UNIVERSITÄT
LEIPZIG

Bachelorseminar

Investigating Core Set-based Active Learning for Text Classification

Leipzig, 11.01.2024

Yannick Brenning

STRUCTURE

1. Motivation
2. Related Work (Text Classification, Active Learning, Core-Set)
3. Approach/Methods
4. Experiment
5. Conclusion

MOTIVATION

- Large amounts of unstructured textual data available
- Efficiently classify documents using active learning

  → Use querying to select instances for labelling

MOTIVATION

- Large amounts of unstructured textual data available
- Efficiently classify documents using active learning
  → Use querying to select instances for labelling

Core-Set (Sener and Savarese [2018])

- Diversity-based query strategy using point distances
- Originally for Computer Vision

MOTIVATION

- Core-Set shows mixed results in text classification
  (Ein-Dor et al. [2020], Prabhu et al. [2021], Liu et al. [2021])

- Mixed results in CV tasks with higher dimensions and higher class numbers
  (Sinha et al. [2019])

TEXT CLASSIFICATION

Idea: Assign a category or class to document or piece of text.

- Subfield of NLP (Natural Language Processing)

TEXT CLASSIFICATION

Idea: Assign a category or class to document or piece of text.

- Subfield of NLP (Natural Language Processing)

Applications:

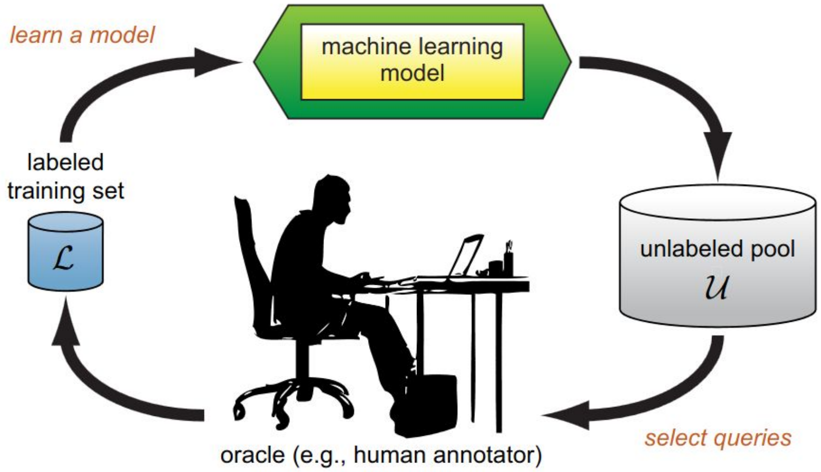
- Spam filtering (binary)
- Sentiment analysis
- News and content categorization (multi-class)
- Information retrieval, document summarization,...

ACTIVE LEARNING

- Subfield of machine learning
- Classifier performs queries on an information source

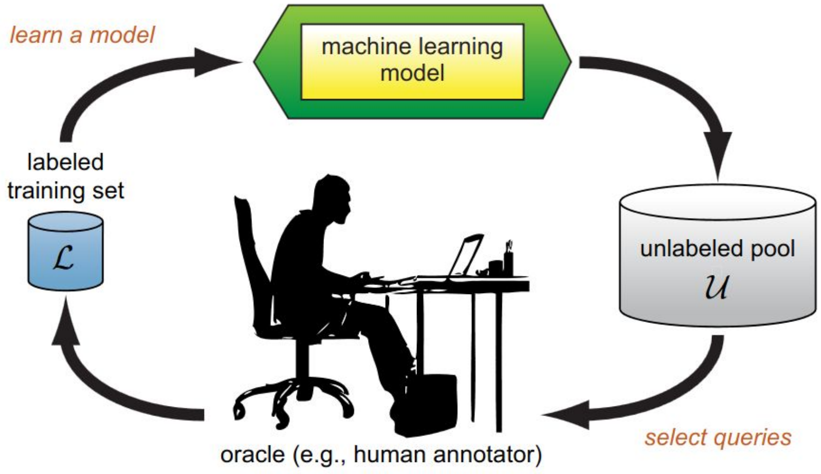
- Reduce total amount of annotated data
  - Large amounts of unlabeled data
  - Manual labeling is expensive
  - Limited annotation resources

