THINK INDIA JOURNAL

ISSN: 0971-1260

Vol-22-Issue-17-September-2019

Challenges and Opportunities in Labeling for Text Classification

Varun Dogra[1], Dr. Sahil Verma[2]

[1]dogra_varun@yahoo.com, [2]sahilkv4010@yahoo.in

[1, 2] Department of Computer Science and Engineering, Lovely Professional University

Abstract

The process of training model in supervised machine learning is more difficult than expected due to the challenges in labeling and annotating data. It has observed that majority of the organization dealing in Artificial Intelligence projects have run in to problems with labeling data to train models. Labeling is commonly done manually by domain experts, which is time consuming task. Many authors have given different approaches to reduce the burden of manual labeling. However, all approaches have been facing different challenges due to increasing volume and shape of the data which further degrades the performance of automation. In order to produce quality in AI projects, the training data must be correctly labelled. The paper presents various challenges and opportunities occur in dealing with unstructured textual data for labeling to produce training data at the expected quality. The paper would also help the readers or scholars to purse their research projects in the area of text analytics or natural language processing.

Index Terms: Natural Language processing, Text classification, Labeling and Machine Learning

1. Introduction

The primary task to prepare a text classifier is to label a collection of training documents, and then apply a supervised machine learning algorithm to train the classifier. Now-a-days, Internet is producing a large amount of data and it can be acquired by scraping, creating or copying from different web sources. A key task in designing and developing a machine learning model is not just to collect large amount of data but also to design strategy to accurately label data to add sense to the data.

Labeling in text data can be described as a way to structure the data depending on its content. This process of structuring involves tagging or labeling to a specific part of text information once it has been pre-processed. Labeling can be performed manually by the domain expert or automated programmed scripts. However, due to large volume of data the manual labeling results ineffective. In case of automated scripts, the algorithm must be able to understand the every piece of the text when it is being processed. The programme has to understand that which label should be assigned to each unit of the textual data. Here, the programmer has to create a script to detect the patterns automatically by running supervised learning algorithms on labelled text-data.

In supervised machine learning, the features and corresponding labels are put into an algorithm during the training process. By the time, algorithm recognizes the relationship between features and their labels. Manual labeling in learning model is complex task. To handle the issue, (Clément and Laurens 2011) have presented idea of using a small labelled data for each class and large unlabeled dataset for building classifier. (Han et al. 2016) proposed classifier using positive and labelled examples. (Liu et al. 2004) provided representative words for every class or tag. These techniques managed to reduce the complexity of labeling task.

In this paper, we have presented the various challenges occur during labeling process for text classification and opportunities to reduce the burden of completely manual labeling in supervised machine learning algorithms. The labeling process is the primary sub-task of text classification once the data pre-processing has made.

1.1 Natural Language Processing

Natural language processing is vast field of research where the philosophy revolves around artificial intelligence, data science and linguistics. There exist real world applications, language translation, question-answering, named-entity recognition. Every application deals with text data in its own way.

These applications of natural language processing or computational linguistics offer a platform for understanding text data processing while diving deep into it using state-of-art machine learning or deep learning methods. In this paper, we are dealing with text classification with respect to its primary

ISSN: 0971-1260

Vol-22-Issue-17-September-2019

task of labeling during trailing of model. Text classification offer interesting applications, like sentiment analysis, fake news detection, topic modelling and news article classification. This further offer us to explore the various phases of text classification along with challenges and opportunities in classifying text documents in to its pre-defined category.

1.2 Text Classification

Text classification describes the way of allocating pre-defined tags to new text data or sentences on the basis of trained classifier on the training data. During training phase, the training examples are labelled with these pre-defined categories. Here, labeling is often done by hand coded rules. (Kim et al. 2006) proposed feature weighing and text normalization per-document method for classification. (Ali et al. 2018) used bi-gram as feature for short text classification. They have mentioned that to achieve accuracy the labeling of text cannot be ignored. (Yi Wang and Xiao-Jing Wang 2005) have given variance-mean feature detection method for reducing dimensions of the feature set to achieve efficient text classification. Fig.1 presents the model of text classification in to two phases, training phase and testing phase. During training phase, labeling play utmost important role in preparing classifier for learning model.

