Volume 6, Issue 10, October 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

An Overview of RNN and CNN Techniques for Spam Detection in Social Media

[1]Gauri Jain*, [2]Manisha, [3]Basant Agarwal
[1, 2] Department of Computer Science, Banasthali Vidhyapeeth, Rajasthan, India
[3] Department of Computer Science & Engineering, SKIT, RTU, Jaipur, Rajasthan, India

Abstract— Spam detection is one area in which the researchers have been working for long. It started with email spam detection and now spam detection on social networking sites is well known research problem. With the rise in use of Internet right from sending email, posting reviews, tweeting status and live media posting, the spammers are trying to make their presence felt everywhere. This paper is an overview of Deep Learning Techniques: Recursive Neural Network (RNN) technique and Convolution Neural Network (CNN) for spam detection. These techniques have an advantage that they are able learn high level features on their own with the help of raw data unlike traditional machine learning classifiers that need hand crafted features for applying classification model.

Keywords— Spam detection, Machine learning, Deep Learning, RNN, CNN, NLP

I. INTRODUCTION

During the recent few years social media has evolved many folds and has become much more interactive, and integral part of our lives. The interaction channels in the social media have changed from traditional media like newspapers and television to mobile phones, social media websites, microblogging sites etc. It has changed the way people communicate with each other on the personal as well as on public from as described in [1]. There are variety of social media sites that offer diverse functionality, some are for common people like Facebook, which started as an experimental social network in the Harvard University by some students, while others like LinkedIn is a network formed by professionals from every field. Many sites are exclusively for sharing videos and pictures media like YouTube, Instagram, Flickr etc. while others focused on blogs where people from varied domains express and share their views. There are even social tagging and news sites like Reddit, Delicious etc. which allow the user to rank the websites on the basis of quality of content and usefulness of the sites. Most recent trend of micro – blogging let people update the real – time status of their daily routine or happenings via app like Twitter which has more than 200 million users exchanging more than 400 million tweets per day [2] where the length of tweets is limited to 140 characters. According to Teen, Social Media and Technology Overview 2015 [3], "More than 24% of the teen are constantly online and 71% of them use more than one social networking site". This ease of sending and receiving data over Internet has resulted in some notorious people sending unwanted messages to large number of recipients over the network trying to take advantage by getting access to their privacy. Initial spread of spams started with email spam. According to M^3AAWG report, the abusive email content amounts to 87.1% - 90.2% of the total email content during 2012 – 2014 [4] which has increased the financial burden by increasing the storage requirement and technological requirement for spam detection. Slowly spams started spreading in every digital media like from mobile network through mobile phone, social networking sites, blogs, review sites etc. An example of spam tweet on the famous microblogging site Twitter is shown in figure 1.

Fig. 1 Sample spam tweets

Due to the increase in the volume of spams message floating on social network, people are keen to find new and more effective ways in order to avoid them or detect them. For e.g. mobile carrier provides the facility to mark a SMS as spam and mobile number from which the message was sent would be added in the spammers list. In future, any message

originating from that number would be marked as spam and won't be delivered to our mobile number. Similarly, most of the social networking sites give an option to mark any message as spam and they would take suitable action like blocking the account if the claim is found to be true. Spam can also be detected while sending the messages at the server level. One such technique involves checking the url of the source of the message with the blacklist [5], [6], whitelist and greylist [7], [8]. Other approach to text classification is automatic spam detection using various techniques. It started with very simple techniques like deriving some simple rules based on the content of the text [9], [10], but this type of technique had a disadvantage of changing the rules with changing the media. Later with the use of machine learning technique, using statistical classifiers gave up to 98% accuracy in spam detection. Some of the significant work related to statistical classifiers were discussed in [11] – [13], where the work was carried out on detection of spam in blogs, Twitter and emails. Some of the main algorithm used for developing the spam detection model includes Support Vector Machine [14], Naïve Bayes [15], classifiers based on Decision Trees [16] and ensemble methods like Random Forest [17]. Inspite of giving good results, these methods has a disadvantage of hand crafted feature engineering on which the classification model works. Some of the feature extraction techniques were proposed in [18,19, 20]. Some hybrid methods were also suggested that used various content based heuristics along with the statistical methods. Some of these heuristics were word count, word count in the title of the page, average word length etc. which are described in [21]. Similar content based techniques were used along with the traditional classifier in [22] to detect spams on Twitter.

