SENTIMENT ANALYSIS ON AMAZON FINE FOOD REVIEWS.

Utkarsh Singh
Information Science and Engineering
R V College of Engineering
Bangalore, India
utkarshsingh.is17@rvce.edu.in

Sagar Raju Shet
Information Science and Engineering
R V College of Engineering
Bangalore, India
sagarrajushet.is17@rvce.edu.in

***Abstract*— Sentiment analysis or opinion mining is the machine study of people's opinions, sentiments, attitudes, and emotions expressed in communication. It has been one of all the foremost active analysis areas in Natural language processing and text mining in recent years. A Kaggle Challenge is hosted,where dataset is organised by SNAP team of Stanford University. The Challenge is about classifying the reviews given by customers on the fine foods. Classifiers based on supervised machine learning algorithms are used to classify the sentiment present in a tweet like Support Vector Machine, Logistic Regression and Random Forest.**

Keywords*— Sentiment analysis, Natural Language Processing

I. INTRODUCTION

Sentiment Analysis is an important task in NLP. Its purpose is to extract a single score from text, which makes it more convenient to analyze a large corpus of text. Various methods have been used to solve sentiment analysis problems, including bag-of-words and n-grams, and the arrival of deep learning, especially recursive neural network, provides a novel and powerful way to extract sentiment from text data.

From Sentiment classification, we can analyze the subjective data within the text and so mine the opinion. Sentiment analysis is that the procedure by which data is extracted from the opinions, appraisals, and emotions of individuals with reference to entities, events and their attributes. In deciding, the opinions of others have a major impact on customers ease, creating decisions with regards to on-line looking, selecting events, products, entities. The approaches of text sentiment analysis generally work a specific level like phrase, sentence or document level. Analyzing a review for the sentiment classification at a fine-grained level, specifically the sentence level during which polarity of the sentence may be given by three categories as positive, negative and neutral.

II. IMPLEMENTATION

A. Dataset used

The dataset was obtained from Kaggle. The Fine Foods datasets consists of 568,454 reviews between October 1999 and October 2012; 256,059 users and 74,258 products. The 52268 reviews have a score of 1, 29769 reviews have a score of 2, 42640 reviews have a score of 3, 80655 reviews have a score of 4, and 363122 reviews have a score of 5. The dataset was split into train and test set in the ratio of 80:20.

B. Data Inspection and Format

The data format is as follows:

product/productId: B001E4KFG0

review/userId: A3SGXH7AUHU8GW

review/profileName: delmartian review/helpfulness: 1/1

review/score: 5.0

review/time: 1303862400

review/summary: Good Quality Dog Food review/text: I have bought several of the Vitality canned dog food products and have found them all to be of good quality. The product looks more like a stew than a processed meat and it smells better. My Labrador is finicky and she appreciates this product better than most.

C. Data Cleaning

Random Under sampling aims to balance class distribution by randomly eliminating majority class examples. This is done until the majority and minority class instances are balanced out. In this approach, we reduce the data from higher class (data with 4 and 5 rating) to match the data with lower class(data with 1 and 2 rating).

D. Data Preprocessing

Data preprocessing is also an important process in which you transform our data before applying the different models on dataset so that you can achieve better results. In this stage , you can perform certain steps on our reviews(text) as mentioned below :-

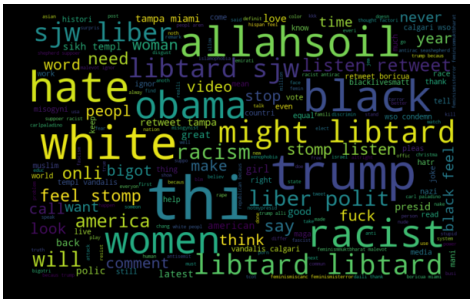
- Begin with defining your own function which will remove the html tags or we can use Beautiful Soup .

