IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 8, August 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.625

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Deep Feature based Text Clustering using K-Means Algorithm

Rutuja Tulashigeri, Dr.B.F.Momin

P.G. Student, Department of Computer Science and Engineering, Walchand College of Engineering, Sangli, India

Associate Professor, Department of Computer Science and Engineering, Walchand College of Engineering, Sangli, India

ABSTRACT: Text Clustering is putting the data into similar data into similar cluster using machine learning algorithms. There are different types of machine algorithms like SVM mode, DBSCAN, K-Means, Agglomerative, Hierarchical Clustering. K-means Clustering is unsupervised clustering algorithm. Unsupervised Learning is where available input data does not have a labeled response. The 'K ' in Kmeans clustering stands for the optimal number of clusters found from data by the method. Text clustering is a critical step in text data analysis and has been extensively studied by the text mining community. Most existing text clustering algorithms are based on the bag-of-words model, which faces the high-dimensional and sparsity problems and ignores text structural and sequence information. Deep learning-based models such as convolutional neural networks and recurrent neural networks regard texts as sequences but lack supervised signals and explainable results.

KEYWORDS: Deep Learning , Feature Extraction , Text Clustering , Confusion Matrix , Preprocessing.

I. INTRODUCTION

To cluster a set of objects means to automatically partition them into clusters (parts), so that objects in the same cluster are as similar to each other as possible. In text clustering the objective is to group texts with similar content together. Such clusters give a overview of a text set and could be used to find structure in for instance the occupation answers of the twins. Text clustering is an application within the large and growing field of Information Retrieval, a sub-area of Language Technology. The search engine2 is the most well known information retrieval tool. Text clustering can also be applied to the documents retrieved by a search engine, so that they can be presented in groups according to content. Recall from the introduction that the objective of clustering is to partition an unstructured set of objects into clusters (parts). One often wants the objects to be as similar to objects in the same cluster and as dissimilar to objects from other clusters as possible. Clustering has been used in many different areas and there exist a multitude of different clustering algorithms for different settings. The field of artificial intelligence (AI) is experiencing an ongoing boom. It has gone from being a relatively obscure technology into being a part of everyday modern life, basically finding its way into almost every industry.

II. LITERATURE SURVEY

[1] In this Paper, Sentences were tagged and normalized using the Longest Common Subsequence (LCS) algorithm for the selection of the most similar subset of sentences.[2] A dataset of one million tweets, freely available on Internet for research purposes, was extracted by using Rapidminer on which text mining is applied through R language. R is an open source language and environment for statistical computation and graphics. Its various packages are used to carry out text processing. [3] Classifying text into emotional labels/intensities is considered a difficult problem. This paper resolves one of the state-of-the-art NLP research emotion and intensity detection tasks using Deep Learning and ensemble implementations**.** [4] To understand the thoughts and feelings of the public, twitter can be considered as one of the best platforms for sentiment analysis. In the existing studies on Covid-19, various word embedding techniques with machine learning and deep learning classifiers has been used for the analysis. [5] In this study, such tweets are being extracted from Twitter using a Twitter API authentication token. The raw tweets are stored and processed using

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

NLP. The processed data is then classified using a CNN classification algorithm. The algorithm classifies the data into three classes, positive, negative, and neutral.[6] In this paper we experimentally quantify the generality versus specificity of neurons in each layer of a deep convolutional neural network and report a few surprising results. [7] Latent Dirichlet Allocation (LDA), a generative probabilistic model for collections of discrete data such as text corpora. We report results in document modeling, text classification, and collaborative filtering, comparing to a mixture of unigrams model and the probabilistic LSI model.[8] The quality of these representations is measured in a word similarity task, and the results are compared to the previously best performing techniques based on different types of neural networks.

III. METHODOLOGY

BBC News Classification: This dataset is the dataset that contains all the information related to news. The news contained in this are from various categories like Regional, Sports, Entertainment, Business, Politics and Technology. Text documents are one of the richest sources of data for businesses. We'll use a public dataset from the BBC comprised of 2225 articles, each labeled under one of 5 categories: business, entertainment, politics, sport or tech. The dataset is broken into 1490 records for training and 735 for testing. The goal will be to build a system that can accurately classify previously unseen news articles into the right category. The competition is evaluated using Accuracy as a metric.

Text clustering includes following steps:
1) Model description: Including feature extraction and selection, demonstrating the data which is suitable in calculation in algorithm.
2) Text Processing involves Text Preprocessing, Feature Extraction , Clustering.
3) Perform Clustering algorithm with K-means clustering model.
4) Evaluate the results.