This paper gives an overview of the Deep learning models for the detection of spam in text data taken from various types of social media. Deep Learning technologies extracts high level feature automatically from the raw data while maintaining data independence. We have discussed in detail two main architectures: Recursive Neural Network (RNN) and Convolution Neural Network (CNN) in section III.

II. BACKGROUND AND RELATED WORK

The work of deep learning came into light with [23] in 2007 but it became very popular with [24] in which model was trained with unlabelled pictures for face detection. LeChun et al. gave a detailed description about deep learning and one of its main algorithm Convolution neural network (CNN) in [25] and how CNN is helpful in processing large datasets which a normal neural network fails to do. Bengio discussed in detail the data representation of abstract features along with the difficulty in training of deep networks in [26]. In an overview given by Li Deng [27] he summarized various architectures, algorithms as well as application related to Deep Learning. This technology is very effective in the field audio/speech, computer vision, language modelling, NLP problems like sentiment classification [28] etc. and the main algorithms of deep learning were Stacked Autoencoders, deep stacking networks like Deep Belief Networks (DBN). Hochreiter proposed Recurrent Neural Network (RNN) which is one of the main architecture of deep network and is useful for supervised learning with the help of learning previous information. One such work was given by Sutskever in [29] where textual sequence was generated based on previous text knowledge. Other main application of RNN is in the area of speech recognition [30] – [32].

Another technique that used deep network is convolution neural network (CNN) which was initially used in the area of computer vision for applications like face detection [33], handwritten character detection [34], handwritten digit detection [35] etc. Recently, the researchers have started using CNN for processing of text and the performance of CNN has given good results especially in the area of sentence classification [36], sentiment analysis of text data [37] and word embedding [38] and sentence modelling [39]. The study in these area have motivated the use of CNN for spam detection since the feature extraction using these technique is independent of the data being processed.

III. DEEP LEARNING APPROACH

Deep Learning also known as hierarchical learning, a progressive and promising technology in the area of Machine Learning, which has taken many steps forward towards the original goal of machine learning that is imitating human brain's intelligence. This technique allows the computer to learn and retain real world knowledge, while learning from real world examples and making subjective decisions. Deep Learning is about learning representations of data in an abstract way which helps in extracting features from the variety of medium such as images, voice and text data.There are many applications already taking advantages in the field of computer vision, language modelling, NLP, topic recognition, etc. The popularity of deep architectures have increased recently due to following:

- The recent developments in the field of hardware abilities as well as software advancements in the form of *high – speed general purpose graphical processing units* (GPGPUs) for processing and new tools and algorithms in the form of software.
- As the volume of data available for training is increasing day by day due to massive use of Internet especially social media, deep learning is becoming more and more important.
- Deep Learning has proved its capabilities by solving many complex applications, which were earlier not touched or limited success was achieved due to technological limitations.

A. Recursive Neural Network -

Recurrent Neural Network (RNN) is one of the popular Deep Learning Architecture that is a special type of feed forward neural network. Like human brains, RNNs are able to process and understand information that has occurred in past i.e. in order to understand the current problem in hand, it looks for the related information in the past. It designed for problems that consists of sequence of input that are related to each other. In a RNN, same processing is applied to each sequence of input with output being dependent on the previous computations and output. After each sequence of input

being processed, an output is generated that might help in the output at the next layer. This is done with the help of "memory" that captures the information at each level. In figure 2, a RNN is shown having a memory *s* with s_t being the memory contents at time t.

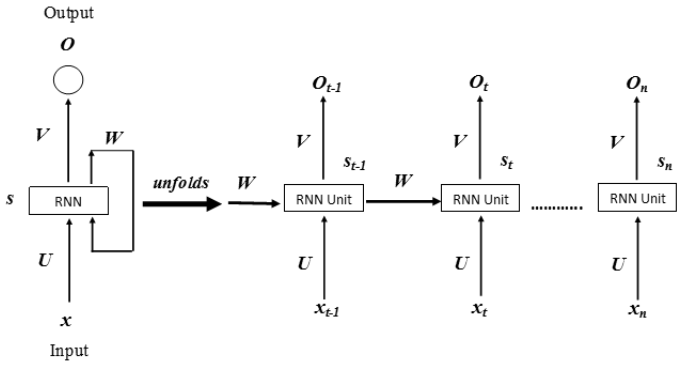


Fig. 2 Basic RNN Structure

The basic working of RNN is as follows and is shown in figure 3:

