"An analytical study of Natural Language Processing in context with Machine Learning"

Article · April 2021

CITATIONS
0

READS
11

2 authors, including:

Dipti H. Domadiya
National Computer College,Jamnagar,Gujarat
19 PUBLICATIONS **14** CITATIONS

SEE PROFILE

“An analytical study of Natural Language Processing in context with Machine Learning”

Shital S Patel Atmiya University Rajkot [1],
Dr. Dipti H Domadia National Computer College Jamnagar [2],
ssspatel369@gmail.com , diptipunjani@gmail.com

ABSTRACT

Natural Language processing study has reached some extent where distinct machine learning algorithms were implemented to obtain better leads to text classification. This paper presents much previous research works in this field of interest. It discusses the different techniques used for text classification so far and summarizes the various methods' benefits and disadvantages. It is observed that each one of the algorithms works well, but some strategies outperform others. Most of the algorithms are often improved by careful
selection of the features which play an essential role within the learning of an algorithm.

KEYWORDS

Multimedia data, unstructured data, Natural Language Processing, Machine Learning, Supervised learning, unsupervised learning, NLP features, and Support vector machines (SVM).

INTRODUCTION

NLP-Natural Language Processing its name suggests that it may be a computer process of understanding the NLP of humans, i.e., a person's language, which may be any language out of the various languages are spoken round the world. So many applications are available in NLP but initial essential application is Text classification and categorization and this is the topic of this paper presentation. It is often subdivided into many areas like User reviews for various services, email spam filters, etc. Many world-known service providers like Google, Amazon, etc., use these techniques to expand their business and provide better services to the users.

Machine learning is just like the human brain, which keeps learning over the years. The machine needs a massive amount of knowledge to be trained upon and learn different features. During trained the algorithm adjusts its parameters using the datasets. Once the algorithm is introduced, it becomes capable of understanding the new dataset. There are ways to coach the machine. Generally, there are two alternative ways supervised and unsupervised of learning algorithms. We tell the device about the info in supervised learning, but we give unlabeled data to the machine in terms of unsupervised learning. Within the latter case, the machine learns to differentiate different clusters or groups within the data, and, supported that, it classifies the data into various labels. Earlier, NLP was done by hand-coding, but now we have a machine-learning algorithm that will process the info in a short time without much effort. Using machine learning techniques, many researchers have worked on NLP. This paper specializes in a few such pieces. SVM is one technique for highly utilized text recognition, and SVM may supervise the learning algorithm. The rest of the paper discusses the works done by many researchers during this area of interest.

LITERATURE SURVEY

In first reference E Asli and B Keller proposed an approach to spot Twitter paraphrases through knowledge-lean techniques. In this, they tried to match two tweets and predict whether or not they are equivalent. For this, they only considered character unigrams and word n-grams. They were considered all the possible combinations and observed performance for the mixture of character bigrams and word unigram is that the best. SVM classifier with linear and an RBF kernel was used. Obtained accuracy was 86.5(linear) and 85.7(RBF), with the f-score to be 67.4 and 66.7, respectively.

Garla et al in second reference, tried to match linear and laplacian SVM for text classification. In this paper, the text is utilized as clinical records. They are Labeled and unlabeled datasets. It had been found that laplacian SVM outperformed both the linear SVM and gaussian SVM. 0.74 Of linear SVM compared with Macro-F1 for laplacian was 0.7 SVM, and linear SVM sensitivity was 0.94 and 0.91, respectively. The F1-score for the two methods is 0.91 and 0.89.

To get the news headlines' emotion, Kirange and Deshmukh applied the SVM algorithm (in 3rd). We had six categories, namely, anger, disgust, fear, joy, sadness, and surprise; for an atypical set of emotion were used, for a typical set of emotions, we have three categories: positive, negative, and neutral. The algorithm was trained for the specific output emotion set. The average accuracy for all the emotions is obtained at 89.95%.

Two different filtering methods for text classification compared SVM with H taira and M Haruno that we can get from 4th reference. The primary one is Mutual Information (MI), and therefore the other is Part-of-speech (PoS). It had been found that Pos outperformed MI. the typical precision and recall is 78.1 and 75.5 for PoS and MI, respectively, for linear SVM. In second-order polynomial, the stock of accuracy and memory is 77.2 and 74.8, respectively.

In 5th reference Performance of two machine learning algorithms is compared by lewis and Ringuette. one may be a bayesian classier called PropBayes. Therefore the other is that the decision tree learning algorithm is applied on two text categorization data sets. The first dataset is that the Reuters dataset containing 21450 newswire stories, and therefore the second dataset is U.S. Foreign Broadcast Information Services. They observed that both algorithms performed well on the Reuters dataset and gave good results. Compare with the PropBayes method decision tree method performed well at high recall levels. They concluded that the algorithms were not so dissimilar and suggested that more focus should tend to the feature selection process because it was found to play an important role within the performance.