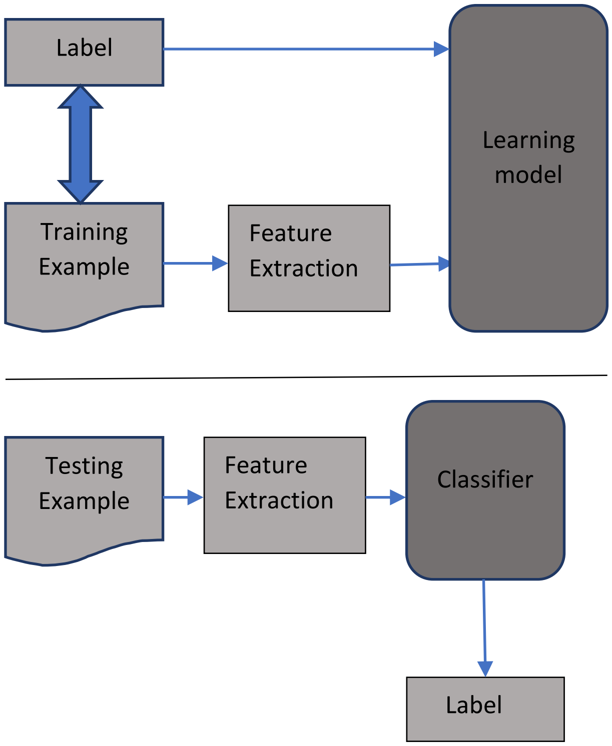

Fig 1. Training and Testing phases of text classification

1.3 Labeling textual-data

Labeling is neither clustering nor topic modelling problems. Labeling refers to understanding natural language, if the sentence is properly understood then it could be better labeling. Since internet have been producing millions of text documents to any particular domain, for the text classification, sample of text documents can be labelled manually. Selecting sample for labeling may follow different approaches depending on domain specific problem. (Rauber et al. 2000) explored self-organizing map for labeling with utmost important keywords. The maps were based on manually created topic-oriented libraries. (Hingmire et al. 2013) proposed LDA based classification without requiring labelled dataset. The topic modelling was done using LDA followed by assigning class label to each topic and then assigning class label to unlabeled documents based on closeness to one of the topic. However, mapping of topic labels to class labels is not feasible in all domains. Some topics may be belonging to multiple classes which arises the situation of multi-label classification.

1.3.1 Multi-class classification

In multi-class classification, a classification takes place with two or more classes where each label are mutually exclusive. The text document is assigned one and only one label.

1.3.2 Multi-label classification

In multi-label classification, the text document may be assigned more than one label. It assumes that the characteristics of data points of text documents are not exactly mutually exclusive. During the labeling process, it is focused that to what labels the text document be correlated. (J. Lee et al. 2019) proposed memetic feature selection for multi-label classification. They have assumed that memetic feature selection was specialized to multi-labeling. According to (Man Lan et al. 2009), multi-label classification was divided into multiple independent binary classification problems. (Pham, Nguyen, and Dinh 2017) focused on supervised learning by implementing semi-supervised multi-label classification technique to exploit unlabeled documents for improving performance. In the next sections, we explore the various challenges and opportunities occur in labeling sentences to its appropriate class or tag during training model.

2. Related Work

In supervised machine learning, each pre-defined class is assigned to its related text documents of training data set. This process is called labeling and it is often done manually during text classification. Text classification can be done using supervised, semi-supervised or unsupervised ways. (Chen et al. 2009) have used Naïve Bayes classifier, (Isa et al. 2008) have explored SVM method, (Ye, Zhang, and Law 2009) have used Naïve Bayes and N-gram methods for text classification using supervised learning.

To reduce the burden of manual labeling for entire set of training documents, (Pham, Nguyen, and Dinh 2017) have used semi-supervised technique for text classification. The authors have used specific features for each class labels selected through greedy approach and unlabeled text documents. Nearest Neighbors algorithm was used to classify new document. (Montañés et al. 2005) preferred rule or scoring based machine learning method for feature selection and automatic labeling of documents. The partial supervised learning from positive and unlabeled examples was used by (Han et al. 2016).

The unsupervised technique does not require any labelled documents for classification purpose. (Ko 2000) have used unsupervised method for categorizing documents to overcome the problems of manual labeling. The documents were categorized based on document similarity measure. (Slonim, Recognition, and Algorithms 2002) used clustering algorithm to categorizing text documents which was implemented using Information bottleneck method. In following section, we have presented the challenges and scope of improvement in labeling documents.

3. Challenges in Labeling: Different approaches

In supervised machine learning, the classifiers are trained by applying algorithm on labelled documents to understand the correlation between label and associated documents. This process is labor-intensive task and require human expertise. (K. Lee et al. 2011) have preferred manual labeling to classify the tweets in to pre-defined categories. But it was lacking in multi-labeling where required. They have mentioned to use application programme to automate the process up to some extent.