- x_t is the input sequence at time t in the form of word vector. Word vector can be 1 hot vector or it can be taken from the Google's word2vec that add semantic meaning to the vectors. These vectors maps the words in the text to multi – dimensional space.
- ***s*** acts as the memory of the network which is passed to the next layer along with the input where $\boldsymbol{s_t}$ is the memory at the time ***t***. $\boldsymbol{s_t}$ is calculated as the function of $\boldsymbol{s_{t-1}}$ and $\boldsymbol{x_t}$ as follows:

$$s_t = f(Ux_t + ws_{t-1})$$

The function *f* is a non – linear function like tanh or ReLU that maps the collective value of vector the current input i.e. x_t and the memory at time *t-1* between certain intervals.

- o_t is the output of the layer at time interval *t* and it is calculated based on the s_t that is memory at the time *t* and the input sequence x_t. The intermediate output is not required in case of spam detection since we need the final class label.

$$o_t = softmax(V\, s_t)$$

- The final out is given by applying a mean pooling function followed by a logistic regression function to the series of memory contents (s_1, s_2, s_3,..., s_t) that gives a probabilistic output for the classifier labels and the values of all the labels add to 1:
- The vectors (U, V, W) are the weight vector that remains constant at each layer.

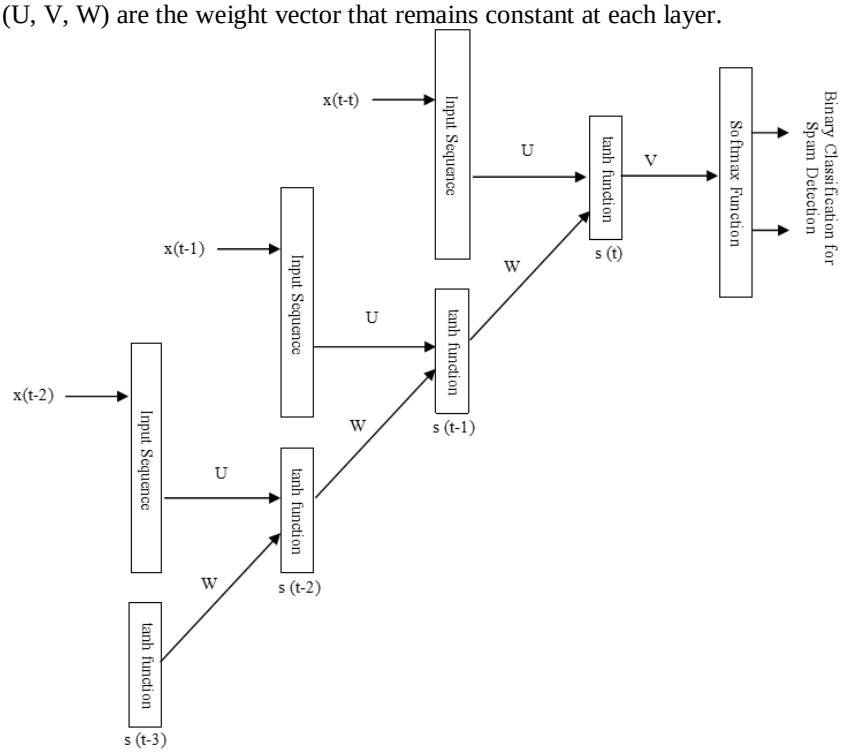


Fig. 3 Working of RNN for classification

B. Convolution Neural Network

Convolution Neural Network (CNN) also known as CovNets is type of feed forward neural network where the architecture is sparsely connected, unlike traditional neural network, which is fully connected. Here, a region of input

layer is connected to a neuron in the next layer. CNNs have initially started to evolve in the area of image recognition and classification and are supporting most of the computer vision systems. More recently, CNN have started giving effective results in the area of NLP like sentence modelling, search query retrieval, sentence classification in the area of sentiment analysis, movie reviews classification, product reviews etc. Unlike the case of images, where the image is represented as a set of 784 pixels (28x28), the text has to be converted into numerical values so as to perform NLP task. This is done by converting the text into word vectors also known as word embeddings, as defined in the previous section. A single dimension matrix is formed corresponding to each token or word and multi – dimensional matrix is formed for the entire text sentence. Architecture of CNN is shown in figure 4. The architecture of CNN consists of four operations: convolution, activation function, pooling stage and application of softmax function. These are shown in figure 4 and are described below and the detailed process is shown in figure 7

