Labeling News Article's Subject Using Uncertainty Based Active Learning

Meet Parekh[1(✉)] and Yash Patel[2]

[1] CBInsights, New York, USA
[2] New Jersey Institute of Technology, Newark, USA

Abstract. In Natural Language Processing, labeling a text corpus is often an expensive task that requires a lot of human efforts and cost. Whereas unlabeled text corpora in varying domains are readily available. For a couple of decades, research efforts have concentrated on algorithms that can be used for labeling the corpus, thus minimizing the number of articles required to be labeled manually. Semi-Supervised Learning and Active Learning have been a great promise for labeling the articles using a trained model. Also, Semi-Supervised learning algorithms and Active learning algorithms have strong theoretical guarantees. This study aims to tag 1183 articles from The New York Times and The Wall Street Journal with the subject (i.e. primary organization related to news articles) employing Active Learning algorithm. We used Active Learning algorithm which uses Random Sampling along with Uncertainty Based Querying. This Active Learning approach is used to train Naïve Bayes classifier using Bag of Words features. This classifier is used to tag 1183 articles of which only 167 required manual review, thus achieving reduction of 85.89% with 78.18% accuracy. Also, for verifying quality of labeled corpus, SVM classifier using same features was trained on labeled corpus giving accuracy of 74.45% on test data.

Keywords: Active learning · Natural language processing · Uncertainty sampling · Naïve bayes · SVM · Labeling

1 Introduction

While unlabeled text corpora are readily available, obtaining labeled corpus is a challenge confronting research community. The primary reason behind the scarcity of labeled corpus is the cost and human efforts incurred in labeling the corpus manually. Corpus used by Ng et al. [11] took university undergraduates one working year to label the corpus. Similarly, Tamas Vardi [12] took four years of efforts to construct Hungarian National Corpus. The Penn Chinese Tree Bank (CTB) [13] took four years to release its first edition CTB-I (annotated Chinese corpus) of 100,000 words. Thus labeling a corpus is a time-consuming effort and also requires high funding as subject experts have to dedicate a considerable amount of time for labeling a corpus.

S. Paiva et al. (Eds.): SmartCity360° 2020, LNICST 372, pp. 200–208, 2021.
https://doi.org/10.1007/978-3-030-76063-2_15

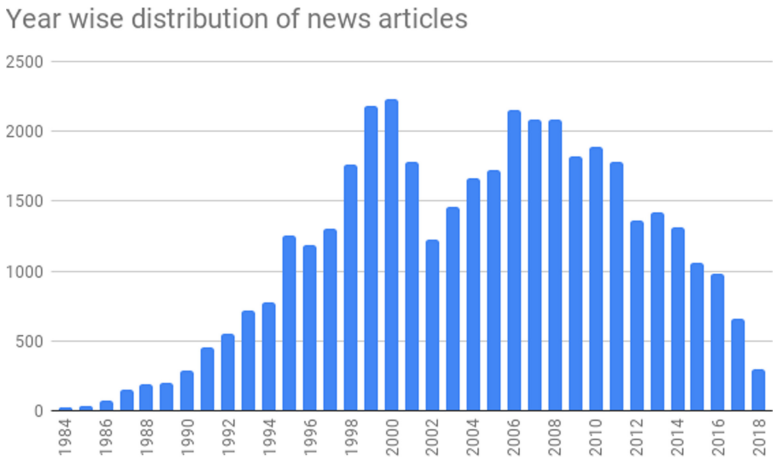

Fig. 1. Year wise distribution of News Articles

To address this issue, researchers have focused on devising algorithms for training models that can label articles. Different algorithms have been proposed by various studies that focus on minimizing the number of articles required to be labeled manually. Another goal for these algorithms is to produce high quality labeled corpus, as the accuracy of models highly depends on the underlying corpus. Thus different aspects of the quality of the corpus must also be taken into consideration while evaluating these algorithms. Literature alludes to semi-supervised learning [14] and Active Learning [1] for labeling unlabeled corpus.