Lewis et al. (6th reference) proposed two machine learning techniques: Widrow-Hoff (W.H.) and, E.G., algorithms. These two algorithms are compared with the documented Rocchio method. The two methods are Information Retrieval (I.R.) algorithms which aren't such a lot known. But, during this work, it had been observed that both these algorithms were way better than Rocchio's algorithm. Dataset used was developed by TREC evaluations. There are three volumes of knowledge. Volume 1 & 2 are used for training, and volume 3 for testing. The F1 score for the Rocchio algorithm is 0.33, and for W.H. and, E.G., it's 0.45 and 0.44, respectively. For the various cases considered, the scores have an equivalent trend.

In 7th Yang and Pedersen did a comparative study for choosing features. They compared Five methods which are document frequency (D.F.), information gain (I.G.), mutual information (MI), chi-square, and term strength (T.S.). The most focus was on dimensionality reduction and checking which algorithm performs best on such a dataset. The info was varied from no removal (only stop word removed) to 98% removal of unique terms. KNN algorithm and linear most petite square fit (LLSF) were used for learning. It had been observed that I.G. and chi-square methods performed best. D.F. did well till 90% removal. T.S. worked fine till 60%. It had been concluded that this type of dimensionality reduction methods are often used for extensive text

Rennie et al. proposed a replacement algorithm called TWCNB in reference no.8. It stands for transformed weighted complement Naive Bayes. The regular Naive Bayes had few assumptions; these assumptions are thanks to how the algorithms are written. These assumptions lead to coffee accuracy and erroneous results for the algorithm. These errors are thanks to 1) dataset being skewed, 2) weight magnitude error. These errors are resolved by taking the complement of the one-vs.-all methods and normalizing the load vector. The opposite modification is formed to the info, which is text during this paper. Text data was modified; firstly, the frequency

transformation was applied to get rid of bias thanks to a text's occurrence. Secondly, the text length modification was wont to remove bias thanks to a text document's length. The results were far better than before. There are many experiments where researchers have tried to mix machine learning techniques with domain knowledge. Even in text categorization, many works are done to incorporate the document knowledge within algorithm training with the info.

IN 9th reference SVM with KNN and naive Bayes algorithms compared by Colas and Brazil. The main focus of this paper was to understand which is the most straightforward algorithm for text categorization? Here we experimented was a binary text classification. We found good results from all algorithms. Both KNN and naive Bayes algorithms were like the SVM. So we concluded that good optimization might be an excellent choice to use for text classification. SVM is advantageous within the case where nonlinearity is required. But in the case of small and datasets, SVM won't be an ethical choice.

Identity has been done by Hassan et al in 10th reference. SVM, and Naive Bayes algorithms are compared. Both the algorithm is trained with the knowledge of the document using histology which is claimed to be the simplest knowledge repositories. The improvement of the Naive Bayes algorithm is significant compared to the development of the SVM. Considering the macro-f1measure, the shared progress is 6.36% and 28.78% for SVM and naive Bayes algorithms, respectively. It had been concluded that a naive Bayes algorithm with text enrichment might be a more sensible choice for text categorization.

NLP IN CONTEXT WITH MACHINE LEARNING

In Natural Language Processing, textual data analyses are done using statistical techniques and Machine learning models. Parts of speech, sentiments, summarize data, classify the emotions, etc., identified by these techniques. The varied approaches could also be expressed as a model and then correlated to different texts, called supervised learning. Many more algorithms work on large data sets to obtain similar data, which's also called an unsupervised model. In supervised machine learning, a batch of text data is labeled and annotated with a summary of what the machine will check for and how each feature should be interpreted. Those documents are employed to "train" a model, which may then be analyzed with untagged text. Several supervised learning algorithms are Decision tree, Naive Bayes, Support vector machine, etc.

Decision Tree:

A decision tree may be a tree structure that consists of root nodes, internal nodes, branches, and leaf nodes. For deciding, one of the simplest methods is decision trees. The result of the leaf nodes represents the decision. These are often used for disambiguating problems from every linguistic level, beginning with ambiguities from phonetics and ending with understanding a dialog.

Naive Bayes:

Naive Bayes classification uses the Bayes rule, which classifies the text-supported probabilities with explicit assumptions between the attributes. Naïve Bayes classifiers are scalable for training data, and therefore the set of variables needs a more number of linear features. Training with maximum likelihood is often accomplished by analyzing the expression of a closed-form. This model of learning is one of the simplest for classifying text.

Support Vector Machine

A support vector machine is an algorithm that evaluates the effective boundary of decision among vectors belonging to a specific group (or category) and vectors not belonging to it. It is often applied to any data encoding vectors. The advantage of the SVM text classification facility is texts got to be converted into vectors to take to build and improve the model alongside supervised machine learning; NLP features are also used. A number of the NLP features are as follows:

Tokenization

Splitting text documents, phrases, sentences, words, etc., into chunks involves Tokenization.