ACTIVE LEARNING



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

ACTIVE LEARNING



- **query strategy** decides which instances to select for labeling
- selecting informative instances important for model’s success

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

CORE-SET

- Used as a pool-based query strategy
- Introduced in 2017 by Sener and Saverese

- Deep learning domain
  - Originally developed for CNNs
  - Computer vision tasks
- Also applied to text classification tasks
  - BERT-based AL (Ein-Dor et al. [2020], Prabhu et al. [2021])

CORE-SET

- diversity-based approach
- selects subset of instances that best cover the total dataset
- k-Center problem solved using greedy approach

CORE-SET

Algorithm 1 k-Center-Greedy

Input: data $\mathbf{x}_i$, existing pool $\mathbf{s}^0$, budget b

Initialize $\mathbf{s} = \mathbf{s}^0$

repeat

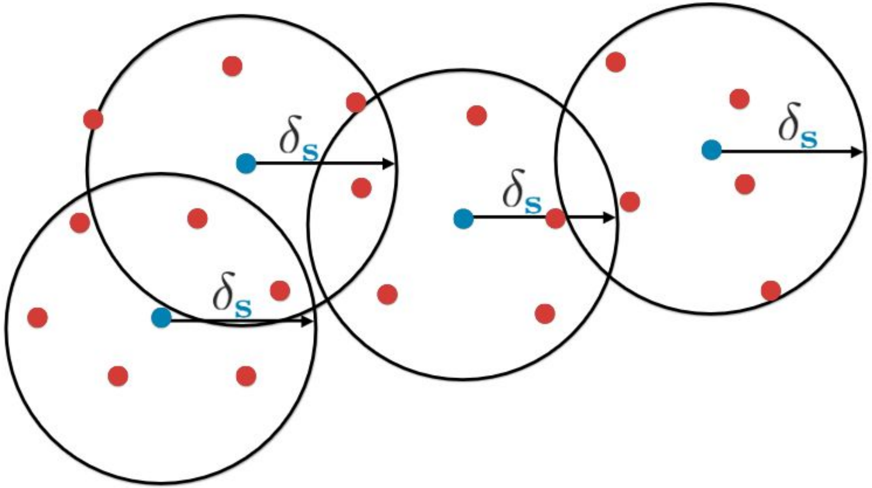
$u = \mathrm{argmax}_{i \in [n] \setminus \mathbf{s}} \min_{j \in \mathbf{s}} \Delta(\mathbf{x}_i, \mathbf{x}_j)$

$\mathbf{s} = \mathbf{s} \cup \{u\}$

until $|\mathbf{s}| = b + |\mathbf{s}^0|$

return $\mathbf{s} \setminus \mathbf{s}^0$

CORE-SET



Source: Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. 2018.

CORE-SET: RECENT RESEARCH

- Mixed results when applied to text classification tasks
  (Ein-Dor et al. [2020], Prabhu et al. [2021], Liu et al. [2021])

- Mixed results in CV tasks with higher dimensions and higher class numbers
  (Sinha et al. [2019])

APPROACH

1. Dimensionality Reduction

2. Uncertainty-Based

3. Class Balance-Based

DIMENSIONALITY REDUCTION APPROACH

- Core-Set may suffer from higher dimensionality (Sinha et al. [2019])
- Phenomenon known as “curse of dimensionality”

Techniques to transform data from high to lower dimension

- Linear techniques
  - PCA, LDA, NMF etc.
- Non-linear techniques
  - Isomap, TSNE etc.

DIMENSIONALITY REDUCTION APPROACH

Algorithm 2 t-SNE with k-Center-Greedy

Input: data $\mathbf{x}_i$, existing pool $\mathbf{s}^0$, budget b

Initialize $\mathbf{s} = \mathbf{s}^0$

repeat

$u = \text{argmax}_{i \in [n] \setminus \mathbf{s}} \min_{j \in \mathbf{s}} \Delta_{TSNE}(\mathbf{x}_i, \mathbf{x}_j)$ ▷ Computing distances on reduced embeddings

