# Impact of Batch Size on Stopping Active Learning for Text Classification

Garrett Beatty†
Department of Computer Science
The College of New Jersey
Ewing, NJ 08628
Email: beattyg2@tcnj.edu

Ethan Kochis†
Department of Computer Science
The College of New Jersey
Ewing, NJ 08628
Email: kochise1@tcnj.edu

Michael Bloodgood
Department of Computer Science
The College of New Jersey
Ewing, NJ 08628
Email: mbloodgood@tcnj.edu

*Abstract*—When using active learning, smaller batch sizes are typically more efficient from a learning efficiency perspective. However, in practice due to speed and human annotator considerations, the use of larger batch sizes is necessary. While past work has shown that larger batch sizes decrease learning efficiency from a learning curve perspective, it remains an open question how batch size impacts methods for stopping active learning. We find that large batch sizes degrade the performance of a leading stopping method over and above the degradation that results from reduced learning efficiency. We analyze this degradation and find that it can be mitigated by changing the window size parameter of how many past iterations of learning are taken into account when making the stopping decision. We find that when using larger batch sizes, stopping methods are more effective when smaller window sizes are used.

## I. INTRODUCTION

The use of active learning has received a lot of interest for reducing annotation costs for text classification [1], [2], [3].

Active learning sharply increases the performance of iteratively trained machine learning models by selectively determining which unlabeled samples should be annotated. The number of samples that are selected for annotation at each iteration of active learning is called the batch size.

An important aspect of the active learning process is when to stop the active learning process. Stopping methods enable the potential benefits of active learning to be achieved in practice. Without stopping methods, the active learning process would continue until all annotations have been labeled, defeating the purpose of using active learning. Accordingly, there has been a lot of interest in the development of active learning stopping methods [3], [4], [5], [6], [7].

Another important aspect of the active learning process is what batch size to use. Previous work has shown that using smaller batch sizes leads to greater learning efficiency [3], [8]. There is a tension between using smaller batch sizes to optimize learning efficiency and using larger batch sizes to optimize development speed and ease of annotation. We analyze how batch size affects a leading stopping method and how stopping method parameters can be changed to optimize performance depending on the batch size.

†These students contributed equally to this paper.

We evaluate the effect batch size has on active learning stopping methods for text classification. We use the publicly available 20Newsgroups dataset[1] in our experiments.

For our base learner, we use the implementation of a Support Vector Machine from the scikit-learn Python library. For our sampling algorithm, we use the closest-to-hyperplane algorithm [3], which has been shown in recent work to compare favorably with other sampling algorithms [9]. We use a binary bag of words representation and only consider words that show up in the dataset more than three times. We use a stop word list[2] to remove common English words.

For analyzing the impact of batch size on stopping methods, we use a method that will stop at the first training iteration that is within a specified percentage of the maximum achievable performance. We denote this method as the Oracle Method, and we will set the percentage to 99 and denote this as Oracle-99. We set the percentage to 99 because it is typical for leading stopping methods to be able to achieve this level of performance (see Table 1 in [5]). Although the Oracle Method cannot be used in practice, it is useful for contextualizing the stopping results of practical stopping methods.

## II. RESULTS

We considered different batch sizes in our experiments, based on percentages of the entire set of training data. The results for batch sizes corresponding to 1%, 5%, and 10% of the training data for the 20Newsgroups dataset are summarized in Table I.

### A. Oracle Results

Looking at Table I, one can see that Oracle-99 needs more annotations with larger batch percents to reach approximately the same F-Measure as with smaller batch percents.

These results are consistent with past findings that learning efficiency is decreased with larger batch sizes [3], [8]. However, an open question is whether changing the parameters associated with actual stopping methods can make them experience less degradation in performance when larger batch sizes are used. In particular, an important parameter of stopping methods is the window size of previous iterations to consider.

[1]http://qwone.com/~jason/20Newsgroups/
[2]Long Stopword List from https://www.ranks.nl/stopwords

| Batch Percent / Stopping Method | 1% | 5% | 10% |
|---|---|---|---|
| Oracle Method | 1514.20 | 2490.40 | 3901.95 |
| | 76.17 | 75.55 | 75.49 |
| BV2009 (Window Size = 3) | 1299.50 | 3877.10 | 6446.70 |
| | 74.44 | 75.17 | 75.19 |
| BV2009 (Window Size = 1) | 1101.75 | 3141.30 | 5089.50 |
| | 74.40 | 75.04 | 75.11 |

TABLE I

STOPPING METHOD RESULTS ON 20NEWSGROUPS FOR DIFFERENT BATCH SIZES USING VARIOUS WINDOW SIZES. THE TOP NUMBER IN EACH ROW SHOWS THE NUMBER OF ANNOTATIONS AT THE STOPPING POINT AND THE BOTTOM NUMBER SHOWS THE F-MEASURE AT THE STOPPING POINT.