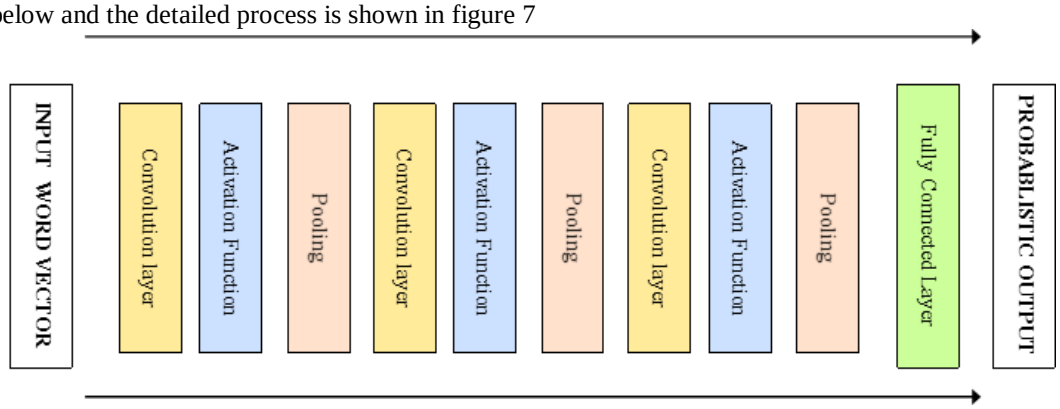


Fig. 4 Architecture of CNN

1. Convolution Layer: This layer is like sliding window function applied over the matrix. The sliding window, acts as a filter for the current region and calculates the value of the matrix depending upon the function used. Convolution layers increases feature abstractness while moving from abstract features to more precise features. In a convolution layer, if $x_i \in \mathbb{R}^n$ represents a word vector having n dimensions. The x_i represents i^{th} word/toekn in the sentence, then the sentence of having m words is represented as:

$$x_{1:m} = x_1 \oplus x_2 \oplus \dots \oplus x_m$$

Where the word vector of different tokens are concatenated with the help of $\oplus$. Convolution layer consists of a filter $\boldsymbol{w} \in \mathbb{R}^{\boldsymbol{hn}}$ having h rows and n columns and filter is applied on the sentence matrix to get a feature. A feature c_i from the window of words is given by applying an activation function on the dot product of filter *w* with the selected window from the input matrix

The filter c_i is applied as many times possible in the word vector and we get a feature map $\boldsymbol{c}$ having length (m-h+1) where m is the number of tokens, h is the filter size.

Convolution process is shown in the figure 5 below:

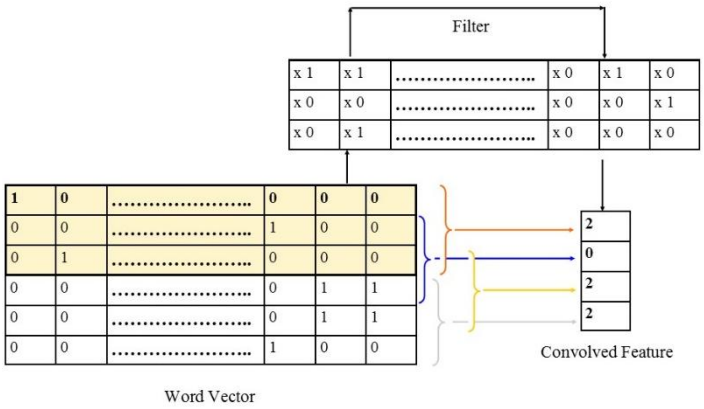


Fig. 5 Convolution Process

This filter slides over 2 to 5 words at a time i.e. the width of the filter consists of whole rows, taking 2 to 5 rows at a time to product feature map. The convolution of different filter with same window gives differ feature map and the CNN model uses different filter with number of features per filter. Therefore, dimension of feature maps is equal to the number of different filter size * number of filter per filter size.