It was utmost important to design text classification techniques that minimizes the efforts of human in terms of cost and time. (Hingmire and Chakraborti 2014) proposed technique brought down the load of human labelers by annotating set of features instead of labeling entire document sets. These features were able to train classifier. The LDA algorithm was used to extract topic based features over entire document sets which represents statistical un-supervised machine learning approach. They defined model as $\acute{z}$ be the topic base for the documents D. The topic labelled to the word-unit $w \in$ W in the location n in document d as given below:

If $z_{d,n} = t, L(t) = c \rightarrow \acute{z}_{d,n}$=c (1)

$L(t)$ Refers the measure that returns class/tag given to the topic t by the human labeler. Human labeler has assigned a tag $c_i \in C$ to a topic t on the basis of possible words units. They have resulted that labeling topics were superior than labeling words as given by (Liu et al. 2004). They preferred NB formulation where every text sentence in trained labelled set D was taken as selected list of words. $w_{d,k}$ Referred as word in location k of text document d, every word comes from the vocabulary $V = \{w_1, w_2, \ldots., w_{|V|}\}$ and it provided collection of words that helps in classification. The pre-defined set of class were given$C = c_1, c_2, \ldots, c_n$. The probability to perform classification was given as Bayesian probability.

$$P(c_j) = \left(\sum_{j=1}^{D} P(c_j|d_i)\,\Big||D|\right) \qquad (2)$$

$P(c_j|d_i)$ Represents posterior probability and c_j is a class d_i is a document. They mentioned that instead of labeling entire document sets, labeling a set of representation words reduced the burden of manual labeling. However, it was mentioned that the load of labeling could be further reduced by annotating representative words by the labelers of domain's expert of different competence.

(Pham, Nguyen, and Dinh 2017) proposed a semi-supervised classification model that exploited the features of each labelled class selected by upgrading LIFT algorithm and LESC for consumption of unlabeled data, then 1NN was used to select appropriate class for new data instance. However, the model for selecting unlabeled data items or removing outliers from resulted clusters were not considered for evaluation of the efficiency of model. It is observed that authors has been finding and exploring techniques that could reduce the manual labeling task and improve efficiency from unsupervised classification models. (Gliozzo, Strapparava, and Dagan 2005) have introduced two unsupervised levels that upgrades the starting categorization level of bootstrapping using Latent semantic to obtain similarity between data items and feature, followed by applying Gaussian mixture to evaluate uniform categorization probabilities for unlabeled examples. In algorithm, only category names were used as initial seeds. However it was required to find optimal technique to get seed features for better performance.

The different authors have presented so far, the techniques for labeling data specific to their research area. There exist certain approaches those are used, preferably in industries to minimize the burden in terms of time and manpower. It is being followed in industries to allocate or broadcasts the labeling task as mentioned in Table 1.

Table 1: Different approaches for Labeling

Approach	**Description**	**Benefits**	**Limitations**
Manual Labeling in-house	Labeling or annotating documents by internal team	Produces accuracy, can apply checks on progress	More time and man power required
Outsourcing	Hiring temporary group of people through freelancing platforms	Task can be done by required competent professionals	Seeks organizing workflow
Crowdsourcing	Working with professionals from crowdsourcing forums	Fast process and cost reduces to some extent	May suffer from expected quality
Outsourced to IT companies	Communicating to third party IT companies for outsourcing	Guaranteed quality	High cost
Programmed	Writing scripts that labels textual data with human intervention	Achieves automated tasks and saves time	Does not produce expected quality

4. Future direction and Conclusion

The paper have mentioned various challenges in labeling text documents to achieve highly accurate classifier for domain specific problems. Generally, the task of labeling is done manually which provides high quality results as per expectations but it takes much time. The authors have found technique to reduce the burden of manual labeling by training model with partially labelled and unlabeled data. The process have compromised with little accuracy which could be back propagated through the model to reduce errors. The un-supervised way of achieving labelled data on the basis of representative words or feature selection through different algorithms have reduce the load in terms of man power and cost. However, accuracy of the training model has reduced as compared to manual labeled data.

In domain specific text mining problems, the labeling of data cannot be achieved accurately through programmed applications. It requires the human expert to disambiguate the certain domain specific words. There are opportunities to design certain libraries for dealing with domain specific data. The research has been evolving for few decades to facilitate the professionals in healthcare, finance, defense and e-commerce through text mining. Internet has been producing data in the form of text which help such professionals to take decision in their particular areas. Here, the role of labeler to understand the workflow in different domains and annotating texts writing domain specific scripts.

References

Ali, Mubashir, Shehzad Khalid, Mazhar Iqbal Rana, and Fizza Azhar. 2018. “A Probabilistic Framework for Short Text Classification.” *2018 IEEE 8th Annual Computing and Communication Workshop and Conference, CCWC 2018* 2018-Janua: 742–47. https://doi.org/10.1109/CCWC.2018.8301712.

Chen, Jingnian, Houkuan Huang, Shengfeng Tian, and Youli Qu. 2009. “Feature Selection for Text Classification with Naïve Bayes.” *Expert Systems with Applications* 36 (3 PART 1): 5432–35. https://doi.org/10.1016/j.eswa.2008.06.054.