$\mathbf{s} = \mathbf{s} \cup \{u\}$

until $|\mathbf{s}| = b + |\mathbf{s}^0|$

return $\mathbf{s} \setminus \mathbf{s}^0$

UNCERTAINTY-BASED APPROACHES

Least Confidence (Lewis and Gale [1994])

- Select instances with lowest classification certainties

Breaking-Ties (Luo et al. [2005])

- Select instances with smallest margin between most likely classes

UNCERTAINTY-BASED APPROACHES

Least Confidence (Lewis and Gale [1994])

- Select instances with lowest classification certainties

Breaking-Ties (Luo et al. [2005])

- Select instances with smallest margin between most likely classes

For any instance $\mathbf{x}_i$, let $\mathbf{p}^*_{j,k}$ denote the probability of the k-th most likely class label for $\mathbf{x}_i$. Then Breaking-Ties attempts to select:

$$\operatorname{argmin}_{\mathbf{x}_i} (\mathbf{p}^*_{i,1} - \mathbf{p}^*_{i,2})$$

UNCERTAINTY-BASED APPROACHES

“Weighted Core-Set”

- Compute BT scores using class label probabilities
- Use weights to combine BT scores with CS distances

“Re-ranked Core-Set”

- For a sample of size n, compute Core-Set of size 2n
- Take n best BT scores

WEIGHTED CORE-SET

Algorithm 3 Weighted k-Center-Greedy

Input: $\mathbf{x}_i$, $\mathbf{s}^0$, b, breaking-ties probabilities $\mathbf{p}_{bt}$

Initialize $\mathbf{s} = \mathbf{s}^0$

repeat

$u = \text{argmax}_{i \in [n] \setminus \mathbf{s}} \min_{j \in \mathbf{s}} \Delta(\mathbf{x}_i, \mathbf{x}_j)$

$\mathbf{s} = \mathbf{s} \cup \{u\}$

until $|\mathbf{s}| = b + |\mathbf{s}^0|$

$\mathbf{s} = 0.8 \cdot \mathbf{s} + 0.2 \cdot \mathbf{p}_{bt}$ ▷ Weigh results using linear combination

return $\mathbf{s} \setminus \mathbf{s}^0$

RE-RANKED CORE-SET

Algorithm 3 Re-ranked k-Center-Greedy

Input: $\mathbf{x}_i$, $\mathbf{s}^0$, b, class probabilities $\mathbf{p}_i$

Initialize $\mathbf{s} = \mathbf{s}^0, r = \emptyset$

repeat

$u = \text{argmax}_{i \in [n] \setminus \mathbf{s}} \text{min}_{j \in \mathbf{s}} \Delta(\mathbf{x}_i, \mathbf{x}_j)$

$\mathbf{s} = \mathbf{s} \cup \{u\}$

until $|\mathbf{s}| = 2b + |\mathbf{s}^0|$ $\triangleright$ Compute Core-Set of size $2b$

repeat

$u = \text{argmin}_{j \in \mathbf{s} \setminus r} \mathbf{p}^*_{j,1} - \mathbf{p}^*_{j,2}$

$r = r \cup \{u\}$

until $|r| = b$ $\triangleright$ Compute the b-highest BT-scores

return r

Where $\mathbf{p}^*_{j,k}$ denotes the probability of the k-th most likely class label for the j-th instance

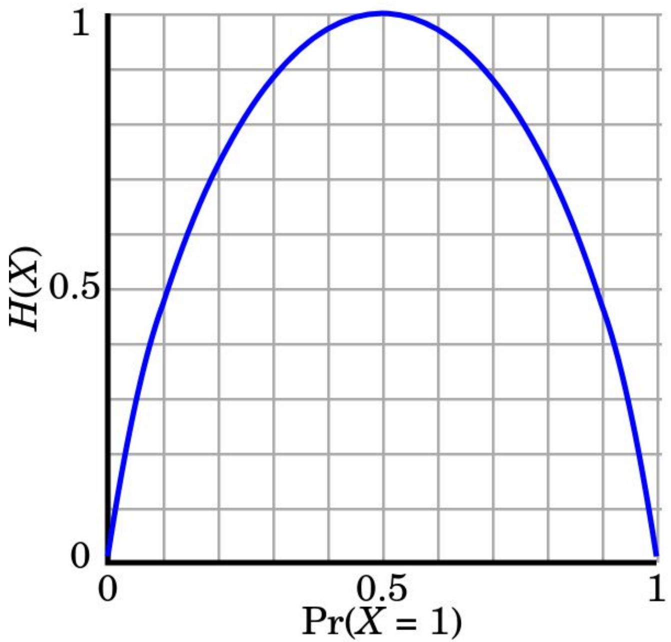
CLASS BALANCE-BASED APPROACH

- Attempt to balance class distribution in Core-Set
- Classes are more balanced if normalized entropy is closer to 1

