Leipzig University
Institute of Computer Science
Degree Programme Computer Science, B.Sc.

Active Learning for Text Classification using Dimensionality Reduction for Core-Set

Bachelor's Thesis

Yannick Brenning
Born Aug. 27, 2002 in Bamberg

Matriculation Number 3732848

1. Referee: Christopher Schröder
2. Referee: Christian Kahmann

Submission date: February 31, 2022

Declaration

Unless otherwise indicated in the text or references, this thesis is entirely the product of my own scholarly work.

Leipzig, February 31, 2022

. . . .
Yannick Brenning

Abstract

This is the LaTeX template for Bachelor and Master theses at Webis. This template contains several hints and conventions on how to structure a thesis, how to cite the work of others, and how to display your results besides plain text.

Contents

Chapter 1

Introduction

Text is one of the most widespread and important sources of information, but extracting data and tangible knowledge from it can be a difficult and expensive task. With the advent of the digital age, enormous amounts of unstructured texts are available with more being generated by the day. Due to this increasingly large amount of textual data, manually processing information at a larger scale becomes infeasible and thus demands the use of computer-driven approaches.

The classification of text, meaning the assignment of a category or class to a document or piece of text is one of the most common and useful ways to gain information from a piece of text. As the amount of available text content continues to grow, text classification tasks become an increasingly important area of research within the field of natural language processing.

Thanks to machine learning and data science, we have been able to develop many methods of extracting information from text, and as a result, perform text classification at a larger scale. This possibility for automated organization of data can enhance insights and decision-making across industries such as healthcare, finance, and social sciences, among many others. Active Learning (AL) is a subfield of machine learning in which the learning algorithm is able to perform queries on an information source in order to reduce the total amount of annotated data. This method can offer significant advantages in improving model performances and especially in reducing labeling costs. Though there is no universally good strategy, AL has been proven to be useful in many cases where annotating data is expensive or the total amount of data is very large (Settles [2009]).

Oftentimes, there is a large amount of unlabelled available data for an AL model to learn from. In this case, selecting the most effective data points to be labelled by the information source from this large pool becomes a crucial, but difficult challenge to overcome.

One attempt at improving the effectiveness of AL in this regard is the Core-Set approach (Sener and Savarese [2018]) This method uses core-set selection to counter the issue of AL ineffectiveness on convolutional neural networks. The proposed approach selects a set of points the pool such that a model learning over this subset can be consistent when provided the remaining data points. The method was shown to have improved results when compared to other approaches in the field of computer vision (Sener and Savarese [2018], Caramalau et al. [2021]), which encompasses tasks that focus on enabling computers to interpret and understand visual information from the world. This field involves a variety of different methods including image classification, object detection, and semantic segmentation, all of which have significant importance in scientific research, classification tasks, and pattern recognition.

However, Core-Set has been shown to have mixed results in cases of text classification using BERT (Prabhu et al. [2021], Ein-Dor et al. [2020]) and binary text classification using DNNs (Liu et al. [2021]). Particularly in the first case, the experiments show that Core-Set performs poorly even when compared to the random sampling strategy. In addition, the approach has even been shown to be less effective in computer vision tasks in cases with higher numbers of classes as well as higher-dimensional data points (Sinha et al. [2019]). The theoretical analysis shown in Sener and Savarese [2018] briefly mentions this within the context of higher class amounts, however it does not attempt to provide a potential solution to the problem.

This thesis aims to explore the possibility of improving the Core-Set approach for text classification tasks. By first explaining Core-Set's functionality and the theoretical reasons for why it tends to underperform in certain classification tasks, I aim to then demonstrate the performance difference in comparison to various baseline approaches on large datasets of text content in order to verify this claim. Furthermore, this thesis looks to improve on the Core-Set approach within the context of text classification tasks and demonstrate this improvement as a part of its experiment.

In the following, Chapter 2 explains the background and related work on the topics of text classification (Section 2.1), active learning in general (Section 2.2), and the Core-Set approach to AL more specifically (Section 2.3). In Chapter 3, I will explain my approach to improving the performance of Core-Set for text classification using dimensionality reduction. In Chapters 4 and 5, I will present my experiment as well as discuss its results. Finally, Chapter 6 will conclude the thesis and provide insights on potential future developments of the method.

Chapter 2

Background/Related Work

2.1 Text Classification

Text classification is one of the most fundamental and important tasks in the field of Natural Language Processing (NLP). As a result, developing efficient automatic text classification methods has become an important research topic.