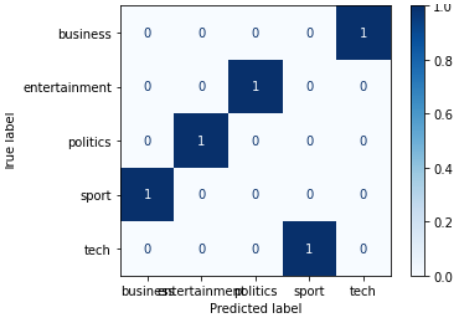
IV. EXPERIMENTAL RESULTS

The Final Result will get generated based on the overall classification and prediction. The performance of this proposed approach is evaluated using some measures like, using the following Confusion Matrix:-

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$



www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

	precision	recall	f1-score	support
business	0.87	0.79	0.83	73
tech	0.85	0.80	0.82	59
politics	0.78	0.83	0.80	52
sport	0.99	0.91	0.94	75
entertainment	0.62	0.82	0.70	39
accuracy			0.83	298
macro avg	0.82	0.83	0.82	298
weighted avg	0.85	0.83	0.84	298

We derive the output as below, in NEWS classification whenever user come across news he will get news classified according to their groups. There are different categories as Business, Entertainment, Politics, Sport, Technology.

	category	headline
0	business	Company profits soar as market demands increase
1	entertainment	New movie release breaks box office records
2	politics	Government passes new healthcare legislation
3	sport	Local team wins championship after intense game
4	tech	Innovative startup introduces groundbreaking t...

	content
0	The company has seen a significant increase in...
1	The latest movie release has broken multiple b...
2	The government has successfully passed new leg...
3	In an intense game last night, the local team ...
4	A new startup has introduced a groundbreaking ...

V. CONCLUSION

The input dataset was mentioned in our research paper.We are implemented the NLP techniques and classification algorithms (i.e.) machine learning algorithm. Then, machine learning algorithms such as K-Means Clustering. Finally, the result shows that the accuracy for above mentioned algorithm and the output is derieved as given above in the paper. Then, categorize the news.

REFERENCES

[1] C. C. Aggarwal and C. K. Reddy, Data Clustering: Algorithms and Applications. Boca Raton, FL, USA: CRC Press, 2013.
[2] N. Ahmad and J. Siddique, “Personality assessment using Twitter tweets,” Procedia Comput. Sci., vol. 112, pp. 1964–1973, Sep. 2017.
[3] T. Ahmad, A. Ramsay, and H. Ahmed, “Detecting emotions in English and Arabic tweets,” Information, vol. 10, no. 3, p. 98, Mar. 2019.
[4] A. Bandi and A. Fellah, “Socio-analyzer: A sentiment analysis using social media data,” in Proc. 28th Int. Conf. Softw. Eng. Data Eng., in EPiC Series in Computing, vol. 64, F. Harris, S. Dascalu, S. Sharma, and R. Wu, Eds. Amsterdam, The Netherlands: EasyChair, 2019, pp. 61–67.
[5] F. Barbieri and H. Saggion, “Automatic detection of irony and humour in Twitter,” in Proc. ICCC, 2014, pp. 155–162.
[6] R. Bhat, V. K. Singh, N. Naik, C. R. Kamath, P. Mulimani, and N. Kulkarni, “COVID 2019 outbreak: The disappointment in Indian teachers,” Asian J. Psychiatry, vol. 50, Apr. 2020, Art. no. 102047.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

[7] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," J. Mach. Learn. Res., vol. 3, pp. 993–1022, Jan. 2003.
[8] P. Boldog, T. Tekeli, Z. Vizi, A. Dénes, F. A. Bartha, and G. Röst, "Risk assessment of novel coronavirus COVID-19 outbreaks outside China," J. Clin. Med.,
[9] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised learning of universal sentence representations from natural language inference data," in Proc. Conf. Empir. Methods Natural Lang. Process., 2017, pp. 670–680.
[10] M. E. Peters et al., "Deep contextualized word representations," in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol., 2018, pp. 2227–2237, doi: 10.18653/v1/N18-1202.
[11] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, arXiv:1301.3781.
[12] S. J. Pan and Q. Yang, "A survey on transfer learning," IEEE Trans. Knowl. Data Eng., vol. 22, no. 10, pp. 1345–1359, Oct. 2010.
[13] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," in Proc. Int. Conf. Neural Inf. Process. Syst., 2014, pp. 3320–3328.
[14] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," in Proc. 56th Annu. Meeting Assoc. Comput. Linguistics, 2018, pp. 328–339.
[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pretraining of deep bidirectional transformers for language understanding," in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol., 2019, pp. 4171–4186.
[16] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," OpenAI Blog, vol. 1, no. 8, pp. 1–24, Feb. 2019.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details