Product Design Improvement System using Aspect Based Sentiment Analysis of Consumer Reviews

Jayanth Bhargav, Kanika Nadkarni, Nikhil Motwani, Vikas Shankarathota
School of Engineering and Applied Sciences, University of Pennsylvania

Abstract

An important aspect of Product Development and Design is taking into consideration consumer feedback in the form of reviews. In the proposed project aspect-based sentiment analysis of consumer reviews is carried out in the domain of laptops to draw insights on various customer opinions that can be used to upgrade designs. SVM, Multinomial Naive Bayes, Stochastic Gradient Descent and Neural network models are built under the Supervised Learning approach. Going ahead, an unsupervised learning approach is developed using Word Embeddings like Word2Vec / Pre-Trained BERT to predict aspects based on a particular type of clustering. Sentiment prediction is done using Lexicon based approach in the unsupervised predictor. These models are tested and their extensions to real-world applications in terms of scalability is discussed.

1 Introduction

Sentiment analysis or opinion mining is the computational study of people's opinions, sentiments, evaluations and emotions. It is one of the most active research areas in Natural Language Processing. A huge amount of written data is generated daily through e-commerce websites and social media. This data has valuable information and user opinions on various goods and services. Sentiment analysis is a key tool for making sense of that data. By opinion mining, companies can get key insights and user feedback for their products and this can help the companies to make better designs and improve their products in future releases.

In this paper, Aspect Based Sentiment Analysis is carried out on Laptop Dataset. Consumer reviews data in the domain of laptops is used to develop models to draw insights on various customer opinions that can be used to upgrade designs. The raw text data of the reviews are pre-processed to create features. Word embeddings are used to project the words into a vector space which enables to capture static or contextualized meanings of the words.

Both supervised and unsupervised/semi-supervised approaches are developed and evaluated. The motivation for developing an unsupervised approach is to eliminate the need for tagging large amount of reviews. The evaluations of the models on the same test data is done to draw insights on different training strategies and their real-world applications with respect to scalability.

2 Literature Survey

(Alghunaim et al. (2015)) target three sub-tasks namely aspect term extraction, aspect category detection, and aspect sentiment prediction. The effectiveness of vector representations over different text data is investigated and evaluation of the quality of domain-dependent vectors is done.

(Ma et al. (2018)) augment the long short-term memory (LSTM) network with a hierarchical attention mechanism consisting of a target level attention and a sentence-level attention. Commonsense knowledge of sentiment-related concepts is incorporated into the end-to-end training of a deep neural network for sentiment classification.

(Sun et al. (2019)) shows the potential of using the contextual word representations from the pre-trained language model BERT, together with a fine-tuning method with additional generated text, in order to solve out-of-domain ABSA.

(He, Ruidan, et al. (2017)) have implemented unsupervised neural attention model for aspect extraction. The models are trained on Restaurant Reviews dataset and have reported F1 scores of about 0.7 to 0.8.

The above research findings give an overview of the state of art models and benchmark accuracies in the domain of Aspect Based Sentiment Analysis (ASBA).

3 Scope and Objectives

The main objective of this project is to develop a model that can almost accurately predict an aspect and the consumer's sentiment towards it with minimal supervision. A few hypotheses have been developed during the course of carrying out the project:

i. Supervised Models perform better than Un-supervised Models

ii. With efficient training strategies and vast data, un-supervised algorithms can also reach considerably high accuracies.

iii. Contextual word embeddings perform better than Static Word embeddings

4 Data Extraction and Processing

Consumer Reviews on Laptops are used for building the models. The main challenge faced was to find tagged dataset for supervised training. The following datasets are used:

Supervised Training:
SemEval Task 5 2016 Tagged Dataset
Datapoints: 2500 Training Samples and 809 Testing Samples
Reviews are broken to separate sentences.
Number of Aspects: 59
Number of sentiments: 3

Unsupervised/Semi-Supervised:
amazon.com Scraped data 2014-2018
Reviews are broken to separate sentences.
Datapoints: 40,100 Training Samples and 200 Testing Samples
Training samples are unlabeled, and Testing Samples are Labeled.