One of the most common applications of text classification is determining whether the opinion associated with a certain document has a positive or negative sentiment, also known as sentiment analysis. This has a wide range of uses, including the possibility for businesses to better gauge customer opinions on products and services (Liu and Zhang [2012]) in order to adapt accordingly. This application is a binary classification task, meaning the classifier has two classes with which each document can be labelled (positive or negative). Similarly, one might apply this binary classification task to the problem of spam filtering in e-mails, text messages and more.

Beyond that, many applications of text classification require multiple classes, such as news and content categorization. In this case, text classification algorithms can organize documents into specific topics or themes (e.g. Sports, Business, Politics, . . .) (Sebastiani [2002]). Other applications include information retrieval, recommender systems, and document summarization (Kowsari et al. [2019]).

Generally, text classification methods can be divided into the following phases: data collection and preprocessing, feature extraction, classifier selection, and model evaluation (Kowsari et al. [2019], Mironczuk and Protasiewicz [2018], Ikonomakis et al. [2005]). . . .

2.2 Active Learning

Active Learning has become an increasingly important field when considering the need for efficient models as well as the labelling bottleneck in various machine learning tasks. Many fields, such as speech recognition, information extraction and classification suffer from this bottleneck as a result of their instance labels being expensive or time-consuming to obtain (Settles [2009]). ...

2.3 Core-Set

The Core-Set approach as proposed in Sener and Savarese [2018] was originally designed for convolutional neural networks (CNNs) to tackle the problem of AL algorithms on large datasets. The empirical studies conducted by Sener and Savarese [2018] show that Core-Set outperforms other approaches within the field of image classification.

The task of active learning is defined as a *core-set selection* problem in which the algorithm selects a smaller subset of points with increased diversity to learn from such that the model can be competitive over a larger dataset. This problem ends up being equivalent to the *k-Center* problem (Sener and Savarese [2018]), which is also referred to as the min-max facility location problem. In practice, Core-Set solves the k-Center problem with a greedy approach and improves upon its robustness by placing an upper limit on the number of outliers. ...

2.4 Dimensionality Reduction

As mentioned earlier, one of the major challenges of the Core-Set approach (among others) is handling data points with a higher dimensionality (Sinha et al. [2019]). Broadly speaking, this is a phenomenon coined by Richard E. Bellman known as the *curse of dimensionality* (François [2007]). As a result, many algorithms have been developed to transform data from a high-dimensional space into a low-dimensional space. This task directly poses another challenge: managing to reduce the dimensionality of the data while still being able to retain the highest possible amount of information.

One such method is the Principal Components Analysis (PCA), which is one of the most popular linear techniques for dimensionality reduction (Van Der Maaten et al. [2009]). In essence, PCA linearly transforms the data into a representation that attempts to closely describe the variance of the initial data (Jolliffe and Cadima [2016]). Other commonly used linear techniques include

Linear Discriminant Analysis (LDA), Multidimensional Scaling (MDS), and Non-negative Matrix Factorization (NMF).

These techniques can be powerful, however they often miss important non-linear structures in high-dimensional data. Therefore, non-linear techniques have been developed such as Isometric Mapping (Isomap) and t-distributed Stochastic Neighbor Embedding (t-SNE). The latter is a relatively modern probablistic approach that has improved upon many other non-linear techniques in creating a single map that reveals structure on many different scales. In addition, it manages to reduce the tendency of SNE to crowd data points together at the center by using a different cost function (Van der Maaten and Hinton [2008]).

First, it converts the high-dimensional Euclidean distances into conditional probabilities, such that similar data points are assigned higher probabilities and dissimilar data points are assigned very low probabilities. It then creates a similar probability distribution over the lower-dimensional map such that the *Kullback-Leibler divergence* (a measure of how one probability distribution differs from another) is minimized.

This method is especially useful for the visualization of high-dimensional data, in which the aim is to display and view the underlying structure in a given dataset. t-SNE plots are strongly influenced by the chosen hyperparameters however, and thus a good understanding for the influence of these parameters is important. Particularly *perplexity*, a measure of the effective number of neighbours, has a complex effect on the resulting reductions. According to Van der Maaten and Hinton [2008], typical values for perplexity are between 5 and 50 and larger/denser datasets often generally require higher perplexities (Van der Maaten [2023]). Other commonly adjusted parameters include the learning rate and number of iterations for the optimization.