Clément, Antoine, and Stéphane Laurens. 2011. “An Alternative to the Lyapunov Exponent as a Damage Sensitive Feature.” *Smart Materials and Structures* 20 (2): 1–34. https://doi.org/10.1088/0964-1726/20/2/025017.

Gliozzo, Alfio, Carlo Strapparava, and Ido Dagan. 2005. “Investigating Unsupervised Learning for Text Categorization Bootstrapping.” *HLT/EMNLP 2005 - Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, no. October: 129–36. https://doi.org/10.3115/1220575.1220592.

Han, Jiayu, Wanli Zuo, Lu Liu, Yuanbo Xu, and Tao Peng. 2016. “Building Text Classifiers Using Positive, Unlabeled and ‘Outdated’ Examples.” *Concurrency Computation* 28 (13): 3691–3706. https://doi.org/10.1002/cpe.3879.

Hingmire, Swapnil, and Sutanu Chakraborti. 2014. “Topic Labeled Text Classification: A Weakly Supervised Approach.” *SIGIR 2014 - Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 385–94. https://doi.org/10.1145/2600428.2609565.

Hingmire, Swapnil, Sandeep Chougule, Girish K. Palshikar, and Sutanu Chakraborti. 2013. “Document Classification by Topic Labeling.” *SIGIR 2013 - Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 877–80. https://doi.org/10.1145/2484028.2484140.

Isa, Dino, Lam Hong Lee, V. P. Kallimani, and R. Rajkumar. 2008. “Text Document Preprocessing with the Bayes Formula for Classification Using the Support Vector Machine.” *IEEE Transactions on Knowledge and Data Engineering* 20 (9): 1264–72. https://doi.org/10.1109/TKDE.2008.76.

Kim, Sang-bum, Kyoung-soo Han, Hae-chang Rim, and Sung Hyon Myaeng. 2006. “Some Effective Techniques for Naive Bayes Text Classification” 18 (11): 1457–66.

Ko, Youngjoong. 2000. “Automatic Text Categorization by Unsupervised Learning.” *Proceedings of the 18th Conference on Computational Linguistics*, 453–59. https://doi.org/10.3115/990820.990886.

Lee, Jaesung, Injun Yu, Jaegyun Park, and Dae Won Kim. 2019. “Memetic Feature Selection for Multilabel Text Categorization Using Label Frequency Difference.” *Information Sciences* 485: 263–80. https://doi.org/10.1016/j.ins.2019.02.021.

Lee, Kathy, Diana Palsetia, Ramanathan Narayanan, Md Mostofa Ali Patwary, Ankit Agrawal, and Alok Choudhary. 2011. “Twitter Trending Topic Classification.” *Proceedings - IEEE International Conference on Data Mining, ICDM*, 251–58. https://doi.org/10.1109/ICDMW.2011.171.

Liu, Bing, Xiaoli Li, Wee Sun Lee, and Philip S Yu. 2004. “Text Classification by Labeling Words.” *Artificial Intelligence* 34 (1): 425–430. http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Text+Classification+by+Labeling+Words#0.

Man Lan, Chew Lim Tan, Jian Su, and Yue Lu. 2009. “Supervised and Traditional Term Weighting Methods for Automatic Text Categorization.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31 (4): 721–35. https://doi.org/10.1109/TPAMI.2008.110.

Montañés, Elena, Irene Díaz, José Ranilla, Elías F Combarro, and Javier Fernández. 2005. “Scoring and Selecting Terms for Text Categorization.” *IEEE Intelligent Systems* 20 (3): 40–47. https://doi.org/10.1109/MIS.2005.49.

Pham, Thi-ngan, Van-quang Nguyen, and Duc-trong Dinh. 2017. “MASS : A Semi-Supervised Multi-Label Classi Fi Cation Algorithm with Speci Fi c Features,” 37–

47. https://doi.org/10.1007/978-3-319-56660-3.

Rauber, a, a Rauber, E Schweighofer, E Schweighofer, D Merkl, and D Merkl. 2000. “Text Classification and Labeling of Document Clusters with Self-Organising Maps.” *OEGAI-Journal* 19 (1): 17–23.

Slonim, Noam, I Pattern Recognition, and Clustering Algorithms. 2002. “Unsupervised Document Classification Using Sequential Information Maximization,” 129–36.

Ye, Qiang, Ziqiong Zhang, and Rob Law. 2009. “Sentiment Classification of Online Reviews to Travel Destinations by Supervised Machine Learning Approaches.” *Expert Systems with Applications* 36 (3 PART 2): 6527–35. https://doi.org/10.1016/j.eswa.2008.07.035.

Yi Wang, and Xiao-Jing Wang. 2005. “A New Approach to Feature Selection in Text Classification,” no. August: 3814-3819 Vol. 6. https://doi.org/10.1109/icmlc.2005.1527604.