Semi-supervised learning, an extension to supervised learning, can be defined as a model that trains itself on labeled dataset to predict the unlabeled dataset. The basic requirement of applying a Semi-Supervised learning is that unlabeled data should be much more than labeled data, else Supervised method could give the necessary output. The primary objective of Semi-Supervised learning is to train a classifier from both labeled and unlabeled data, such that it is better than the supervised classifier trained on the labeled data alone. Nigam et al. [23] showed that unlabeled data, when used in conjunction with a small amount of labeled data, can produce considerable improvement in learning accuracy. A classifier learns a function based on the provided labeled data, and extends this model to unlabeled data. The learning can be done in one of the many ways including self-training, probabilistic generative models, co-training, graph-based models, semi-supervised support vector machines, and so on.

Active Learning defers from Semi-Supervised learning in labeling unlabeled data, where Active Learning tries to determine most informative examples and queries them for manual review. Cohn et al. [1] defines Active Learning as learning in which learning algorithm has some form of control over which examples it will be trained on. Generally in Active Learning, learner selects a set of examples to be labeled and based on the learning strategy it sends some examples

for manual review. Strategies can be devised to reduce cost, improving accuracy of learner, selecting most informative example, selecting examples that learner is uncertain about etc. Most of these strategies can be categorized into two broad categories viz. 1. Query based Committee and 2. Uncertainty based sampling which are described in further detail in Sect. 2. Although Active Learning reduces human efforts in labeling the unlabeled data [15,16,20,22], yet certain degree of human intervention is required in both.

Also, some Active Learning algorithms are shown to have strong theoretical guarantees. Mellow Active Learner have shown to have upper bound for the separable data [2], Agnostic Active Learning algorithm have been shown to have upper bound [3–6] as well as lower bound [7].

This study focuses on labeling subjects (i.e., the organization which is the primary subject of the article) of the News Paper articles using Active Learning algorithm. Generally, headlines are the subject of newspaper articles; however, nouns mentioned in the headline may not necessarily have relevance with the rest of the article, whereas on the other hand nouns not mentioned in the headline might have relevance with the article. For example, an article mentions "Microsoft" in its headline, but the article primarily focuses on the product launched by "Google". On the contrary, headlines of some of the articles may not mention "Microsoft", but the article may very much relate to it. Such a labeling can be useful if one wishes to use this corpus to perform sentiment analysis about a particular organization.

For this study, articles containing the word "Microsoft" at least once were extracted from The New York Times and The Wall Street Journal. These articles were labeled by Active Learning algorithm using Random Uncertainty Sampling [27] with Naïve Bayes classifier. We used Bag of Words model to train Naïve Bayes classifier. Accuracy of classification was measured on a separate test data, and it was measured as the function of number of articles labeled manually. Also to evaluate the quality of labeled corpus, SVM classifier using same features i.e. Bag of Words was trained on labeled corpus and evaluated on test data.

Outline of this paper is as follows, Sect. 2 describes details about Active Learning and other related work, Sect. 3 gives a description of corpus used, Sect. 4 describes Design and Implementation of the study, Sect. 5 underlines the Results, Sect. 6 highlights Future Works that could follow and Section Acknowledges the support received for this study.

2 Related Work

Active learning is learning in which learning algorithm has some form of control over which examples it will be trained on. In Active Learning, learner selects a set of examples to be labeled manually based on learning strategy. General motivation behind selecting examples is to select those examples which are most informative.

Based on this motivation different strategies can be devised to select examples. One basic motivation behind selecting examples is to reduce the cost

incurred in labeling the articles. While many studies assumes that cost for labeling each example is same, this may not always be the case [32]. Thus a separate heuristic may be required to estimate the cost of labeling an example. Haertel et al. [17] devised cost sensitive heuristic function to measure the hourly cost incurred in labeling articles and used this heuristic to select articles which incurred low cost, thus achieving 73% reduction in hourly cost over random selection. Ringger et al. [21] further improvised heuristic function developed by Haertel et al. [17] by estimating the parameters for heuristic function based on a statistical study.