Chapter 3

Approach

As mentioned, distance-based methods such as Core-Set can be ineffective in higher dimensions. To overcome this challenge, I propose the application of t-SNE as a dimensionality reduction technique on the training data before employing the Greedy Core-Set strategy. In order to optimize my use of t-SNE, I will first examine the effects of different hyperparameters on the reductions. Based on these examinations, I will select the optimal t-SNE parameters to use for the experiment.

Next, I will examine the effect of pointwise probabilities on the Core-Set algorithm. Core-Set currently operates using the distances between its points – in this experiment, I propose an approach where Core-Set considers the probability distributions in addition to the distances.

To accomplish this, I will be using an uncertainty-based approach. Uncertainty sampling, which was introduced by Lewis and Gale [1994], essentially attempts to select the unlabelled examples with the lowest classification certainty. In this case, I specifically opted for an approach similar to that of Breaking-Ties (Luo et al. [2005]), where the instances with the smallest margin between their top two most likely classification probabilities are selected, essentially "breaking the tie" between the two most likely classes. The results of this calculation will then be multiplied as a weight vector with the distances normally used by Core-Set.

This approach resulted in two different implementation ideas: the first, which we will call "Weighted Greedy Core-Set" (WCS), calculates the uncertainties prior to executing the standard Core-Set algorithm, then multiplies the resulting Core-Set distances and probabilities using weights (in this case, I used a 80-20 weighting in favour of the Core-Set distances).

The second, which I will call "Ranked Weighted Greedy Core-Set" (RWCS), opts to first compute a Core-Set twice the size of the original sample size n, then "re-ranks" the resulting distances according to Breaking-Ties uncertainties and finally selects the points with the n highest uncertainties.

Finally, I will take into account the class distributions within each Core-Set and select points based on the representation of each class.

The experiment will be conducted with two models on three well-known classification datasets and will measure the performances throughout 5 runs per combination. I have chosen two state-of-the-art transformers, BERT (Devlin et al. [2019]) and SetFit (Tunstall et al. [2022]) as models for this experiment.

Chapter 4

Experiment

4.1 Data

This experiment was conducted across three datasets commonly used in the field of AL. These datasets are of three different types: sentiment analysis (S), questions (Q), and news (N). For binary classification, I used **Movie Review**, a sentiment analysis dataset (Pang and Lee [2005]) containing 5,331 positive and 5,331 negative samples. For multi-class text classification, **AG's News** (Zhang et al. [2015]), comprised of 120,000 training samples and 7,600 test samples, and **TREC** (Li and Roth [2006]), a question dataset containing 5,500 training samples and 500 test samples.

The test set was only provided in the case of AG's News and TREC, in the case of Movie Review I employed a split of the 10,662 samples myself. ...

4.2 Experiment Setup

As mentioned earlier, setting t-SNE's hyperparameters and exploring the resulting differing reduction behaviours has been shown to often be necessary (Wattenberg et al. [2016]). Van der Maaten [2023] mentions that examining the resulting plots is one of the best ways to assess the quality of the visualizations, in addition to comparing the Kullback-Leibler divergences.

Below, I have demonstrated the effect of t-SNE on a training set of the MR dataset (Pang and Lee [2005]) containing 9,596 samples with a feature dimension of 768. The reductions are completed using perplexity values of 5, 30, 50 and 100 with iteration counts of 1,000 and 5,000 respectively. Note that the samples have been embedded using BERT and then normalized prior to the dimensionality reductions.

Starting at a perplexity of 30, the points begin to cluster more and more tightly. The clusters do not seem to separate in any of the runs – it is important to note, however, that distances and sizes of the clusters do necessarily carry any meaning when examining the reduction plots (Wattenberg et al. [2016], Van der Maaten [2023]).

It is also worth noting that the differences between 1,000 and 5,000 iterations appear quite small despite the relative increase in computations. ...

The experiment is conducted using 20 queries on 25 instances, with a validation set size of 10%. The experiment is averaged over five runs of queries per combination of dataset, model and query strategy.

4.3 Experiment Results

The following figures show the learning curves for each combination of dataset, model and query strategy.