The next subsection shows how decreasing the window size parameter can help to reduce the degradation in performance that stopping methods experience with larger batch sizes.

### B. Comparing BV2009 with the Oracle Method

We denote the stopping method published in [5] as BV2009. This stopping method will stop the active learning process if the mean of the three previous kappa agreement values between consecutive models is above a threshold. For larger batch percents, note that BV2009 stops later than the optimal Oracle Method point.

We ran BV2009 with smaller window sizes for each of our different batch sizes. Our results are summarized for a window size of one in the row "BV2009 (Window Size = 1)" in Table I. When using a window size of one, BV2009 is able to stop with a smaller number of annotations than when using a window size of three. This is done without losing much F-Measure. The next subsection provides an explanation as to why smaller window sizes are more effective than larger window sizes when larger batch sizes are used.

### C. BV2009 Window Size Discussion

We set $n$ to be the window size that the user has defined. Kappa is an agreement metric between two models. Therefore, BV2009 needs $n + 1$ models to be generated before it begins to check if the average is above the threshold. This does not necessarily mean it stops after $n + 1$ models have been generated. Rather, it represents the first point in the active learning process at which BV2009 even has a chance to stop.

When using larger batch percents, fewer models are generated than when using smaller batch percents. This gives any stopping method less points to test whether or not to stop. We also note that kappa agreement scores are generally low between the first few models trained. This, combined with fewer points to stop at, causes BV2009 to stop somewhat sub-optimally when using very large batch percents. Usage of very large batch sizes, such as 10% of the data, is not common so sub-optimal performance of stopping methods in those situations is not a major problem.

### III. CONCLUSION

Active learning has the potential to significantly reduce annotation costs. Two important considerations in the active learning process are when to stop the iterative process of asking for more labeled data and how large of a batch size to use when asking for additional labels during each iteration. We found that stopping methods degrade in performance when larger batch sizes are used. The degradation in performance is larger than the amount that can be explained due to the degradation in learning efficiency that results from using larger batch sizes. An important parameter used by stopping methods is what window size of earlier iterations to consider in making the stopping decision. Our results indicate that making the window size smaller helps to mitigate the degradation in stopping method performance that occurs with larger batch sizes.

### REFERENCES

[1] A. Mishler, K. Wonus, W. Chambers, and M. Bloodgood, "Filtering tweets for social unrest," in *Proceedings of the 2017 IEEE 11th International Conference on Semantic Computing (ICSC)*. San Diego, CA, USA: IEEE, January 2017, pp. 17–23. [Online]. Available: http://ieeexplore. ieee.org/stamp/stamp.jsp?tp=&arnumber=7889498&isnumber=7889486

[2] M. Bloodgood and K. Vijay-Shanker, "Taking into account the differences between actively and passively acquired data: The case of active learning with support vector machines for imbalanced datasets," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*. Boulder, Colorado: Association for Computational Linguistics, June 2009, pp. 137–140. [Online]. Available: http://www.aclweb.org/anthology/N/N09/N09-2035.pdf

[3] G. Schohn and D. Cohn, "Less is more: Active learning with support vector machines," in *Proceedings of the 17th International Conference on Machine Learning (ICML)*, 2000, pp. 839–846.

[4] F. Laws and H. Schütze, "Stopping criteria for active learning of named entity recognition," in *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, Manchester, UK, August 2008, pp. 465–472. [Online]. Available: http://www.aclweb. org/anthology/C08-1059

[5] M. Bloodgood and K. Vijay-Shanker, "A method for stopping active learning based on stabilizing predictions and the need for user-adjustable stopping," in *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*. Boulder, Colorado: Association for Computational Linguistics, June 2009, pp. 39–47. [Online]. Available: http://www.aclweb.org/anthology/W09-1107

[6] M. Bloodgood and J. Grothendieck, "Analysis of stopping active learning based on stabilizing predictions," in *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*. Sofia, Bulgaria: Association for Computational Linguistics, August 2013, pp. 10–19. [Online]. Available: http://www.aclweb.org/anthology/W13-3502

[7] A. Vlachos, "A stopping criterion for active learning," *Computer Speech and Language*, vol. 22, no. 3, pp. 295–312, 2008.

[8] K. Brinker, "Incorporating diversity in active learning with support vector machines," in *Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), August 21-24, 2003, Washington, DC, USA*, T. Fawcett and N. Mishra, Eds. AAAI Press, 2003, pp. 59–66.

[9] M. Bloodgood, "Support vector machine active learning algorithms with query-by-committee versus closest-to-hyperplane selection," in *Proceedings of the 2018 IEEE 12th International Conference on Semantic Computing (ICSC)*. Laguna Hills, CA, USA: IEEE, January 2018.