Another strategy is to select examples that possibly boosts up the accuracy of the classifier. This would be particularly useful if the purpose of the study is just to train classifier and use it for some task. Becker et al. [19] have used f-complement score to select examples that can potentially increase the f-1 score of classifier. On the similar lines, Thompson et al. [24] selected examples that would help maintain f-1 score of classifier.

At its base, general motivation is selecting examples that can be considered most informative for classifier. General strategies to achieve this purpose can be broadly categorized in two categories viz. 1. Query Based Committee and 2. Uncertainty Sampling [27].

In Query Based Committee methods, committee of classifiers are trained. These committee of learners are inquired to label examples. If there is disagreement within committee then this example is considered to be difficult and thus can be highly informative for classifier. Hachey et al. [16] used Query Based Committee method to determine examples that would need manual review, and found that examples selected by query based methods also had disagreement amongst group of annotators. Thus proving that these examples are complex and can be difficult to learn on. Dagan and Engelson [25] have used Query Based Committee method and found that it reduces cost of annotation. Similarly, Song et al. [33] considered examples on the margin of SVM to be the most informative examples, as they could change the separating hyper-plane.

In Uncertainty Sampling method, classifier gives certainty score for each example along with the label. If certainty score is below certain threshold then example is queried for manual review. Reason for selecting examples that the classifier is uncertain about, is that these examples can indeed be highly informative as they form boundary for probability based classifiers [26]. Lewis et al. [27] selected examples having probability of classification near 0.5 for manual review. In this study we have used Uncertainty Sampling along with Naïve Bayes classifier to determine examples that would be queried for manual review. Analogous to Uncertainty Sampling is Confidence Based Sampling, where learning algorithm requests manual review for examples for which confidence of classifier is low [30].

3 Dataset Description

Corpus containing 40144 news articles was constructed from The New York Times and The Wall Street Journal. All of these articles contained word

"microsoft" either in metadata or body or headline of the article. Articles of The New York Times and The Wall Street Journal were extracted from ProQuest library, additionally articles of The Wall Street Journal were also extracted from Factiva Library. Each article contained a headline, publishing date, miscellaneous metadata, text, author, acknowledgments, and copyright details.

Figure 1 shows the distribution of these news articles over the years. It was observed that the majority of these articles i.e., 34750, were published between 1995–2015. Due to limitations in resources, we only focused on labeling articles for the year of 1996, which contained 1183 articles.

For initial training of classifier, 19 articles were labeled manually by subject experts. On top of that 55 articles from year 1998 were labeled manually by subject experts which were then used as test data.

4 Design and Implementation

Figure 2 shows the setup for this study. This section describes details about different steps of our experimental setup.

As described in Sect. 3, each articles in corpus contained headline, publishing date, miscellaneous metadata, text, author, acknowledgments, and copyright details. However much of these details were not required for the purpose of this study, so required data needed was extracted from these articles. Also text was required to be broken into tokens which would be needed in the Bag of Words model.

Firstly headline, publication data, and article text were extracted for each articles ignoring rest of information. Also date for each articles were in different format, so it was converted to standard yyyy-mm-dd format.