2. Non – Linearity (ReLU): After each convolution operation as described, a non – linear activation function is applied to introduce non – linearity in the network as shown in the equation (1). Since the convolution operation is linear operation involving element wise matrix multiplication and addition, a non – linear function like Rectified Linear Unit (ReLU) or hyperbolic tangent (tanh) applied. These function maps the feature value to a specific continuous real value range. For eg. , the ReLU is a non – linear function that replaces all the negative values in the feature map with zero.

3. Pooling or Sub Sampling: After the convolution layers, we get the resultant matrix from the various filters, which are not uniform in size due to various filter size. Therefore, pooling is applied so as to get the uniform output that can be fed as input in the classification layer. Secondly, it reduces the dimension of the convolution layer output while keeping the important features intact thus controlling overfitting. Pooling layer produces a part or sample from the result of the convolution layers and thus increase spatial abstractness. It can be applied over the resultant matrix as a whole or taking a window at a time. The most common pooling function is ***max*** function. It takes the maximum cell value as output from the window of cells.

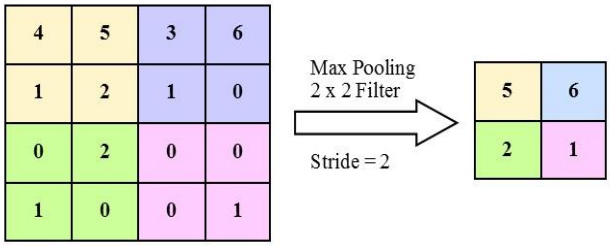


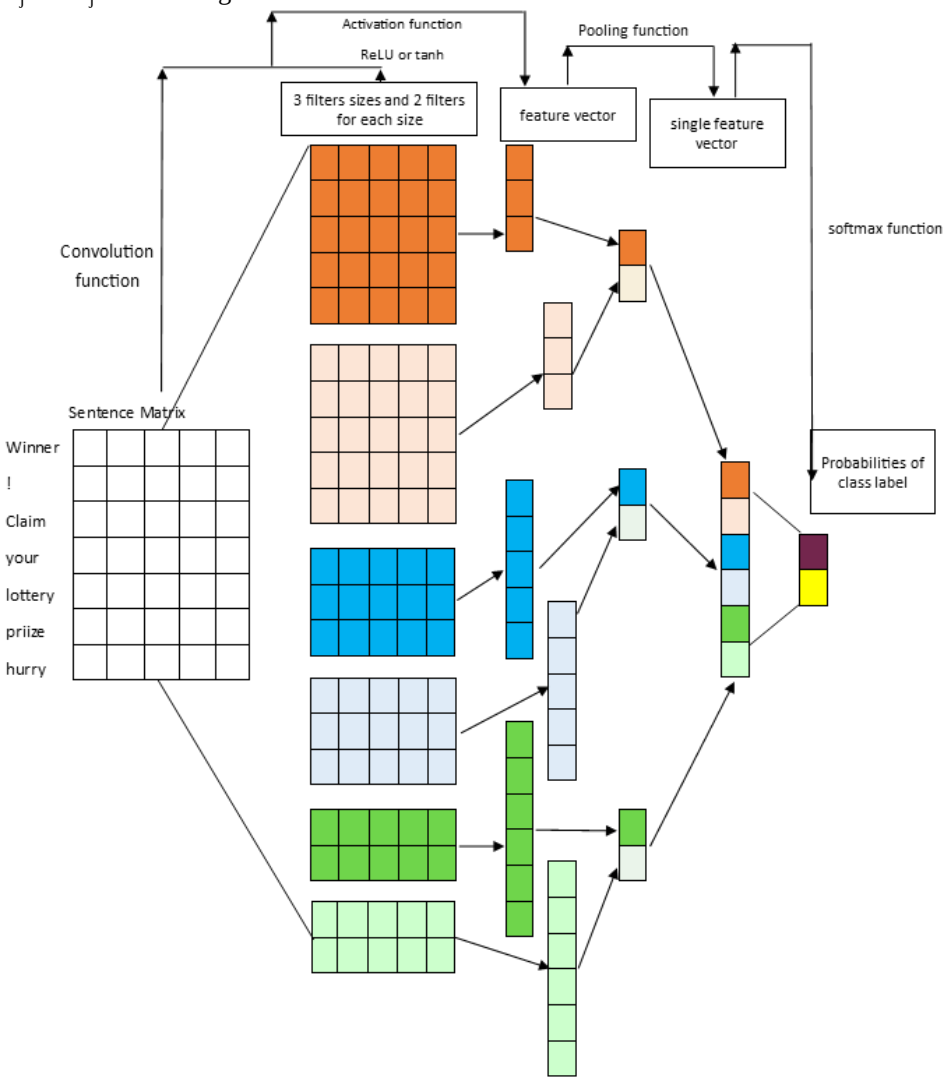
Fig. 6 Max Pooling Function

In the figure 6, a filter of 2x2 is applied with the stride of 2.

4. Classification (Fully Connected Softmax Layer): This is a fully connected multi – layer perceptron layer that takes max pooling layer as an input and the N dimensional vector as output, where N is the number of class labels in the classification problem. In case of spam detection, the value of N = 2. This layer uses softmax activation function for classification. The speciality of this type of function is that, it takes a vector of arbitrary real – valued numbers and maps it to the vector of values between [0,1] and all the values adds to 1. This function gives the probabilistic value of the classes to which an input text belongs and is given by:

$$p(j|x) = \frac{e^{x^T w_j + b_j}}{\sum_{k=1}^{K} e^{x^T w_k + b_k}}$$

Where w_j and b_j are the weights and bias vector of the k^{th} class.



IV. CONCLUSIONS

