Sentiment Analysis Using LSTM

Sushama Khanvilkar[1], Chrisdesmond Pereira[2], Pavan Chaudhari[3], Akash Ate[4], Akshay Serrao[5]

Computer Department, Xavier Institute of Engineering, Mumbai, India

[1]sushma.k@xavierengg.com

[2]chrispereira997@yahoo.com

[3]pavanprojectbee@gmail.com

[4]akashate2@gmail.com

[5]akserrao@gmail.com

***Abstract*— In sentiment analysis and opinion mining; the thoughts, opinions, emotions, and sentiments of a person are analyzed and evaluated from the text. It is an important research field in natural language processing and is largely used in web mining, text mining and data mining. The focus of this field has greatly extended to businesses where the success of a particular product or brand largely depends upon knowing the customer's needs and satisfactions.**

The growth of sentiment analysis corresponds with the growth of social media, as people are nowadays connected to each other through social media fields such as reviews, micro-blogs, twitter, forum discussions, blogs and through other social networks. The greatest boon in the history, for the people of this generation, is that now we have machines with the capacity to store a large amount of data and to process on it (Big Data), thereby leading to major new discoveries that is changing the way we live. Python is a programming language developed by Guido Van Rossum, it is a largely implemented language for sentiment analysis as it has support for wide varieties of open source packages. Researchers use machine learning and deep learning concepts to solve many drawbacks of the existing models.

***Keywords*— Sentiment Analysis, Word Vectorization, LSTM, GloVe, Amazon Review Dataset**

I. INTRODUCTION

The internet plays an important role in learning about the product, or getting suggestions. Everyday millions of reviews are generated on the internet about an event, product, or a person. E-commerce websites generate a huge amount of reviews which on analysis show the current popularity of the product in the market. So these digital reviews play an important role in boosting up worldwide communications between consumers and regulating consumer buying patterns. Major companies corresponding to Amazon, Flipkart, etc. use sentiment analysis for gathering real insights referring to the performance of the products sold on their platform. The assortment of reviews into positive and negative sentiment is needed for extraction of valuable insights from a large collection of reviews. But understanding and handling of reviews become very difficult due to their huge number and size.

The field that identifies and derives the opinion of the given review is called sentiment analysis. The analysis involves natural language processing (NLP), computational linguistics, text analytics and classifying the polarity of the opinion to identify and extract subjective information from source materials. Classification of reviews into positive, negative, or neutral is accomplished by using neural networks in deep learning.

A. Aim and Objective

The main objective of this project is to develop a model to perform sentiment analysis on a set of reviews. This model can further help in decision making; as others impression have a substantial effect on customers ease in making choices with regard to online shopping, products, choosing events, entities etc. The objectives are:

i. To develop an accurate, efficient, and easy to use model for sentiment analysis.

ii. To provide an accurate sentiment analysis result which would help potential customers in the near future.

iii. To present and discuss the proposed algorithms that we will use for sentiment analysis.

II. RELATED WORK

With the overwhelming number of reviews produced each day, there are ways to analyze the sentiment behind these reviews, and many methods have been created for the depiction of the sentences through stacking RNNs, LSTMs, CNN and other complex architectures. For the representation of the sentence, Pang et al. [1] used a bag of words feature extraction method to predict the sentiment class of movie reviews. Their result shows that SVM is better than Naive Bayes, but not by a large margin.

In Ye et al. [2] research, they used three supervised machine learning models that are SVM, Naive Bayes and N-gram model to classify sentiment of reviews that were acquired from travel websites of popular destinations. The study found that N-gram and SVM performed better than Naive Bayes by a large margin.

In the previous few years, due to the state-of-the-art performance of deep learning models in pattern recognition and computer vision, these models are being implemented in performing NLP tasks and have shown better results in comparison to the commonly used machine learning methods for NLP [3].

Socher et al. [3], introduce Stanford Sentiment Treebank and RNTN which when trained on new tree bank outperforms all previous methods on several parameters used for evaluation in state-of-the-art-models, that classify the sentiment of a single sentence. In Sepp Hochreiter study, RNN has a problem when dealing with long sentences or long time lags, this problem is the vanishing gradient problem, and it is solved by using LSTM [4].

Dai and Lei [5] use a variety of datasets and perform document classification on it. They found that LSTMs initialized from scratch are not so good as compared to LSTMs pre-trained by recurrent language models or sequence autoencoders. Our work is using LSTM and the word-embedding performed by using the GloVe pretrained vectors [6].