$$H(X) = -\sum_{x \in X} p(x) \log p(x)$$

$$H_{norm}(X) = H(X) / \log n$$

CLASS BALANCE-BASED APPROACH



Source: https://en.wikipedia.org/wiki/Binary_entropy_function

RESEARCH QUESTIONS

1. Can we use dimensionality reduction to improve Core-Set for text classification tasks?

2. Can we improve Core-Set using an uncertainty-based approach?

3. How do class imbalances impact Core-Set's performance?

EXPERIMENT: DATA

Movie Review Dataset (Pang and Lee [2005])

- Sentiment analysis dataset (binary classification)
- 10,662 movie reviews

AG’s News Dataset (Zhang et al. [2015])

- Multi-class news dataset
- 127,600 news articles

TREC Dataset (Li and Roth [2006])

- Question classification
- 6,000 questions

EXPERIMENT: CLASSIFIERS

- BERT (Bidirectional Encoder Representations from Transformers)

- SetFit (Sentence Transformer Fine-tuning)
  - Based on sentence transformers
  - Fine-tuned using contrastive representation learning

EXPERIMENT: SETUP

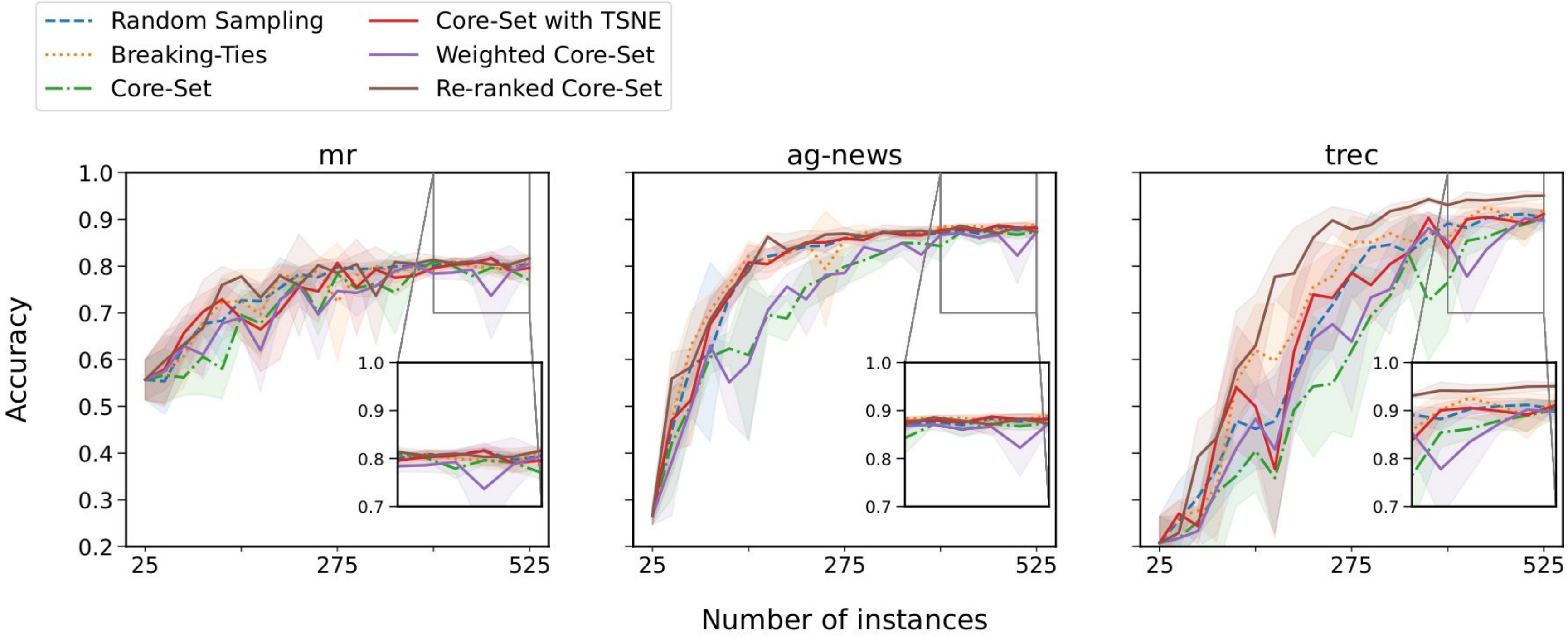
- BERT (Bidirectional Encoder Representations from Transformers)