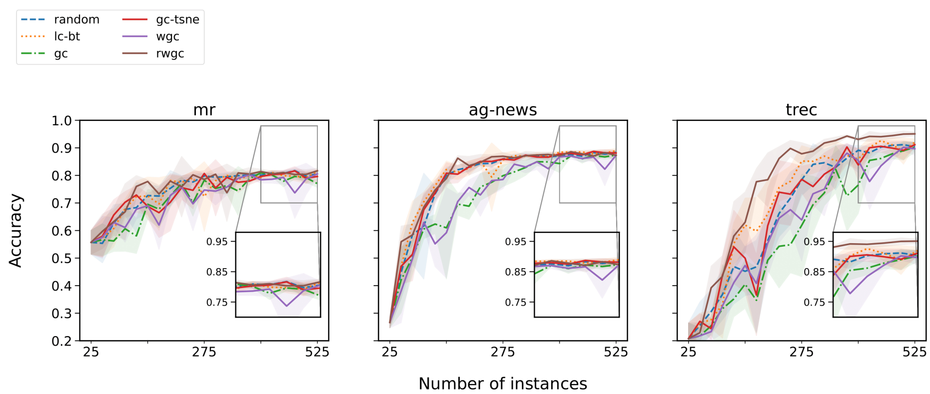

Figure 4.1: Active Learning curves of BERT on each dataset when combined with six query strategies: Random Sampling (RS), Breaking-Ties (BT), Greedy Core-Set (CS), Greedy Core-Set with t-SNE (CS-TSNE), Weighted Greedy Core-Set (WCS), and Ranked Weighted Greedy Core-Set (RWCS). The line represents the mean accuracy and the surrounding tubes represent the standard deviation over five runs.

The SetFit learning curves generally outperform the BERT model and seem to show much less variation between the different query strategies. When looking at the query strategies, we can see that specifically in the case of BERT, Core-Set seems to underperform especially in earlier iterations, even when compared to RS. This further reinforces the sentiment that Core-Set has its weaknesses when considering the task of text classification. We can see that in most cases, the three variations of Core-Set seem to be on par with or more performant than the original Core-Set. Even so, BT looks to be a strong

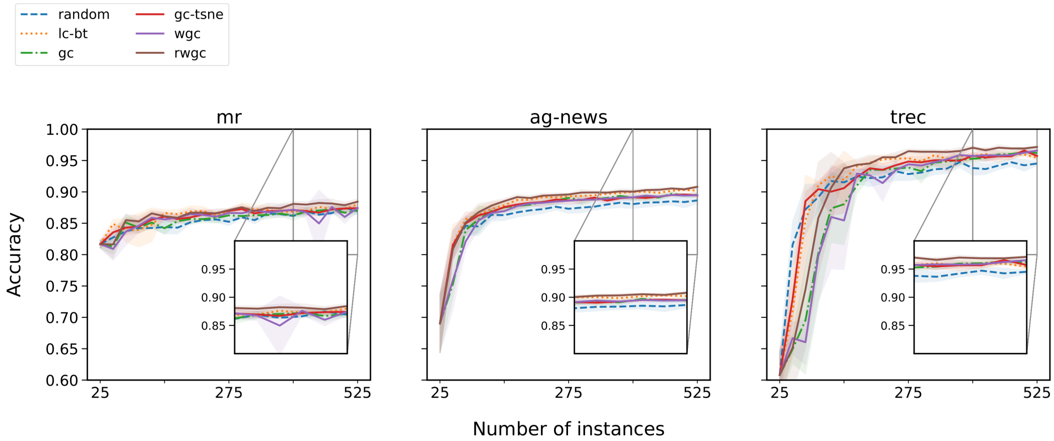

Figure 4.2: Active Learning curves of SetFit on each dataset when combined with six query strategies: Random Sampling (RS), Breaking-Ties (BT), Greedy Core-Set (CS), Greedy Core-Set with t-SNE (CS-TSNE), Weighted Greedy Core-Set (WCS), and Ranked Weighted Greedy Core-Set (RWCS). The line represents the mean accuracy and the surrounding tubes represent the standard deviation over five runs.

