Development of Textual Analysis using Machine Learning to Improve the Sentiment Classification

Abha Kiran Rajpoot¹, Parma Nand², Ali Imam Abidi³
Sharda University, Greater Noida, India

¹akrajpoot@gmail.com, ²parma.nand@sharda.ac.in, ³ali.abidi@sharda.ac.in

Abstract. Rapid development in Internet and the increase in online information, the technology demanded for intelligently classifying the textual data has become significant role in Information Retrieval Process. Based on given query in the search box, the response from the internet has made open to the public. Thus, the scope of text mining is being explored by several researchers. Sentiment Analysis is one of the most popular process for analysing user opinions and feelings, and since, online communication has become the fast ever growing medium for expressing thoughts, therefore, there have been development in text classification to improve sentiment analysis. In this paper, some of the prior works on sentiment analysis and the advancements in text classification have been discussed.

Keywords - Text classification, Sentiment analysis, Deep learning, Text mining.

1. INTRODUCTION

Artificial Intelligence and Machine Learning are the growing stars of technology in today’s times. Their usage can be found everywhere, there might be any area, where the duo is not there. They have been bringing along different techniques and models that are very much effective and fast, providing optimized solutions to the given problem statements. Rapid development in Internet and therefore the increase in online information, has increased the demands and usage of one of the branches of AI that is, Natural Language Processing (NLP). With the help of NLP, machines can easily read and understand the human language which makes it a powerful tool to analyse and understand human’s sentiments and thoughts. Text classification is a commonly used NLP task in applications such as automatic news article categorization, product reviews etc. into well-defined topics and sentiment analysis [1].

In today’s times, online communication has become one of the most widely used practice by everyone to deliver thoughts and feelings about something to others. This rise has made text classifiers, one of the most important tools for people to track and organize information. Sentiment Analysis, a part of text classification is the most widely used process for all forms of online communications.

Due to its ability to understand the opinions and conclude the sentiments of the given statement, Sentiment Analysis can find its uses in the areas [2] like:

- Online Shopping: Purchasing behavior analysis based on prior reviews on products,
- Movie Reviews: Review about a movie (or) shows using Internet Movie Database (IMDb),
- Government Policies: Opinion mining on established policies,
- E-Learning Systems: Student’s opinion analysis to enhance the learning services,
- Recommendation Systems: It recommends the users based on their classification of reviews,
- Security Systems: It assists to detect the spams.

In this paper, we discuss some of the works related to sentiment analysis and some of the new developments in text classification techniques which are aiding to more efficient techniques of sentiment analysis in today’s times. The dissection of the paper is as follows: Section 2 talks about sentiment analysis and its process; Section 3 highlights some of the advancements in text classification methods; Section 4 discusses some of the prior works on sentiment analysis; finally, the paper concludes in Section 5.
2. SENTIMENT ANALYSIS

Sentiment analysis is defined as the process of extracting the information of opinions, views, and emotions from heterogeneous sources such as tweets and database sources via natural language processing system. Based on opinions given in text, the classes like ‘positive’ (or) ‘negative’ (or) ‘neutral’ will be categorized. This information analysis is known as subjectivity analysis, opinion analysis and inspection analysis. The words such as opinion, sentiment, view, and belief have difference in context [3]. They are defined as follows:

a) Opinion: Experts have different opinions i.e., dispute communication
b) View: Opinion based on subjectivity
c) Belief: Opinion based on acceptance and assents
d) Sentiment: Opinion represents the feelings

The process involved in sentiment analysis can be depicted from Fig. 1.

1. Data Collection. User sentiments can be found in public platforms like blogs, twitter and/or Facebook posts, YouTube comments, reviews etc. These sentiments are generally the opinions and feelings towards specific posts or things.

2. Feature Extraction. Feature extraction is an important process in defining the sentiment analysis.

3. Sentence Opinion. After feature extraction, every opinion is examined for its subjectivity i.e., only the subjective expressions are preserved while the objective expressions are removed.

4. Results. The final sentiments are classified further into two main groups called, positive (+) and negative (-). The sentiments can be good – bad or like – dislike. The results are then presented in the form of charts and graphs.