- SetFit (Sentence Transformer Fine-tuning)

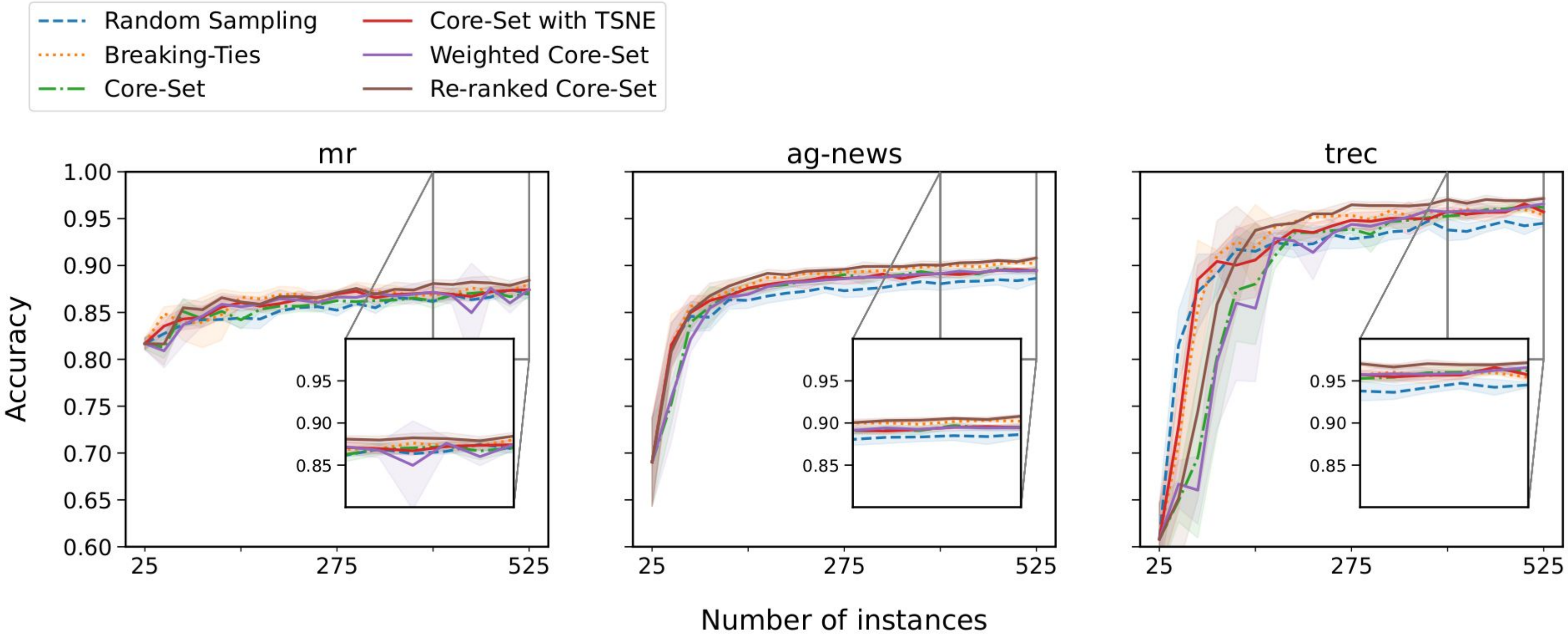
- 20 queries on 25 instances
- 5 runs of queries per combination of dataset, model, and query strategy