III. DESIGN AND IMPLEMENTATION

A. Data Set

The initial stage is to collect data, and the dataset used is Amazon Product Reviews Dataset [7][8]. We are using cell phones and accessories category dataset to train the model. It has 194,439 reviews in total. After removing the duplicate reviews there are 194186 reviews. Fig. 1 shows the distribution of reviews in each category. From this data it can be observed that the reviews with score 1 and 2 are not distinguishable, and this same difficulty is present for reviews with score 4 and 5. Therefore, our solution is to label the reviews as positive, neutral, and negative by grouping score 1 and 2 as negative, score 3 as neutral and score 4 and 5 as positive. Fig. 2 shows the number of reviews after grouping them into positive, negative and neutral categories. After labelling the review scores, we notice that this dataset is quite imbalanced, i.e., about 76.4% of the reviews are positive. To solve this problem, we use undersampling technique. After dropping the reviews there are 85,384 reviews.

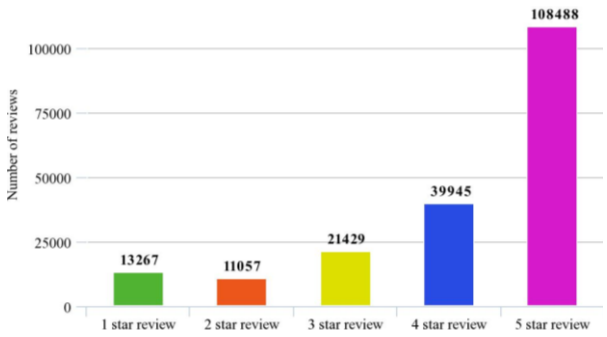

Fig. 1 Category wise number of the reviews.

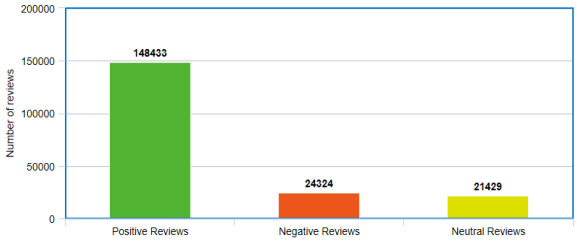

Fig. 2 Number of reviews in each category before dropping.

B. Data Pre-processing

From the dataset we used review text and overall feature as these two features are important to develop the model. Using python code other useless features were removed except for review text and overall rating. Before the reviews are used, they have to be tokenized. The term tokenization means the process of dividing a sentence or, a paragraph into meaningful elements such as phrases, words, etc. These tokenized words are then further considered separately and lemmatization is performed over them. Lemmatization is the mechanism of considering the exact form of a word so that they can be evaluated as a single unit i.e., it normally aims to remove the inflectional endings and just specifies the dictionary or base form of that word, which is called as the lemma. Furthermore, by using regular expressions, this data is refined.

C. Word Embeddings

In case of modelling of language and feature learning, a technique called word embedding is used. This technique maps words from a vocabulary into vectors of real numbers. This technique involves encoding of vectors in a multi-dimensional space. Here the closeness in vector space relates to how similar the words mean. In the embedding vector, all the latent features of a word show a particular dimension. During mapping of words to vectors, patterns and linguistic consistency is used. Based on the task, the optimal dimension is selected i.e., syntactic undertaking corresponding to named entity recognizing or POS tagging is more effective on using smaller dimensions, while with regard to further semantic tasks just as sentiment analysis is more pragmatic on adopting larger dimensions. Visualization of the learned vectors is performed by extruding them down to two dimensions which are done by the t-SNE dimensionality reduction technique [9]. Fig. 3 shows the visualization of vector representation of the words.

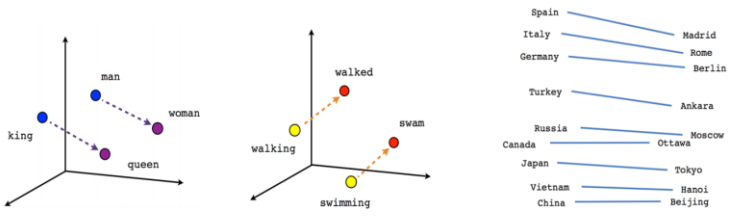

Fig. 3 Word Vector Representation.