Dataset	**Model**	**Query Strategy**					
		RS	BT	CS	CS-TSNE	WCS	RWCS
AGN	BERT	0.884 ± 0.004	**0.889** ± **0.01**	0.874 ± 0.012	0.881 ± 0.01	0.873 ± 0.011	0.873 ± 0.022
	SetFit	0.886 ± 0.006	0.902 ± 0.004	0.895 ± 0.003	0.895 ± 0.003	0.895 ± 0.005	**0.908** ± **0.002**
MR	BERT	0.806 ± 0.011	0.815 ± 0.009	0.77 ± 0.015	0.796 ± 0.02	0.806 ± 0.014	**0.817** ± **0.01**
	SetFit	0.869 ± 0.006	0.88 ± 0.005	0.871 ± 0.005	0.874 ± 0.005	0.874 ± 0.007	**0.884** ± **0.005**
TREC	BERT	0.904 ± 0.018	0.92 ± 0.014	0.902 ± 0.021	0.912 ± 0.009	0.897 ± 0.027	**0.951** ± **0.008**
	SetFit	0.945 ± 0.004	0.954 ± 0.006	0.962 ± 0.004	0.957 ± 0.007	0.966 ± 0.004	**0.972** ± **0.003**

Table 4.1: Final accuracy per dataset, model, and query strategy. We report the mean and standard deviation over five runs. The best result per dataset is printed in bold.

Dataset	**Model**	**Query Strategy**					
		RS	BT	CS	CS-TSNE	WCS	RWCS
AGN	BERT	0.79 ± 0.015	0.8 ± 0.009	0.735 ± 0.02	0.792 ± 0.007	0.731 ± 0.014	**0.805** ± **0.01**
	SetFit	0.865 ± 0.007	0.881 ± 0.004	0.871 ± 0.005	0.875 ± 0.004	0.87 ± 0.003	**0.884** ± **0.002**
MR	BERT	0.75 ± 0.007	0.746 ± 0.011	0.718 ± 0.009	0.741 ± 0.017	0.72 ± 0.004	**0.759** ± **0.007**
	SetFit	0.854 ± 0.002	0.864 ± 0.005	0.856 ± 0.003	0.862 ± 0.003	0.858 ± 0.007	**0.866** ± **0.003**
TREC	BERT	0.674 ± 0.029	0.709 ± 0.008	0.594 ± 0.022	0.676 ± 0.037	0.629 ± 0.024	**0.771** ± **0.021**
	SetFit	0.914 ± 0.011	**0.923** ± **0.007**	0.896 ± 0.015	0.919 ± 0.005	0.893 ± 0.012	0.918 ± 0.015

Table 4.2: Final AUC per dataset, model, and query strategy. We report the mean and standard deviation over five runs. The best result per dataset is printed in bold.

contender in all cases and manages to perform similarly or better than CS-TSNE and WCS. The largest difference can be seen on the TREC-6 dataset, with RWCS outperforming all other strategies.

In 4.1, I show the reported mean accuracy and standard deviation for each combination of dataset, model and query strategy. The results show small improvements in accuracy for the dimensionality reduction approach, as well as Weighted Core-Set. However, the Ranked-Weighted Core-Set approach achieves the highest accuracy in almost all instances, being outperformed only on the AG-News dataset with BERT by the Breaking-Ties strategy.

4.2 contains the reported mean AUC (area under curve) and standard deviation for each combination of dataset, model and query strategy. The top performing AUC results, similarly to the accuracy table, show RWCS as having the best results in almost all cases.

As also made evident by the learning curves earlier, the accuracy and AUC improvements are most drastic in the case of BERT on the TREC dataset, where RWCS achieves an improvement in Core-Set's mean accuracy of 4.9 percent, and an improvement in AUC of 17.7 percent.

Chapter 5

Discussion

...

Chapter 6

Conclusion

...

Bibliography

Razvan Caramalau, Binod Bhattarai, and Tae-Kyun Kim. Sequential graph convolutional network for active learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 9583–9592. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.00946. URL `https://openaccess.thecvf.com/content/CVPR2021/html/Caramalau_Sequential_Graph_Convolutional_Network_for_Active_Learning_CVPR_2021_paper.html`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/n19-1423. URL `https://doi.org/10.18653/v1/n19-1423`.

Liat Ein-Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim. Active learning for BERT: an empirical study. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 7949–7962. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.EMNLP-MAIN.638. URL `https://doi.org/10.18653/v1/2020.emnlp-main.638`.

Damien François. High-dimensional data analysis. 2007.

M Ikonomakis, Sotiris Kotsiantis, Vasilis Tampakas, et al. Text classification using machine learning techniques. *WSEAS transactions on computers*, 4 (8):966–974, 2005.

Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura E. Barnes, and Donald E. Brown. Text classification algorithms: A survey. *Inf.*, 10(4):150, 2019. doi: 10.3390/info10040150. URL `https://doi.org/10.3390/info10040150`.

David D. Lewis and William A. Gale. A sequential algorithm for training text classifiers. In W. Bruce Croft and C. J. van Rijsbergen, editors, *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. Dublin, Ireland, 3-6 July 1994 (Special Issue of the SIGIR Forum)*, pages 3–12. ACM/Springer, 1994. doi: 10.1007/978-1-4471-2099-5_1. URL `https://doi.org/10.1007/978-1-4471-2099-5_1`.

Xin Li and Dan Roth. Learning question classifiers: the role of semantic information. *Nat. Lang. Eng.*, 12(3):229–249, 2006. doi: 10.1017/S1351324905003955. URL `https://doi.org/10.1017/S1351324905003955`.

Bing Liu and Lei Zhang. A survey of opinion mining and sentiment analysis. In Charu C. Aggarwal and ChengXiang Zhai, editors, *Mining Text Data*, pages 415–463. Springer, 2012. doi: 10.1007/978-1-4614-3223-4_13. URL `https://doi.org/10.1007/978-1-4614-3223-4_13`.

Qiang Liu, Yanqiao Zhu, Zhaocheng Liu, Yufeng Zhang, and Shu Wu. Deep active learning for text classification with diverse interpretations. In Gianluca Demartini, Guido Zuccon, J. Shane Culpepper, Zi Huang, and Hanghang Tong, editors, *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*, pages 3263–3267. ACM, 2021. doi: 10.1145/3459637.3482080. URL `https://doi.org/10.1145/3459637.3482080`.

Tong Luo, Kurt Kramer, Dmitry B. Goldgof, Lawrence O. Hall, Scott Samson, Andrew Remsen, and Thomas Hopkins. Active learning to recognize multiple types of plankton. *J. Mach. Learn. Res.*, 6:589–613, 2005. URL `http://jmlr.org/papers/v6/luo05a.html`.

Marcin Mironczuk and Jaroslaw Protasiewicz. A recent overview of the state-of-the-art elements of text classification. *Expert Syst. Appl.*, 106:36–54, 2018.

doi: 10.1016/j.eswa.2018.03.058. URL https://doi.org/10.1016/j.eswa.2018.03.058.

Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In Kevin Knight, Hwee Tou Ng, and Kemal Oflazer, editors, *ACL 2005, 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 25-30 June 2005, University of Michigan, USA*, pages 115–124. The Association for Computer Linguistics, 2005. doi: 10.3115/1219840.1219855. URL https://aclanthology.org/P05-1015/.

Sumanth Prabhu, Moosa Mohamed, and Hemant Misra. Multi-class text classification using bert-based active learning. In Eduard C. Dragut, Yunyao Li, Lucian Popa, and Slobodan Vucetic, editors, *3rd Workshop on Data Science with Human in the Loop, DaSH@KDD, Virtual Conference, August 15, 2021*, 2021. URL https://drive.google.com/file/d/1xVy4p29UPINmWl8Y7OospyQgHiYfH4wc/view.

Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47, 2002. doi: 10.1145/505282.505283. URL https://doi.org/10.1145/505282.505283.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=H1aIuk-RW.

Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 5971–5980. IEEE, 2019. doi: 10.1109/ICCV.2019.00607. URL https://doi.org/10.1109/ICCV.2019.00607.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. Efficient few-shot learning without prompts. *CoRR*, abs/2209.11055, 2022. doi: 10.48550/arXiv.2209.11055. URL https://doi.org/10.48550/arXiv.2209.11055.

Laurens Van der Maaten. Laurens Van der Maaten home page. https://lvdmaaten.github.io/tsne/, 2023. Accessed: 2023-09-24.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Laurens Van Der Maaten, Eric O Postma, H Jaap van den Herik, et al. Dimensionality reduction: A comparative review. *Journal of Machine Learning Research*, 10(66-71):13, 2009.

Martin Wattenberg, Fernanda Viégas, and Ian Johnson. How to use t-sne effectively. *Distill*, 2016. doi: 10.23915/distill.00002. URL `http://distill.pub/2016/misread-tsne`.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 649–657, 2015. URL `https://proceedings.neurips.cc/paper/2015/hash/250cf8b51c773f3f8dc8b4be867a9a02-Abstract.html`.