- Query strategy baselines: Random Sampling, Breaking-Ties, Core-Set

EXPERIMENT: RESULTS (BERT)



EXPERIMENT: RESULTS (SETFIT)



EXPERIMENT: RESULTS

Dataset	Model	Query Strategy					
		RS	BT	CS	CS-TSNE	WCS	RCS
AGN	BERT	0.884 ± 0.004	$\mathbf{0.889 \pm 0.010}$	0.874 ± 0.012	0.881 ± 0.010	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	$\mathbf{0.908 \pm 0.002}$
MR	BERT	0.806 ± 0.011	0.815 ± 0.009	0.77 ± 0.015	0.796 ± 0.020	0.806 ± 0.014	$\mathbf{0.817 \pm 0.010}$
	SetFit	0.869 ± 0.006	0.88 ± 0.005	0.871 ± 0.005	0.874 ± 0.005	0.874 ± 0.007	$\mathbf{0.884 \pm 0.005}$
TREC	BERT	0.904 ± 0.018	0.92 ± 0.014	0.902 ± 0.021	0.912 ± 0.009	0.897 ± 0.027	$\mathbf{0.951 \pm 0.008}$
	SetFit	0.945 ± 0.004	0.954 ± 0.006	0.962 ± 0.004	0.957 ± 0.007	0.966 ± 0.004	$\mathbf{0.972 \pm 0.003}$

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.

EXPERIMENT: RESULTS

Dataset	Model	Query Strategy					
		RS	BT	CS	CS-TSNE	WCS	RCS
AGN	BERT	0.790 ± 0.015	0.800 ± 0.009	0.735 ± 0.020	0.792 ± 0.007	0.731 ± 0.014	$\mathbf{0.805 \pm 0.010}$
	SetFit	0.865 ± 0.007	0.881 ± 0.004	0.871 ± 0.005	0.875 ± 0.004	0.87 ± 0.003	$\mathbf{0.884 \pm 0.002}$
MR	BERT	0.750 ± 0.007	0.746 ± 0.011	0.718 ± 0.009	0.741 ± 0.017	0.720 ± 0.004	$\mathbf{0.759 \pm 0.007}$
	SetFit	0.854 ± 0.002	0.864 ± 0.005	0.856 ± 0.003	0.862 ± 0.003	0.858 ± 0.007	$\mathbf{0.866 \pm 0.003}$
TREC	BERT	0.674 ± 0.029	0.709 ± 0.008	0.594 ± 0.022	0.676 ± 0.037	0.629 ± 0.024	$\mathbf{0.771 \pm 0.021}$
	SetFit	0.914 ± 0.011	$\mathbf{0.923 \pm 0.007}$	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.

EXPERIMENT: RESULTS

- Core-Set underperforming in early iterations with BERT
- Minor improvements with dimensionality reduction and uncertainty weights
- Re-ranked core-set especially effective on TREC
- Uncertainty-based approach generally performant in all instances

EXPERIMENT: OUTLOOK

- Examine class-balanced Core-Sets
- Consider other dimensionality reduction techniques
  - Effect of reducing to different dimensions
- Combining different approaches (reduction, probabilities, class balances)
- Examine effect of hyperparameters when performing reduction

CONCLUSION

Experiment

- Using dimensionality reduction with Core-Set
- Examined Core-Set in conjunction with pointwise probabilities
- 2 models (BERT, SetFit), 3 datasets

Findings

- Minor improvements of Core-Set in nearly all cases
- Re-ranking improves Core-Set’s efficiency

REFERENCES

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.

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.

REFERENCES

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 Weber, 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.

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.

REFERENCES

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.

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.

REFERENCES

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.

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.

REFERENCES

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/.

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.

IMPLEMENTATION DETAILS

- NumPy https://numpy.org/
- Scikit-Learn https://scikit-learn.org/stable/
- Matplotlib https://matplotlib.org/
- Small-Text https://github.com/webis-de/small-text
- HuggingFace https://huggingface.co/SetFit

IMPLEMENTATION DETAILS

t-SNE hyperparameters:

- No. of components:2
- Perplexity: 30
- No. of iterations: 1000
- Initialization Method: PCA

Model Names:

- BERT-base-uncased
- Paraphrase-MPNET-base-v2