D. LSTM

LSTM incorporates multiple gating functions in their state dynamics which addresses one of the most common problem in RNN called vanishing gradient problem. LSTM processes a sequence of inputs as pairs $(x_i, y_i) \cdots (x_z, y_z)$. For each pair (x_i, y_i) and at every time step t, an LSTM maintains a hidden vector h_t and a memory vector m_t which are responsible for controlling the state updates and outputs to produce a

target output y_i based on the past state of the x_i input (i.e., $x_1 \cdots x_i-1$). The computations at time step t are as follows:

$$\mathbf{g}^u = \sigma(\mathbf{W}^u * \mathbf{h}^t-1 + \mathbf{I}^u)$$
$$\mathbf{g}^f = \sigma(\mathbf{W}^f * \mathbf{h}^t-1 + \mathbf{I}^f)$$
$$\mathbf{g}^o = \sigma(\mathbf{W}^o * \mathbf{h}^t-1 + \mathbf{I}^o)$$
$$\mathbf{g}^c = \tanh(\mathbf{W}^c * \mathbf{h}^t-1 + \mathbf{I}^c)$$
$$\mathbf{m}^t = \mathbf{g}^f \odot \mathbf{m}^t-1 + \mathbf{g}^u \odot \mathbf{g}^c$$
$$\mathbf{h}^t = \tanh(\mathbf{g}^o \odot \mathbf{m}^t)$$

Where σ is the logistic sigmoid function, W^u, W^f, W^o, W^c are recurrent weight matrices and I^u, I^f, I^o, I^c are projection matrices. In contrast to the standard RNNs, the LSTM computed gates g^u, g^f, g^o , and g^c play a major role in learning important features out of the input/computed data by keeping the computed values as long as needed in the memory vector m_i. The g^f forget gate is able to remove components of the previous memory vector m_i whereas the gate g_c is able to write new content to the new memory vector m_i modulated by the input gate g^u. The output gate g^o controls what is then read from the new memory vector mi onto the hidden vector h_i [10].

E. Implementation

Keras provides an excellent high level abstraction for both multicore and GPU implementations, making the implementation of different network architectures fairly easy. All training experiments were carried out on Google Colab notebook that has CUDA 9.2 with an NVIDIA Tesla K80 graphics card. The dataset was split into 70% training set and 30% test set. The layers in the model consists of two LSTM and one hidden layer. From Fig. 4 we can observe that the model is overfitting after 5th epoch and the accuracy at that point is 66 percent. Fig. 5 shows the loss of the model on the test set.

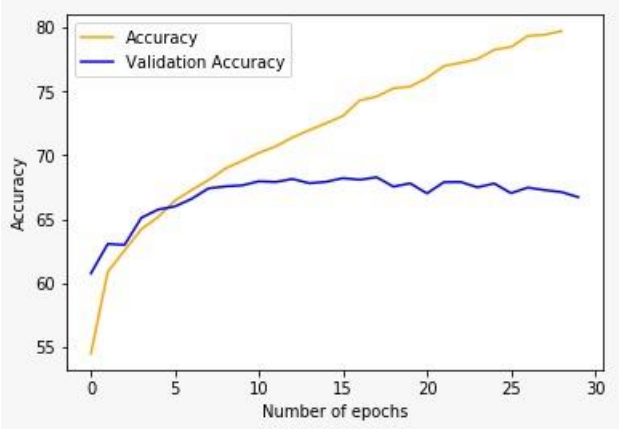

Fig. 4 Accuracy on the train and test set.

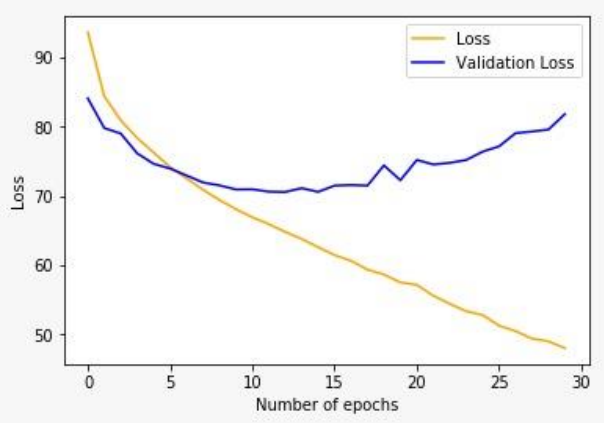

Fig. 5 Loss on the test set.

IV. RESULTS

The model is fed with the review from the user and the sentiment of the review is calculated. Fig. 6 shows the output for a positive review. Fig. 7 shows the output for a negative review. Fig. 8 shows the output for a neutral review.