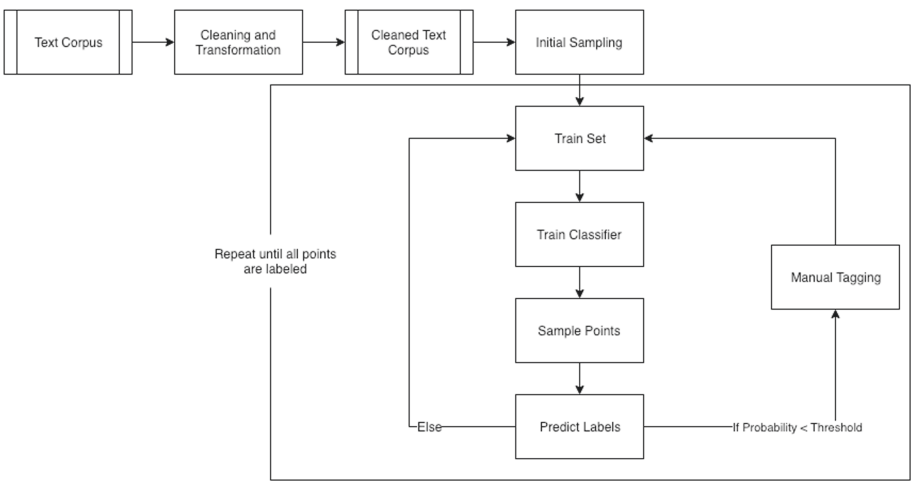

Fig. 2. Experimental setup

Thereafter NLTK [10] PUNKT [28] tokenizer was used to split article text and headline into tokens. Also NLTK "stopwords" corpus was used to remove tokens that were stop words. Remaining tokens were then used to create Bag of Words model where for each word corresponding frequency of occurrence of word were stored.

Once Bag of Words data was ready, Naïve Bayes classifier from sklearn [29] was trained on initial 19 labeled articles. Once classifier was trained we used Random Uncertainty Sampling to query articles for manual review which classifier was uncertain about.

To achieve this we randomly shuffled unlabeled data and fed it into classifier one by one. Articles that classifier was uncertain about were queried for manual review. Lewis et al. [27] used points having probability of certainty near 0.5 for manual review. Lewis et al. reasoned that these were the points that classifier was most uncertain about and formed sort of boundary for probabilistic classification, and if classifier was trained on these examples with correct label then classifier would be able to adjust its boundaries of classification. On similar lines we used the threshold of 0.6, such that if probability of majority class was less than 0.6 then it was queried for manual review. While rest of the articles having probability of classification above 0.6 were directly assigned respective labels. And this labeled articles were incorporated in training data.

After each manual review, classifier was retrained on new training data which included both, articles that were manually reviewed and articles that were directly assigned labels by the classifier. And then accuracy of classifier was measured on test data and plotted as a function of number of articles that were manually reviewed as shown in Fig. 3.

All 1183 articles of the corpus were labeled using this method. Another concern with Active Learning method is the quality of labeled corpus, i.e. how well would other classifiers work on corpus labeled with Active Learning [31]. To address this concern, we trained SVM classifier on this corpus using same features i.e. Bag of Words. We then evaluated accuracy of SVM on test data, results of which are described in Sect. 5.

5 Results

Figure 3 plots accuracy as a function of number of articles labeled manually. It can be observed from Fig. 3 that maximum accuracy achieved by Naïve Bayes on test data model was 78.18%. Also, learning process required 167 articles to be reviewed and labeled manually i.e. 14.11% required manual labeling. Thus model achieved reduction of 85.89% in labeling articles manually.

To measure the quality of labeled corpus, we trained SVM on labeled corpus using same features i.e. Bag of Words and measured its accuracy on test data. SVM gave accuracy of 74.45% on test data which is comparable with maximum accuracy achieved by Naïve Bayes i.e. 78.18% while labeling the corpus. This reflects that the labeling of the corpus is of good quality.

