

TEAM MEMBERS-

1. Gopi Krishnan N (gnaras2@uic.edu)
2. Mrudula Borkar (mborka2@uic.edu)

ABSTRACT –

Aspect based sentiment analysis systems receive input in a set of texts which consider a particular entity. It analyzes different features/ aspects of the product and focuses on finding important aspects and then finds its polarity expressed on those aspects. Polarity of per aspect is how positive, negative or neutral the opinions are on average per aspect term. ABSA makes it easier to identify and determine the sentiment towards specific aspects in given input text. Various research prototypes are proposed for ABSA but there is no established task decomposition.

This system proposes the following workflow –

1. Data preprocessing

2. Building a classifier model and

3. Evaluation of the classifier model.

INTRODUCTION –

Sentiment analysis which is also known as opinion mining which is the automated process of understanding an opinion about a given subject from written or spoken language. It is contextual text mining which extracts subjective material and understanding sentiment of the subject. A lot of information is available in unstructured format and with the help of sentiment analysis systems, this unstructured information can be transformed into a structured data by expressing opinions. This transformed data can be useful for commercial applications like marketing analysis, product reviews and customer service.

Aspect based sentiment analysis refers to determining the sentiments expressed on different aspects. An aspect is an attribute of entity. The advantage of aspect-based sentiment analysis is the possibility to capture nuances about the objects of interest. Usually, while analyzing the sentiments in subject, it becomes mandatory to know polarity as well as aspect of the subject.

Sentiment Analysis and Opinion mining are the interchangeable terms for some researchers while some researchers consider it as different notions. From the view of data mining, this can be considered as multi-step classification problem. Firstly, determination of the type such as – document level, sentence level or aspect level. And second, identify the polarity of the document, sentence or aspect as positive, neutral or negative. Document level SA determines aspect of whole document as positive or negative by considering the whole document as information unit. Sentence level SA determines the opinion of the sentence. Aspect based SA aims to identify the aspects of entities being used in expressing sentiments. It determines the sentiment expressed by the author towards each aspect of the entity. Aspect based SA is much critical. From past few decades, there are multiple aspect-based sentiment analysis systems are developed for variety of entities such as customer services, restaurants, reviews, travels etc.

Aspect based sentiment analysis systems receive a set of text discussing about an entity. The system attempts to find the main frequently discussed aspects of the entity to find the expressed sentiment towards each aspect and their summary of polarity.

DATA PREPROCESSING-

Let us first discuss the features of the given dataset. Two datasets are available i.e. Restaurant reviews and Laptop reviews with the following columns –

1. Example_id : Unique ID for each review
2. Text : Reviews mainly focusing on laptops and restaurants
3. Aspect_Term : The term to be classified as positive, neutral or negative
4. Aspect_Location: Start and end location of aspect term in the review
5. Sentiment Label: 1 (Positive) 0 (Neutral) -1 (Negative)

Preprocessing is important step for data analysis since it is used for extracting knowledge from unstructured text data. This step has to be carried out of the given columns – Text, Aspect Term.

Steps –

1. Data cleaning performed on Text column and Aspect term column

   i. Text is converted to lower case
   ii. Removal of stop words
   iii. Removal of punctuation marks
   iv. Noisy words were removed (Ex: [comma])

2. Feature engineering

    i. Generating embedding matrix using TFIDF model.
    ii. Comparing the text with a window size of 3 from the aspect term with positive and negative lexicons (Prof. Bing Liu's list of lexicons) and assigning +1 if a word in a text is found in positive lexicon and -1 if found in negative lexicon and 0 if not found in both the lists and a score of 0 for aspect term.
    iii. The above scores are averaged and multiplied with sentiment polarity of each word in the window. This value is appended to the end of embedding matrix generated in step 1.

Example for steps b and c:

Cleaned Text: "again same problem right speaker work properly".

Aspect term: "right speaker"

Since we have considered a window size of 3 from the aspect term, all the words in the above text are considered for calculating the lexicon and sentiment polarity.

1. Lexicon(again) = 0 and SentiP = 0.1
2. Lexicon(same) = 0 and SentiP = 0.4
3. Lexicon(problem) = -1 and SentiP = -0.6
4. Lexicon(right) = 0 and SentiP = 0
5. Lexicon(speaker) = 0 and SentiP = 0
6. Lexicon(work) = 1 and SentiP = 0.2
7. Lexicon(properly) = 1 and SentiP = 0.6

Sum of Lexicon = 1 and Sum of SentiP = 0.7

Value to be appended at the end of TFIDF matrix is 0.7. (Product of lexicon sum and SentiP sum)

MODELS-

Once the embedding matrix is generated using the above method, we trained our below models using 10fold cross validation on both (food and laptop) datasets and calculated the respective accuracy, F-score, precision and recall for positive, negative and neutral class.

1. Multinomial Naïve Bayes
2. CNN
3. SVM

EXPERIMENT RESULTS-

Laptop reviews data:

Classification Technique	Accuracy	Positive Precision	Positive Recall	Positive F Score	Negative Precision	Negative Recall	Negative F Score	Neutral Precision	Neutral Recall	Neutral F Score
Naïve Bayes	0.6532	0.7539	0.7012	0.7181	0.7288	0.6651	0.6941	0.4478	0.5814	0.4925
CNN	0.6012	0.6945	0.7230	0.7085	0.6235	0.6306	0.6271	0.3780	0.3369	0.3563
SVM	0.7539	0.7909	0.8019	0.7964	0.7250	0.8129	0.7665	0.7357	0.5764	0.6464

Food reviews data:

Classification Technique	Accuracy	Positive Precision	Positive Recall	Positive F Score	Negative Precision	Negative Recall	Negative F Score	Neutral Precision	Neutral Recall	Neutral F Score
Naïve Bayes	0.6232	0.1077	0.7012	0.6785	0.5119	0.4923	0.5132	0.3397	0.5301	0.4135
CNN	0.5833	0.604	0.9953	0.7518	0.2388	0.0148	0.0277	0.5666	0.0157	0.0304
SVM	0.7228	0.7931	0.8022	0.7979	0.6618	0.6739	0.6678	0.6364	0.6088	0.6223

CONCLUSION-

From the above experimental results, we can state that SVM performs better on both the data set compared to Multinomial Naïve Bayes and CNN. Additionally, the project also makes us realize the importance of feature engineering, which is a crucial part when it comes to text classification.

FUTURE WORK-

1. CNN can still be tuned to perform better with more layers, filters and reworking the weighting layer.
2. Implementing resampling which will definitely improve precision, recall and F1 scores, since the data is highly imbalanced.
3. Come up with a better word embedding technique, which involves domain-based lexicons.
4. Try Recurrent neural network and LSTM.

REFERENCES –

1. Amlaan Bhoi, Sandeep Joshi. Various Approaches to Aspect Based Sentiment Analysis
2. Bhumika Jadav, Vimalkumar Vaghela. Sentiment Analysis using Support Vector Machine based on Feature Selection and Semantic Analysis.