Pre-Processing:

Word vectors are positioned in the vector space such that words that share common contexts in the corpus are located close to one another in the space. In the case of word2vec, the reviews have been lemmatized and stop words have been removed.

BERT, which stands for Bidirectional Encoder Representations from Transformers is designed to pretrain deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. Preliminary analysis was done to find out that BERT performs better when stop words are not removed due to the fact that it allows more context to be identified in the sentence. Thus, when using pre-trained BERT embeddings, the review text is taken as it is.

5 Supervised Training

The problem statement is divided into two parts: (i) Predicting aspects and (ii) Predicting sentiments of each aspect in the review.

The general pipeline of supervised training is shown in Figure 1.

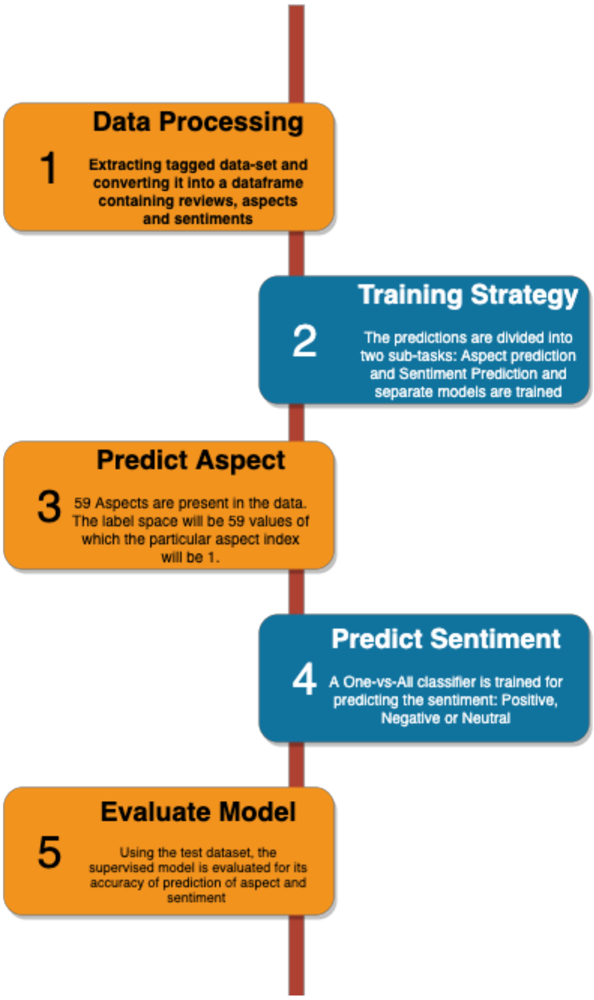
Data Processing
1
Extracting tagged data-set and converting it into a dataframe containing reviews, aspects and sentiments
Training Strategy
2
The predictions are divided into two sub-tasks: Aspect prediction and Sentiment Prediction and separate models are trained
Predict Aspect
3
59 Aspects are present in the data. The label space will be 59 values of which the particular aspect index will be 1.
Predict Sentiment
4
A One-vs-All classifier is trained for predicting the sentiment: Positive, Negative or Neutral
Evaluate Model
5
Using the test dataset, the supervised model is evaluated for its accuracy of prediction of aspect and sentiment

Figure 1: Supervised Training Flowchart

5.1 Aspect Prediction using classifiers:

Four models, namely, L_1 loss SVM, L_2 loss SVM, Multinomial Naïve Bayes and Stochastic Gradient Descent Classifier are trained.

The feature space X is of dimension $R_{|V|}$ where |V| is the vocabulary size. A TF-IDF model is created in which each sentence represents TF-IDF scores of each token that is present in the sentence. The label space y is of dimension R_{59} where y[i] =1 if the aspect 'i' is specified for the review in the dataset.

As this is a multi-class classification problem, One-Vs-All approach is used in which a classifier is trained for each aspect a_i, to predict whether the sentence contains the aspect a_i or not

where i = {1, 2, 3, …. 59}

In this way there are 59 classifiers, one for each aspect a_i and each classifier c_i would predict whether the aspect a_i is present in the review or not.

