**Dataset used:**

The dataset used for aspect based sentiment analysis is taken from SemEval-2016 Task5 which was mentioned in the research paper. The dataset is available for different domain in different languages. We took Consumer electronics dataset in English language. The dataset file is in XML format. There are multiple review ids and each review contains multiple sentences. Each sentence further contains a text body, and multiple opinions. Each opinion in itself contains a target, category and polarity.

The target refers to phrase in the sentence for which polarity is given.

The category tag contains entity-attribute pair which is nothing but the aspect of a sentence.

Polarity is actually the sentiment of the sentence. We have 3 kinds of polarities here:

Positive: that means review is good.

Negative : review was bad.

Neutral: overall sentiment of the sentence is neither good nor bad.

**Data Preprocessing:**

Firstly the data we were given is in XML format, so our very first task is to parse the data out. For that we used Element Tree parser. After parsing, the textual content of the data is tokenized using spacy tokenizer and further refined which involves conversion into lower-case and punctuation removal. We have used glove embeddings to represent each token into it’s vector space. For some of the tokens of the parsed textual content, glove embeddings were not available, so we classified them into OOV tokens i.e. Out of Vocabulay tokens. These tokens mostly contains mis-spelled words or words which are used in social media chats and for whom embeddings are not available in English dictionary. We tried to correct most of the mis-spelled words using the python’s built-in speller which auto-corrects the tokens given the context. This way we reduced the number of OOV tokens from around 250 to 50!

**Data Representation:**

The data representation involves the input format in which the data is represented so that it becomes acceptable by our architecture. Our architecture has LSTM infact a bi-LSTM which takes fixed size inputs. So we fixed the number of sentences in each review to 35 and the sentence length to 100. We used ‘pad’ keyword to accommodate this and polarity label as 0 for the sentences which contains only ‘pad’ tokens. For the tokens which are classified as Out of Vocabulary, we gave them ‘oov’ token. Glove embedding’s are used to represent each token into it’s vector space and an integer token is assigned to each token which will be later used to map token to embeddings. Also the category tag in our XML data is actually the entity-attribute pair and the aspect vector is defined as mean of the entity and attribute as mentioned in the research paper.

And one o=more thing to add is that for the cases where sentence\_text, category are same but polarity is different for a particular review, we took average over polarity of these sentences, and used for our case.