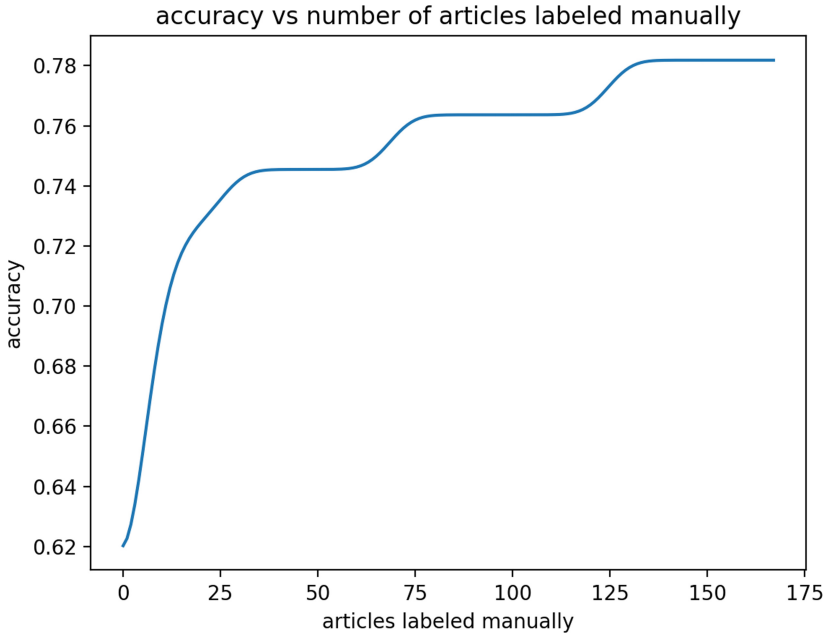

Fig. 3. Accuracy vs Number of articles tagged manually

6 Future Works

Random Sampling was used in this study to generate samples for training. Although reduction of 85.89% was observed using this method, there are few studies which suggests that purely Random Sampling may not be very effective strategy [17,19]. Thus, in future works we would like to further investigate different heuristics for selecting samples for training, and compare its result with pure Random Sampling.

For the purpose of this study, we evaluated the quality of labeled corpus by measuring the accuracy of SVM trained using same features i.e. Bag of Words. However for future studies, we would like to investigate training classifier using different features from the one used by classifier in Active Learning algorithm.

Also, we would like to use this strategy to label all 40144 articles and then use articles related to Microsoft for performing sentiment analysis, thus evaluating sentiments for Microsoft over a timeline of 24 years.

Acknowledgments. We want to thank New York University Library, ProQuest, and Factiva for making this news article corpus open to students for academic research. We would also like to thank our colleagues Manthan Shah of Pace University and Amod Panchal of Rutgers University, for manually tagging news articles.

References

1. Cohn, D., Atlas, L., Ladner, R.: Improving generalization with Active Learning. Mach. Learn. **15**(2), 201–221 (1994)
2. Hanneke, S.: Teaching dimension and the complexity of active learning. In: Bshouty, N.H., Gentile, C. (eds.) COLT 2007. LNCS (LNAI), vol. 4539, pp. 66–81. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-72927-3_7
3. Balcan, M.F., Beygelzimer, A., Langford, J.: Agnostic active learning. J. Comput. Syst. Sci. **75**(1), 78–89 (2009)
4. Dasgupta, S., Hsu, D.J., Monteleoni, C.: A general agnostic Active Learning algorithm. In: Advances in Neural Information Processing Systems, pp. 353–360 (2008)
5. Koltchinskii, V.: Rademacher complexities and bounding the excess risk in Active Learning. J. Mach. Learn. Res. **11**, 2457–2485 (2010)
6. Hanneke, S.: A bound on the label complexity of agnostic Active Learning. In: Proceedings of the 24th International Conference on Machine learning, pp. 353–360, June 2007
7. Beygelzimer, A., Dasgupta, S., Langford, J.: Importance weighted Active Learning. In: Proceedings of the 26th Annual International Conference on Machine Learning, pp. 49–56, June 2009
8. Settles, B.: Active Learning. In: Synthesis Lectures on Artificial Intelligence and Machine Learning, vol 6. no. 1, pp. 1–114 (2012)
9. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J.R., Bethard, S., McClosky, D.: The Stanford CoreNLP natural language processing toolkit. In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55–60, June 2014
10. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python: Analyzing Text With the Natural Language Toolkit. O'Reilly Media Inc., Sebastapol (2009)
11. Ng, H.T., Lee, H.B.: Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In: Proceedings of the 34th Annual Meeting on Association for Computational Linguistics, pp. 40–47. Association for Computational Linguistics, June 1996
12. Váradi, T.: The Hungarian National Corpus. In: LREC (2002)
13. Xue, N., Chiou, F. D., Palmer, M.: Building a large-scale annotated Chinese corpus. In: COLING 2002: The 19th International Conference on Computational Linguistics (2002)
14. Board, R.A., Pitt, L.: Semi-supervised learning. Mach. Learn. **4**(1), 41–65 (1989)
15. Tomanek, K., Wermter, J., Hahn, U.: An approach to text corpus construction which cuts annotation costs and maintains reusability of annotated data. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pp. 486–495, June 2007
16. Hachey, B., Alex, B., Becker, M.: Investigating the effects of selective sampling on the annotation task. In: Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005), pp. 144–151, June 2005
17. Haertel, R.A., Seppi, K.D., Ringger, E.K., Carroll, J.L.: Return on investment for Active Learning. In Proceedings of the NIPS Workshop on Cost-Sensitive Learning, vol. 72, December 2008
18. Zhu, X.J.: Semi-supervised learning literature survey. University of Wisconsin-Madison Department of Computer Sciences (2005)