5.2 Sentiment Prediction using classifiers:

For predicting sentiments, the same feature space that was used for training aspects is utilized. Sentiments have three values namely: positive, negative and neutral. A One-vs-All approach is used to train a classifier for each value of the sentiment, that is: there are three classifiers: one for positive, one for negative and another for neutral sentiments.

To train a classifier for predicting positive sentiments, the label space y is of dimension R_{59} where y=1 if sentiment is positive or 0 otherwise. Similarly, negative and neutral sentiment classifiers are developed.

5.3 Neural Networks:

Two neural network models were built and trained, one for predicting an aspect in the sentence and another for predicting the sentiment of the aspect.

5.3.1 Neural Network model for predicting an aspect in the sentence:

The assumption was that the sentence would have only one aspect in it. So, a single (sentence, aspect) pair is being trained on this model. Sentences are converted into TD - IDF scores of each token in the sentence. Aspect categories are one-hot encoded and hence the label space would be R_{59}.

The architecture of the model is as follows:

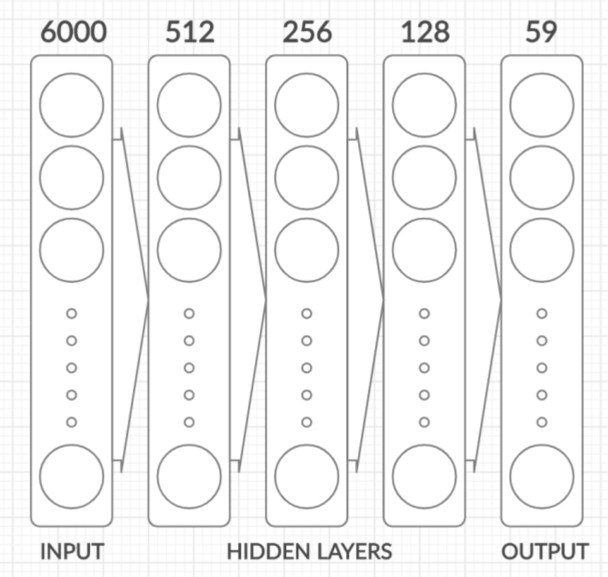

Figure 2: NN Architecture for Aspects

The model has an input layer of 6000 units, three hidden layers are present in the model with 512, 256 and 128 neurons respectively. ReLu activation is being used in the hidden layers. Then there is an output layer of size 59 (number of aspects) with SoftMax activation. The model is trained on 100 epochs.

5.3.2 Neural Network model for predicting sentiment in the sentence:

There can be positive, negative and neutral sentiments in the dataset. The feature space would be the TD-IDF scores of the sentences. There can be positive, negative and neutral sentiments in the dataset. The feature space would be the same as used in the neural network model above.

The architecture of this neural network model is similar to the architecture of neural network model for predicting an aspect, the only difference is that there would be 3 nodes in the output layer as the sentiment has only three values, namely, positive, negative and neutral.

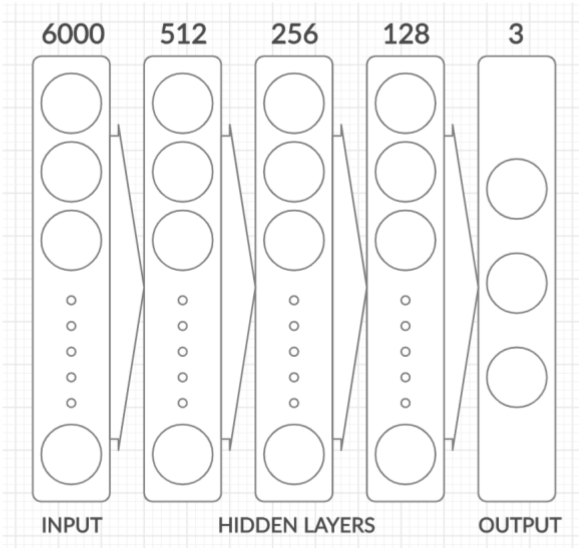

Figure 3: NN Architecture for Sentiments

The above models are trained for 100 epochs each and later tested using the test data. The model performances are discussed in the results section of the paper.