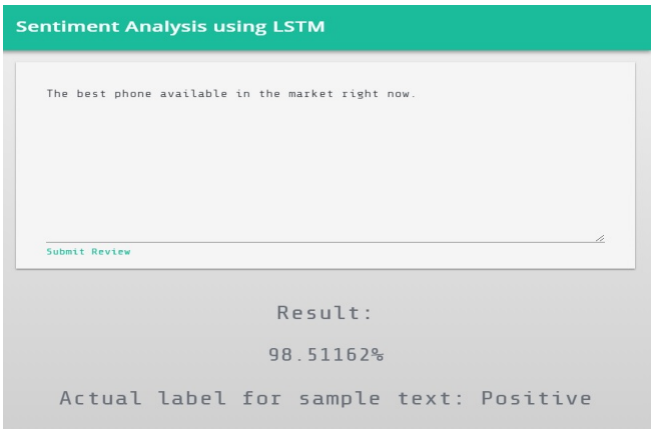

Fig. 6 Output for positive review.

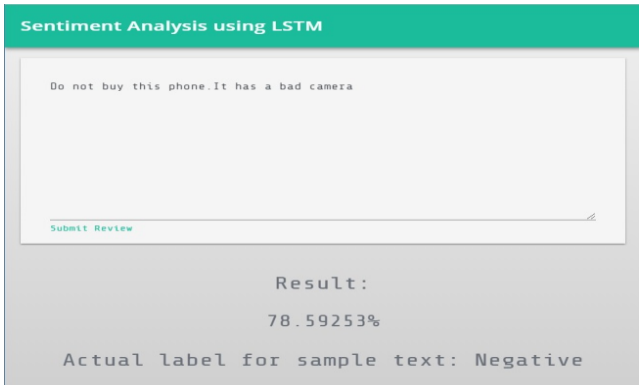

Fig. 7 Output for negative review.

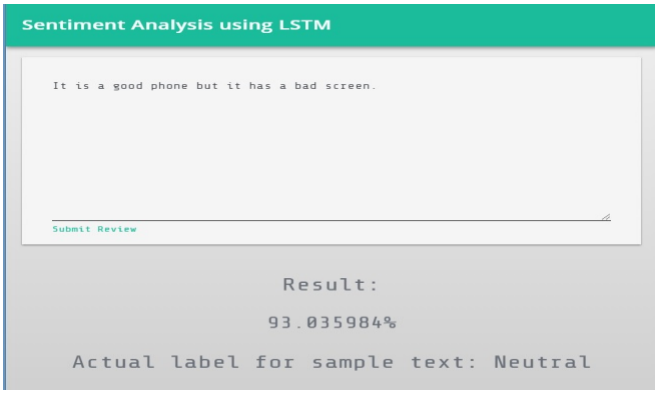

Fig. 8 Output for neutral review.

V. CONCLUSION

An evolutionary shift from offline markets to digital markets has increased the dependency of customers on online reviews to a great extent. Online reviews have turned into a platform for building trust and influencing consumer buying patterns. With such dependency there is a need to handle such large volume of reviews and present credible reviews before the consumer. In this project, we have studied various types of algorithms to perform sentiment analysis such as CNN, RNN and LSTM. NLP a field of computer science and AI is understood and varies preprocessing techniques has been implemented. These are then used to find the sentiment of a review. In the future, the work can be extended to perform intent analysis of reviews which will provide delineated intent of review to the consumer, hence better classification of the product.

ACKNOWLEDGMENT

We express sincerest gratitude to our project guide, Head of Department Sushama Khanvilkar, who has been a source of inspiration and has made it possible for us to initiate and carry on with this project.

REFERENCES

[1] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 79–86, 2002.
[2] Qiang Ye, Ziqiong Zhang, and Rob Law. Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. Expert systems with applications, 36(3):6527–6535, 2009.
[3] R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, C. Potts et al., "Recursive deep models for semantic compositionality over a sentiment treebank," in Proceedings of the conference on empirical methods in natural language processing (EMNLP), vol. 1631, 2013, p. 1642.
[4] S. Hochreiter and J. Schmidhuber. Long short-term memory. Neural computation, 9(8):1735–1780, 1997.
[5] A. M. Dai and Q. V. Le, "Semi-supervised sequence learning," in Proc. Advances Neural Information Processing Systems, 2015, pp. 30
[6] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation.
[7] Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering R. He, J. McAuley WWW, 2016.
[8] Image-based recommendations on styles and substitutes J. McAuley, C. Targett, J. Shi, A. van den Hengel SIGIR, 2015.
[9] L.J.P. van der Maaten and G.E. Hinton. Visualizing data using t-SNE. Journal of Machine Learning Research, 9(Nov):2431–2456, 2008.
[10] Al-Smadi, M., Talafha, B., Al-Ayyoub, M. et al. Int. J. Mach. Learn. & Cyber. (2018).