- Removal of punctuations or limited set of special characters like . or "," or removing whitespaces or etc from the reviews.
- Check if the word is made up of english letters and is not alpha-numeric.
- Check if the length of word is greater than 2 (as there is no adjective in 2 letters).
- uppercase to lower case conversion.
- Removal of Stopwords(commonly used word such as "a", "the" etc to be ignored).
- Stemming(Using Porter Stemmer) the word of each reviews.
- Tokenization is done for each review.

III. ALGORITHMS USED

To examine how well the given sentiments are distributed across the train dataset, 'word clouds' were plotted to understand the common words. A word cloud is a visualization wherein the most frequent words appear in large size and the less frequent words appear in smaller sizes. The word cloud obtained for both negative and positive tweets is shown below.

Positive



Negative

A. Extracting Features from Raw Dataset

Score of 1 and 2 usually denotes bad reviews whereas a score of 4 and 5 maps to good reviews. Since, rating 3 cannot be predicted, whether the user gave a positive or negative sentiment, we have taken it as negative.

B. Extracting Features from Cleaned Reviews

i) Word2Vec:Word2Vec is the current state of the art technique for feature extraction. Word2Vec is a shallow, two-layer neural network which is trained to reconstruct linguistic contexts of words.It takes as its input a large corpus of words and produces a vector space, typically of several hundred dimensions, with each unique word in the corpus being assigned a corresponding vector in the space.

Word2Vec model is trained on the cleaned reviews using gensim library.Word2Vec works on cosine similarity. If the cosine distance is less, the two words are similar in meaning.

A Dictionary is prepared with 'word' as key and its 'vector as value.

IV MODEL BUILDING

Each Dataset review is padded to length of maximum length of any review,and an embedding matrix is prepared for reviews vs word vector. The cleaned dataset is one-hot encoded and it is split into train,test as 80%,20% ratio.

- Ist layer is the embedding layer which embeds the word2vec and output shape is (1565,300).
- The second layer flattens the embedding to a 1D tensor, which allows to feed into a fully connected Dense Layer.
- The dense layer is a fully connected layer with 16 nuerons and uses ReLU as activation function.
- Dropout of 0.5 is used to improve generalization and prevent overfitting.
- Finally, the last layer is the output layer which does the binary classification in positive and negative classes and uses Sigmoid as activation function.

Model Summary :

Layer (type)	Output Shape	Param #
embedding_1 (Embedding)	(None, 1565, 300)	16914000
flatten_1 (Flatten)	(None, 469500)	0
dense_1 (Dense)	(None, 16)	7512016
dropout_1 (Dropout)	(None, 16)	0
dense_2 (Dense)	(None, 2)	34

Total params: 24,426,050
Trainable params: 24,426,050
Non-trainable params: 0

D. Analysis: -

The accuracy obtained in all the models implemented is roughly in the same range. Although the Neural Network Model implemented has an edge over the others. The Accuracy Obtained is 82.451

V. CONFUSION MATRIX

[108006 1]

[661 5684]]

	precision	recall	f1-score	support
0	0.93	1.00	0.96	8905
1	0.96	0.03	0.06	684
micro avg	0.93	0.93	0.93	9589
macro avg	0.94	0.52	0.51	9589
weighted avg	0.93	0.93	0.90	9589

The confusion matrix was obtained for logistic regression.

ACKNOWLEDGEMENT

We are grateful to Dr. B.M. SAGAR, HoD of ISE, for giving us an opportunity to explore such an interesting domain of natural language processing. It was a great learning experience. We would also like to thank everyone else who supported us and guided us through.

REFERENCES

[1] https://www.kaggle.com/snap/amazon-fine-food-reviews

[2] https://www.analyticsvidhya.com/blog/2016/07/practical-guide-data-preprocessing-python-scikit-learn/

[3] https://towardsdatascience.com/data-pre-processing-techniques-you-should-know-8954662716d6

[4] Tony Mullen and Nigel Collier, “Amazon Food Reviews Analysis ” by Stanford University (2010).