6 Unsupervised Approach using Pre-Trained Embeddings

The unsupervised model with pre-trained embeddings aims to learn a set of average word embeddings for each a user specified number of aspects. These average word embeddings will be learnt by the model and will be used in the prediction process. The model will also output the top words that are most closely related to the average aspect embedding. Looking at these top words, several aspects have been defined as follows:

Ergonomics – Relating to the comfort of using the laptop and it's peripheral components like the keyboard, mouse, etc.

Performance – Refers to the aspect of the review which deals with the hardware specifications and how well the laptop performs.

Build Quality – Related to customers talking about how well the laptop is put together and its durability

Nan – No major aspect identified

Service – Relating to the customers sentiment about the customer support and service-related issues.

Money – This aspect term is used to identify reviews that are related to the cost, discount, or value for money the user has spoken about in the review.

Usability – The term usability is used to classify reviews which talk focus on the ease of use of the laptop under consideration.

The model returns the average aspect embedding matrix, along with the top n words most closely related to the average aspect using cosine distance.

Each input sample to Attention Based Aspect Extraction (ABAE) is a list of indexes for words in a review sentence. Filtering of Non-aspect words is done by down-weighting them using an attention mechanism and then constructing a sentence embedding is done from weighted word embeddings. Sentence embeddings are reconstructed as a linear combination of aspect embeddings from T: the aspect embedding matrix. This process of dimension reduction and reconstruction, where ABAE aims to transform sentence embeddings of the filtered sentences (z_s) into their reconstructions (r_s) with the least possible amount of distortion, preserves most of the information of the aspect words in the K embedded aspects.

We first construct a vector representation of the sentence. For each word w_i in the sentence, we compute a positive weight a_i which is the probability that the word w_i leads to finding out the aspect in the sentence. The weight a_i is computed by an attention model, which is conditioned on the embedding of the word e, w_i as well as the global context of the sentence.

Reconstruction of Sentence – The reconstruction is a linear combination of aspect embeddings from T: where r_s is the reconstructed vector representation, p_t is the weight vector over the aspects. p_t can simply be obtained by reducing z_s from d dimensions to K dimensions and then applying a SoftMax non-linearity that yields normalized non-negative weights.

'w' is the weighted matrix parameter, and b is the bias vector: which are learned as part of the training process. The general pipeline for Unsupervised training is depicted in Figure 4.

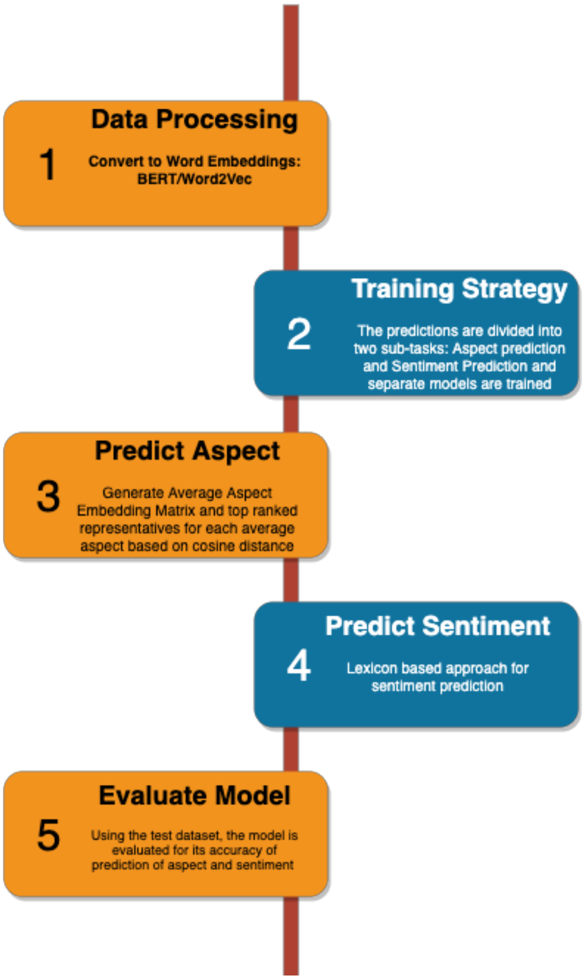
1
Data Processing
Convert to Word Embeddings: BERT/Word2Vec
2
Training Strategy
The predictions are divided into two sub-tasks: Aspect prediction and Sentiment Prediction and separate models are trained
3
Predict Aspect
Generate Average Aspect Embedding Matrix and top ranked representatives for each average aspect based on cosine distance
4
Predict Sentiment
Lexicon based approach for sentiment prediction
5
Evaluate Model
Using the test dataset, the model is evaluated for its accuracy of prediction of aspect and sentiment