The overview suggest that the deep learning techniques have potential for effective text classification problems like spam detection because these models work effectively on raw data by learning high level features on its own. The data related to social media is very diverse in form and most of the time no language structure is followed, people tend to use slangs and short forms. Therefore, it is difficult to use handcrafted feature extraction for use in traditional text classification algorithms. The future work includes the experimentation of spam detection using RNN and CNN deep learning architectures and comparing it with traditional classification algorithms.

REFERENCES

[1] Van Dijck, José. The culture of connectivity: A critical history of social media. Oxford University Press, 2013.

[2] Boyd, D., & Ellison, N. (2008). Social network sites: Definition, history, and scholarship. Journal of Computer Mediated Communication, 13(1), 210—230

[3] Lenhart, Amanda. "Teens, social media & technology overview 2015." Pew Research Center 9 (2015).

[4] MAAWG. Messaging anti-abuse working group. Email metrics report. Q1 2012 to Q2 2014. Available at https://www.m3aawg.org/sites/default/files/document/M3AAWG_2012-2014Q2_Spam_Metrics_Report16.pdf

[5] Thomas, Kurt, et al. "Design and evaluation of a real-time url spam filtering service." Security and Privacy (SP), 2011 IEEE Symposium on. IEEE, 2011.

[6] Kim, Jangbok, Kihyun Chung, and Kyunghee Choi. "Spam filtering with dynamically updated URL statistics." IEEE Security & Privacy 4 (2007): 33-39.

[7] Levine, John R. "Experiences with Greylisting." CEAS. 2005.

[8] González-Talaván, Guillermo. "A simple, configurable SMTP anti-spam filter: Greylists." computers & security 25.3 (2006): 229-236.

[9] Ahmed, Shabbir, and Farzana Mithun. "Word Stemming to Enhance Spam Filtering." *CEAS*. 2004.

[10] Mccord, Michael, and M. Chuah. "Spam detection on twitter sing traditional classifiers." *International Conference on Autonomic and Trusted Computing*. Springer Berlin Heidelberg, 2011.

[11] Kolari, Pranam, et al. "Detecting spam blogs: A machine learning approach."*Proceedings of the National Conference on Artificial Intelligence*. Vol. 21. No. 2. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.

[12] Wang, Alex Hai. "Don't follow me: Spam detection in twitter." Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on. IEEE, 2010.

[13] Tretyakov, Konstantin. "Machine learning techniques in spam filtering." *Data Mining Problem-oriented Seminar, MTAT*. Vol. 3. No. 177. 2004.\

[14] Hearst, Marti A., et al. "Support vector machines." IEEE Intelligent Systems and their Applications 13.4 (1998): 18-28.

[15] Rish, Irina. "An empirical study of the naive Bayes classifier." IJCAI 2001 workshop on empirical methods in artificial intelligence. Vol. 3. No. 22. IBM New York, 2001.