For example: "Tokenization is the feature of Natural Language Processing" can be tokenized as ["Tokenization", "is" , "the", "feature", "of", "Natural", "Language", "Processing"]

Part of Speech tagging

Part of Speech Tagging is to spot any POS token; whether it might be noun, pronoun, adverb, adjective, etc., to spot entities, processing opinions, and retrieving them, there's a requirement of POS tagging.

Sentiment analysis

Sentiment analysis is referred to as finding or classifying emotions like positive, negative, and neutral. It's also referred to as opinion mining. Positive words indicate happiness, excitement, well, kindness, great, excellent, etc. Negative words mean badly, sadness, ugly, hate, dispute, etc. Neutral indicates when there are not any emotions.

Unsupervised learning

Unsupervised machine learning trains a model without labeled data. A number of the unsupervised techniques are Clustering, Latent Semantic Indexing and Matrix factorization, etc.

Clustering: Clustering may be a process where similar items group together. Each group with similar objects is understood as a cluster. The various clustering algorithms are K-means, Hierarchal clustering, etc.

Latent Semantic Indexing:

This approach compares commonly occurring words and phrases with each other, and it's used to return search results that are not the exact search term.

Matrix Factorization:

This method decomposes the matrix into a smaller matrix and is used to find latent factors.

CONCLUSION AND FUTURE SCOPE

The survey on different machine learning algorithms for the appliance of text mining is, to be more specific, the most focus is text categorization. Most machine learning algorithms are used, and every one has its advantages and drawbacks like SVM, which is that the best algorithm but still slower in some cases where a dataset is large and algorithms like KNN and Naive Bayes are better choices for learning. Therefore, one main thing to require far away from this survey is that no technique is best or lesser than the others but is like one another. It all depends on the appliance. Another thing to note here is that the importance of feature selection. Some algorithms can perform far better if features are selected carefully. The second thing to watch during this survey is that the SVM and Naive Bayes algorithm's indisputable fact is the simplest algorithms for text categorization. In most of the papers discussed here, these both have the very best accuracy. SVM performs well due to its ability to capture nonlinearity. Naive Bayes may be a robust algorithm and combined with some modifications like it camper form far better. Its other advantage is that it's straightforward to implement. These papers suggest that the researchers should specialize in feature selection more and also try different algorithms on sample data before choosing the algorithm together algorithm may go better than the opposite. More research goes on during this area by the authors also. One is applying the modified Naive Bayes algorithm on opinion mining (user reviews on different services).

REFERENCES

1. Eyecioglu, A. & Keller, B. (2015). “Twitter paraphrase identification with simple overlap features and SVMs." Proceedings of the 9th international workshop on semantic evaluation (SemEval@NAACL-HLT 2015). Association for Computational Linguistics(64–69).

2. V. Garla, C. Taylor, C. Brandt, "Semi-supervised clinical text classification with Laplacian SVMs: An application to cancer case management," J.Biomed. Informat., vol. 46, no. 5, pp. 869-875, Oct. 2013.

3. D. Kirange and R. Deshmukh, "Emotion classification of news headlines using SVM," Asian Journal of Computer Science and Information Technology, pp. 104-106, 2012.

4. H. Taira and M. Haruno. "Feature selection in SVM text categorization." In Proceedings of AAAI-99, 16th Conference of the American Association for Artificial Intelligence, pages 480–486, Orlando, US, 1999. AAAI Press, Menlo Park, US.

5. Lewis, D.D., Ringuette, M.: "A Comparison of Two Learning Algorithms for Text Categorization." In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, pp. 81–93. Las Vegas, U.S. (1994).

6. Lewis, D., Schapire, R., Callan, J. and Papka, R. (1996). "Training Algorithms for Linear Text Classifiers." In Hans-Peter Frei, Donna Harman, Peter Schauble and Ross Wilkinson (eds.), SIGIR '96: Proceedings of the 19th Annual International ACM SIGIR Conference Research and Development in Information Retrieval, 298-306.

7. Y. Yang and J. Pedersen. "A comparative study on feature selection in text categorization." In International Conference on Machine Learning (ICML), 1997.

8. J. Rennie, L. Shih, J. Teevan, and D. Karger. “Tackling the poor assumptions of na¨ıve Bayes text classifiers." In International Conference on Machine Learning (ICML) 20, 2003.

9. F. Colas, P. Brazdil, "Comparison of SVM and some older classification algorithms in text classification tasks," in IFIP-AI 2006, World Computer Congress, Santiago de Chile, Chile, Springer, vol. 217, 1861–2288, 2006, pp. 169–178.

10. Hassan, S., Rafi, M., & Shaikh, M.S. (2011). “Comparing SVM and naive Bayes classifiers for text categorization with Wikitology as knowledge enrichment." 14th IEEE International Multi topic Conference.