Figure 4: Unsupervised Training Flowchart

Training Process: ABAE is trained to minimize the reconstruction error. For each input sentence, we randomly sample m sentences from our training data as negative samples. We represent each negative sample as n_i which is computed by averaging its word embeddings. The objective is to make the reconstructed embedding r_s similar to the target sentence embedding z_s while different from those negative samples. Therefore, the unregularized objective J is formulated as a hinge loss that maximizes the inner product between r_s and z_s and simultaneously minimize the inner product between r_s and the negative samples:

Regularization and Loss Functions: Hinge loss to maximize the inner product between encoded representation and reconstructed representation of positive samples and minimize the same for negative samples. Regularization terms were added to encourage uniqueness for aspect embeddings.

Sentiment Prediction: For each review sentence the positive, negative, or neutral sentiment associated with it must be predicted. This is also done without using tagged data as it prevents the need for human intervention. The approach to do this has been adapted from Hutto, C.J, & Gilbert, E. (2014, May) [5]. The approach uses a combination of qualitative and quantitative methods, to first construct and empirically validate a gold standard list of lexical features (along with their associated sentiment intensity measures) which are specifically attuned to sentiment in microblog-like contexts. These lexical features are combined with consideration for five general rules that embody grammatical and syntactical conventions for expressing and emphasizing sentiment intensity.

Tagged dataset is used to evaluate the performance of the unsupervised model developed.

7 Results and Discussions

The accuracies of the supervised training models are plotted for neutral, positive and negative sentiments. It is found that while Positive and Negative sentiments are being classified with an approximate accuracy of 70%, the neutral sentiment has a lower accuracy of about 50% only. These results are on a labelled training set of around 2500 datapoints. However, with larger data points and more efficient state of the art models using RNN, CNN etc. accuracies of labelled training have reached 90%. However, with a simple 3-layer Feedforward Neural Network the Aspect Prediction Accuracy has reached 71.5%. The state-of-the-art models have achieved around 90% with LSTM and BERT-pairs using auxiliary sentence. [1]