Fig. 1. Process of Sentiment Analysis

3. DEEP LEARNING BASED TEXT CLASSIFICATION

There are different text classification models based on machine learning. Some of these are Naïve Bayes Theorem, Logistic Regression, Decision Trees and Random Forests, SVM etc. However, the number of datasets and moreover, the complexity of the data is exponentially increasing day by day. This exponential complexity requires more developments in the traditional text classification methods to provide more vigorous and accurate results. Deep Learning is the growing machine learning method, initially started in 1990s [4], which is based on artificial neural networks. Unlike these previous ML
techniques, deep learning approaches follow the hierarchical architecture where the output of the lower layer becomes the input of the higher level i.e., these approaches are purely based on the datasets and not on the user’s training. Due to this approach, deep learning methods prove to provide much more accurate results than the existing models, when it comes on classifying large and/or complex data.

Sentiment Analysis combined with deep learning-based models make it a strong tool for understanding texts and data beyond simple definitions and determining the actual mood and/or feeling of the user. For example, we have sentence,

“This is a good update, but it works very slow and above all, it is not supported on the phone.”

This sentence shows a mixture of positive and negative comments. With deep learning, that is, with advanced networks, it is very easy for sentiment analyser tool to understand the sentiments correctly.

There are mainly 3 broad categories of deep learning models, which are Deep Belief Network, Recurrent Neural Network and Convolution Neural Network.

3.1 DBN (Deep Belief Network)
DBN is the most general text classifier. Its property of extracting unique features of the texts makes it a high learning text classifier. Some of the most common applications of DBN can be found in [5, 6].

In a general DBN model, there is mainly an input layer and an output layer, and in between there are many hidden layers, each containing Restricted Boltzmann’s Machine (RBM). That is why, the model is sometimes called as a stack of RBMs. The pictorial representation of the DBN model can be seen from Fig. 2. The input layer is the visible layer for the first hidden layer, and further each hidden layer becomes the visible layer for its subsequent layer. There are undirected connections between them but not within them unlike RNN. These machines are purely non-linear in nature, and therefore they produce patterns of high variations in the training data.

3.2 RNN (Recurrent Neural Network)
RNN is sometimes compared with the Frontal Lobe of a human brain; since like Frontal Lobe, RNN also uses its internal memory while processing the input data. Some of the commonly used application areas of RNN model can be found in [7, 8].

RNN basically works on sequence of data and perform classification based on the current and past observations. Their connections form a directed graph where one layer is connected to its previous layers which allow the data to flow back to the previous layers as well. The pictorial representation of the RNN model can be seen in Fig. 3. Thus, in this way, working on the past events and learning from the past events, the model gives out better semantic analysis of the text.

![General architecture of RNN](image1)

Fig. 3. General architecture of RNN

However, traditional RNN model seems less efficient though due to some of its weaknesses like vanishing gradient and hence it is integrated with LSTM (Long Short-Term Memory) [9]. LSTMs are a novel form of RNN that have the property of retaining a memory for longer durations of time and therefore provide better results when compared with traditional RNN-based models.

### 3.3 CNN (Convolution Neural Network)

Like RNN, CNN is considered as the heart of Computer Vision systems. CNN was originally being extensively used for the purposes of image classification. However, recent advancements show its use cases in NLP and text classification as well. Some of the application areas of CNN can be found in [10, 11].

![General architecture of CNN](image2)

Fig. 4. General architecture of CNN

The pictorial representation of the general CNN architecture can be seen in Fig. 4. The first layer is known as the Sentence Matrix where the input word is a 2D matrix represented as, s X d, where s is the length
of the sentence and d is the dimension of the word vector. The next layer is known as the Convolutional Layer. This layer produces feature maps for each word which are obtained after performing convolutions or filtering over the sentence matrix with the help of varying height filters known as filter matrix. The next layer is known as Max Pool where max pooling is executed over each feature map, generated in the previous layer, to get the largest number of that vector. After max pooling, the generated largest number of each vector is then concatenated forming a 1D feature vector, and this is received as an input for the next layer known as SoftMax Layer. The output vector received is used as a classifier for the sentence.

3.4 Some Hybrid Models
Considering the complexity and the size of data sets, there are other different deep learning models commonly known as hybrid models which are the combination of the traditional deep learning-based models. Some of these are:

3.4.1 R-BLS and G-BLS [12]. In this paper authors have proposed two novel models derived from Broad Learning System (BLS). Recurrent BLS is based on Recurrent Neural Network model and Gated BLS is based on Long Term Short Memory model. The analysis showed that the proposed methods proved to produce results with high accuracy and faster training time.