[16] Safavian, S. Rasoul, and David Landgrebe. "A survey of decision tree classifier methodology." (1990).

[17] Dong, Yan-Shi, and Ke-Song Han. "A comparison of several ensemble methods for text categorization." Services Computing, 2004.(SCC 2004). Proceedings. 2004 IEEE International Conference on. IEEE, 2004.

[18] Basant Agarwal, Namita Mittal, "Semantic Feature Clustering for Sentiment Analysis of English Reviews", In IETE Journal of Research, Taylor Francis, Vol: 60 (6), pages 414-422, 2014.

[19] Agarwal B, Mittal N (2016) Prominent feature extraction for sentiment analysis. Springer book series: socio-affective computing series. Springer, Berlin

[20] Agarwal B, Mittal N, Bansal P, Garg S (2015) Sentiment analysis using common-sense and context information. Comput Intell Neurosci. doi:10.1155/2015/715730

[21] A. Ntoulas, M. Najork, M. Manasse & D. Fetterly. Detecting Spam Web Pages through Content Analysis. WWW'2006.

[22] M. Mccord and M. Chuah "Spam Detection on Twitter Using Traditional Classifiers", ATC'11, Sept 2-4, 2011, Banff, Canada.

[23] G.E. Hinton and R.R. Salakhutdinov, Reducing the Dimensionality of Data with Neural Networks, Science, 28 July 2006, Vol. 313. no. 5786, pp. 504 - 507.

[24] Le, Quoc V. "Building high-level features using large scale unsupervised learning." 2013 IEEE international conference on acoustics, speech and signal processing. IEEE, 2013.

[25] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." Nature521.7553 (2015): 436-444.

[26] Bengio, Yoshua. "Learning Deep architectures for AI." Foundations and trends in Machine Learning 2.1 (2009): 1-27.

[27] Deng, Li. "A tutorial survey of architectures, algorithms, and applications for deep learning." APSIPA Transactions on Signal and Information Processing 3 (2014): e2.

[28] Glorot, Xavier, Antoine Bordes, and Yoshua Bengio. "Domain adaptation for large-scale sentiment classification: A deep learning approach." Proceedings of the 28th International Conference on Machine Learning (ICML-11). 2011.

[29] Sutskever, Ilya, James Martens, and Geoffrey E. Hinton. "Generating text with recurrent neural networks." Proceedings of the 28th International Conference on Machine Learning (ICML-11). 2011.
[30] Graves, Alex, Abdel-rahman Mohamed, and Geoffrey Hinton. "Speech recognition with deep recurrent neural networks." 2013 IEEE international conference on acoustics, speech and signal processing. IEEE, 2013.
[31] Graves, Alex, and Navdeep Jaitly. "Towards End-To-End Speech Recognition with Recurrent Neural Networks." ICML. Vol. 14. 2014.
[32] Hinton, Geoffrey, et al. "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups." IEEE Signal Processing Magazine 29.6 (2012): 82-97.ss
[33] Lawrence, Steve, C. Lee Giles, Ah Chung Tsoi, and Andrew D. Back. "Face recognition: A convolutional neural-network approach." IEEE transactions on neural networks 8, no. 1 (1997): 98-113.
[34] Ciresan, Dan Claudiu, Ueli Meier, Luca Maria Gambardella, and Jurgen Schmidhuber. "Convolutional neural network committees for handwritten character classification." In 2011 International Conference on Document Analysis and Recognition, pp. 1135-1139. IEEE, 2011.
[35] Ciresan, Dan Claudiu, Ueli Meier, Luca Maria Gambardella, and Jürgen Schmidhuber. "Deep, big, simple neural nets for handwritten digit recognition." Neural computation 22, no. 12 (2010): 3207-3220.
[36] Kim, Yoon. "Convolutional neural networks for sentence classification." arXiv preprint arXiv:1408.5882 (2014).
[37] dos Santos, Cícero Nogueira, and Maira Gatti. "Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts." COLING. 2014.
[38] Wang, Peng, et al. "Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification."Neurocomputing 174 (2016): 806-814.
[39] Kalchbrenner, Nal, Edward Grefenstette, and Phil Blunsom. "A convolutional neural network for modelling sentences." arXiv preprint arXiv:1404.2188(2014).