19. Becker, M., Hachey, B., Alex, B., Grover, C.: Optimising selective sampling for bootstrapping named entity recognition. In: ICML-2005 Workshop on Learning with Multiple Views, pp. 5–11, August 2005
20. Ringger, E., et al.: Active learning for part-of-speech tagging: accelerating corpus annotation. In: Proceedings of the Linguistic Annotation Workshop, pp. 101–108, June 2007
21. Ringger, E.K.: Assessing the costs of machine-assisted corpus annotation through a user study. In: LREC, vol. 8, pp. 3318–3324, May 2008
22. Ngai, G., Yarowsky, D.: Rule writing or annotation: cost-efficient resource usage for base noun phrase chunking. In: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, pp. 117–125. Association for Computational Linguistics, October 2000
23. Nigam, K., McCallum, A., Thrun, S., Mitchell, T.: Using EM to classify text from labeled and unlabeled documents (No. CMU-CS-98-120). Carnegie-Mellon Univ Pittsburgh PA School of Computer Science (1998)
24. Thompson, C.A., Califf, M.E., Mooney, R.J.: Active learning for natural language parsing and information extraction. In: ICML, pp. 406–414, June 1999
25. Dagan, I., Engelson, S.P.: Committee-based sampling for training probabilistic classifiers. In: Machine Learning Proceedings 1995, pp. 150–157. Morgan Kaufmann (1995)
26. Engelson, S.P., Dagan, I.: Minimizing manual annotation cost in supervised training from corpora. In: Proceedings of the 34th Annual Meeting on Association for Computational Linguistics, pp. 319–326. Association for Computational Linguistics, June 1996
27. Lewis, D.D., Gale, W.A.: A sequential algorithm for training text classifiers. In: SIGIR 1994, pp. 3–12. Springer, London (1994). https://doi.org/10.1007/978-1-4471-2099-5_1
28. Kiss, T., Strunk, J.: Unsupervised multilingual sentence boundary detection. Comput. Linguist. **32**(4), 485–525 (2006)
29. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
30. Li, M., Sethi, I.K.: Confidence-based Active Learning. IEEE Trans. Pattern Anal. Mach. Intell. **28**(8), 1251–1261 (2006)
31. Baldridge, J., Osborne, M.: Active learning and the total cost of annotation. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pp. 9–16, July 2004
32. Settles, B., Craven, M., Friedland, L.: Active learning with real annotation costs. In: Proceedings of the NIPS Workshop on Cost-sensitive Learning, pp. 1–10, December 2008
33. Song, H., Yao, T., Kit, C., Cai, D.: Active learning based corpus annotation. In: CIPS-SIGHAN Joint Conference on Chinese Language Processing, Chicago (2010)