3.4.2 (DNN) model together. First, the system extracts semantic words with the help of word embedding model followed by word feature construction. For the classification, the system uses DBN model and Back-Propagation NN model.

3.4.3 Bi-LSTM + CNN [14]. This paper proposes a new hybrid model which integrates the benefits of Bi-LSTM and CNN along with an attention mechanism. The architecture consists of mainly 4 layers. First layer is responsible for text embedding, for which the model uses Word2Vec Technique. This is then passed to the next layer called Convolutional Layer where the convolution process or filtering takes place and 1D feature vector is received as an input to the next layer which is Bi-LSTM Layer. The attention mechanism is used to filter the data sequence produced by the previous layer for the final classification.

3.4.4 Using Capsules [15]. Capsule Networks are extensively used in the field of image processing. But now, capsules are also expanding to text classification as well. There has been different research on classifying texts with the help of capsule networks. Authors in this paper explore capsule networks to the field of text classification. The model consists of 4 layers. First layer is the input layer where text embedding takes place. Next layer is the Convolutional Capsule layer where feature vector is produced by convolutions of filters. Third layer is the Text Capsule Layer, where the comparison between Capsule Network with dynamic routing and static routing is shown.

3.4.5 Extended Capsule Networks [16]. This paper proposes a new version of CapsNets (stated previously). This is known as Interpretable Capsule Network. The proposed model consists of 4 layers. First Layer is the Embedding Layer, followed by the Convolutional Layer. The vector features obtained are passed to the Multi Head Attention Layer, where each vector produces one primary capsule. Each primary capsule represents one semantic meaning. These primary capsules are then passed to the final layer which produces high level semantic meanings using dynamic routing.

4. PRIOR WORKS ON SENTIMENT ANALYSIS

Prior works on sentiment analysis in view of text sources, the classification tasks pertain in the following three forms:
4.1 Document Level Sentiment Analysis

Every document comprises of opinions/opinionated words. classification for these documents is done in two approaches, namely, supervised and unsupervised learning. Document Level Sentiment Analysis or Document Sentiment Analysis basically gives out the overall opinion of the complete document, either a positive or negative, i.e., in this case, the whole document is considered as a single entity. Some of the recent prominent researches for Document Level Sentiment Analysis include:

a) Using Neutrosophic Set and Particle Swarm Optimization [17]
b) Using Hierarchical Interaction Networks with Rethinking Mechanism [18]

4.2 Sentence Level Sentiment Analysis

It operates on extracting opinion words from each sentence. Through this, every sentence is determined as either being a positive sentiment or negative sentiment or neutral sentiment. Some of the recent research for Sentence Level Sentiment Analysis include:

a) Using LeNFEM, that is, Lexicon-pointed hybrid N-gram Features Extraction Model [19]
b) Using Deep CNN and LSTM model [20]

4.3 Aspect Level Sentiment Analysis

Aspect/Attribute-based classification operates in situations related to customer reviews. Consider an instance where customer reviews of a mobile phone with attributes like battery capacity, camera-resolution, processor frequency etc. Some of the recent researches for Aspect Level Sentiment Analysis include:

a) Using Artificial Bee Colony algorithm [21]
b) Using Interactive Rule Attention Network [22]

4.4 Sentiment Lexicon Acquisition

It is a widely popular and critical resource for a large number of sentiment analysis algorithms prevalent today. This technique involves acquisition of the sentiment lexicon manually, using dictionary and corpus-based approaches. Some of the recent researches for Sentiment Lexicon Acquisition include:

a) Using Corpus-Generated Polarity Seed Words [23]
b) Using Training Optimization-Based Method [24]

5. CONCLUSION

Recent fields of technology is inclined and the data are dynamic in nature. Online communication and therefore sentiment analysis are growing parallelly and rapidly. However, sentiment analysis brings along many challenges as well. Issues like performance issues, manipulating of models, handling typographical errors, ambiguity, inferences, multi-lingual reviews etch are some of the challenges that sentiment analysis face/may face. In most cases, the incorrect judgements are easily given by classifiers. Therefore, classification needs to be improved for better prediction as well as recognition accuracy. Some advanced text classifier is required to analyze and interprets the sentiments of the users. With the help of deep learning-based approaches, abundant amount of information will be easily analysed. And even, unstructured data shall also increase the amount of training data and expand the field of data coverage.