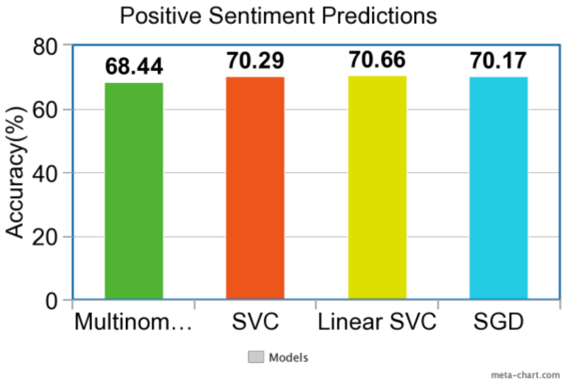

Figure 5: Positive Sentiment Accuracies

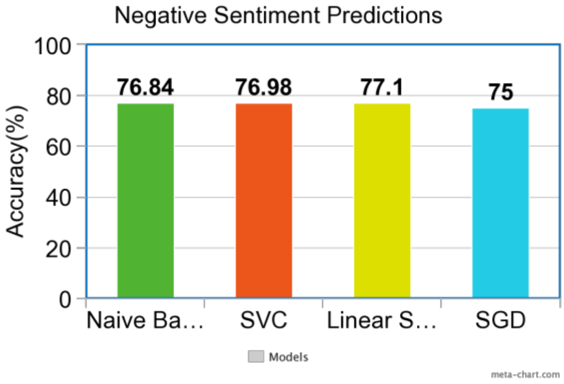

Figure 6: Negative Sentiment Accuracies

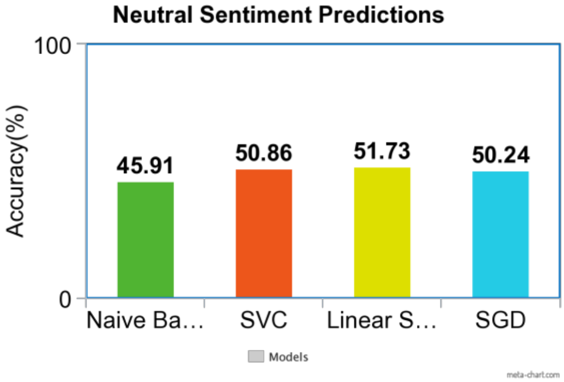

Figure 7: Neutral Sentiment Accuracies

The accuracies of aspects are summarized in the Table 1.

Model	Accuracy
L1 SVM	0.4888
L2 SVM	0.4876
Naïve Bayes	0.4282
SGD	0.4764
Neural Nets	0.3650

Table 1: Aspect Prediction Accuracies

The accuracies are not up to the state-of-the-art models as the network architectures and features used are not complex enough and well adapted to the domain.

In the unsupervised training, it is found that BERT Embedding performs better than Word2Vec due to contextualized nature.

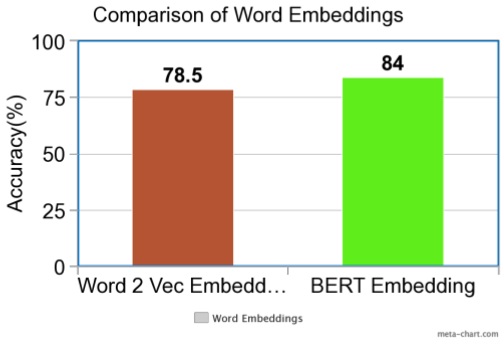

Figure 8: Comparison of Word embeddings

The Figure 9 depicts the learning curve of the Unsupervised model on train data which indicates that more training data improves accuracy.

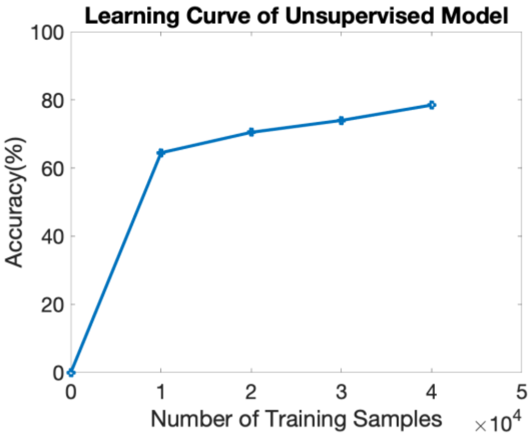

Figure 9: Learning Curve of Unsupervised model

The model has a sentiment prediction accuracy of 82.5%, which could be improved with domain specific lexicon.

8 Conclusions

The supervised models have reached accuracy levels of around 90% in state-of-the-art models using LSTM/RNN architectures. The main goal of this project was to develop supervised models and eventually focus more on unsupervised training techniques to eliminate the need to label vast amount of data. The aspect prediction accuracy has reached around 80% which is 10% lesser than state-of-the-art supervised models. Considering the arduous task of tagging enormous datasets, it would be feasible to have an unsupervised model with a reasonable trade-off in prediction accuracy which is practically scalable. Further improvements on the unsupervised approach can be made by domain adapting the word embeddings.

References

[1] Sun, Chi, Luyao Huang, and Xipeng Qiu. "Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence." *arXiv preprint arXiv:1903.09588* (2019).

[2] MA, Y.; PENG, H.; CAMBRIA, E.. Targeted Aspect-Based Sentiment Analysis via Embedding Commonsense Knowledge into an Attentive LSTM. AAAI Conference on Artificial Intelligence, North America, apr. 2018.

[3] He, R., Lee, W. S., Ng, H. T., & Dahlmeier, D. (2017, July). An unsupervised neural attention model for aspect extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 388-397).

[4] Hutto, C. J., & Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Eighth international AAAI conference on weblogs and social media.