REFERENCES

[1] Khan, S.U.R., ISLAM, M.A., Aleem, M. and Iqbal, M.A., 2018. Temporal specificity-based text classification for information retrieval. Turkish Journal of Electrical Engineering & Computer Sciences, 26(6), pp.2915-2926.
[2] Morbidoni, C., Cucchiarelli, A. and Ursino, D., 2018, June. Leveraging linked entities to estimate focus time of short texts. In Proceedings of the 22nd International Database Engineering & Applications Symposium (pp. 282-286).

[3] Jatowt, A., Yeung, C.M.A. and Tanaka, K., 2015. Generic method for detecting focus time of documents. Information Processing & Management, 51(6), pp.851-868.

[4] Dechter, Rina. "Learning while searching in constraint-satisfaction problems." (1986): 178-1

[5] Jiang, Mingyang, et al. "Text classification based on deep belief network and softmax regression." Neural Computing and Applications 29.1 (2018): 61-70

[6] Srilakshmi, V., K. Anuradha, and C. Shoba Bindu. "Optimized deep belief network and entropy-based hybrid bounding model for incremental text categorization." Internation-al Journal of Web Information Systems (2020).

[7] Buber, Ebubekir, and Banu Diri. "Web Page Classification Using RNN." Procedia Computer Science 154 (2019): 62-72.

[8] Liu, Pengfei, Xipeng Qiu, and Xuanjing Huang. "Recurrent neural network for text classification with multi-task learning." arXiv preprint arXiv:1605.05101 (2016).

[9] S. Hochreiter, “Long Short Term Memory,” Neural Comput., vol. 9, no. 8, pp. 1–32, 1997.

[10] Kim, Y. "Convolutional neural networks for sentence classification., arXiv." preprint (2014)

[11] Ce, Peng, and Bao Tie. "An Analysis Method for Interpretability of CNN Text Classification Model." Future Internet 12.12 (2020): 228.

[12] Novel Efficient RNN and LSTM-Like Architectures: Recurrent and Gated Broad Learning Systems and Their Applications for Text Classification - IEEE Journals & Magazine

[13] Wang, Tianshi, et al. "Bi-LSTM model to increase accuracy in text classification: combining Word2vec CNN and attention mechanism." Applied Sciences 10.17 (2020): 5841.

[14] Jang, Beakcheol, et al. "A multi-label text classification method via dynamic semantic representation model and deep neural network." Applied Intelligence 50.8 (2020): 2339-2351

[15] Jain, Amita, et al. "Senti-NSetPSO: large-sized document-level sentiment analysis using Neutrosophic Set and particle swarm optimization." Soft Computing 24.1 (2020): 3-15.

[16] Wei, Lingwei, et al., "Hierarchical Interaction Networks with Rethinking Mechanism for Document-level Sentiment Analysis." arXiv preprint arXiv:2007.08445 (2020).

[17] Mutinda, James, Waweru Mwangi, and George Okeyo. "Lexicon-pointed hybrid N-gram Features Extraction Model (LeNFEM) for sentence level sentiment analysis." Engineering Reports (2020): e12374.

[18] Nguyen, Huy, and Minh-Le Nguyen. "A deep neural architecture for sentence-level sentiment classification in twitter social networking." International Conference of the Pacific Association for Computational Linguistics. Springer, Singapore, 2017.

[19] Zhang, Ming, et al. "Attention-based word embeddings using Artificial Bee Colony algorithm for aspect-level sentiment classification." Information Sciences 545 (2021): 713-738.

[20] Lu, Qiang, et al. "Interactive rule attention network for aspect-level sentiment analysis." IEEE Access 8 (2020): 52505-52516.

[21] Sanagar, Swati, and Deepa Gupta. "Unsupervised Genre-Based Multidomain Sentiment Lexicon Learning Using Corpus-Generated Polarity Seed Words." IEEE Access 8 (2020): 118050-118071.

[22] Du, Maokang, Xiaoguang Li, and Longyan Luo. "A Training-Optimization-Based Method for Constructing Domain-Specific Sentiment Lexicon." Complexity 2021 (2021).