ENSEMBLE FEATURE SELECTION FOR DOMAIN ADAPTATION IN SPEECH EMOTION RECOGNITION

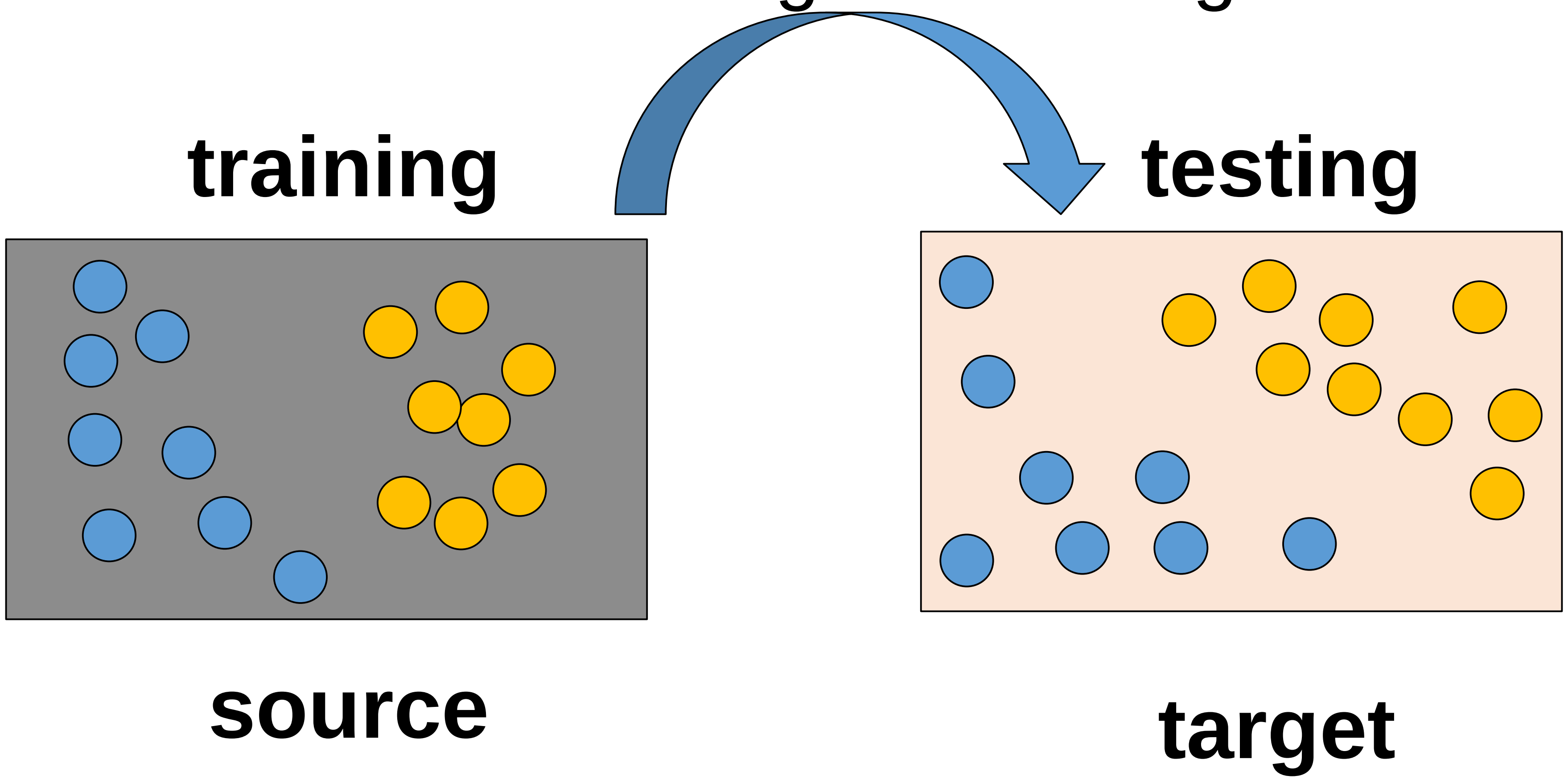
Mohammed Abdelwahab and Carlos Busso

Multimodal Signal Processing (MSP) lab
The University of Texas at Dallas
Erik Jonsson School of Engineering and Computer Science

March 2017

The Problem

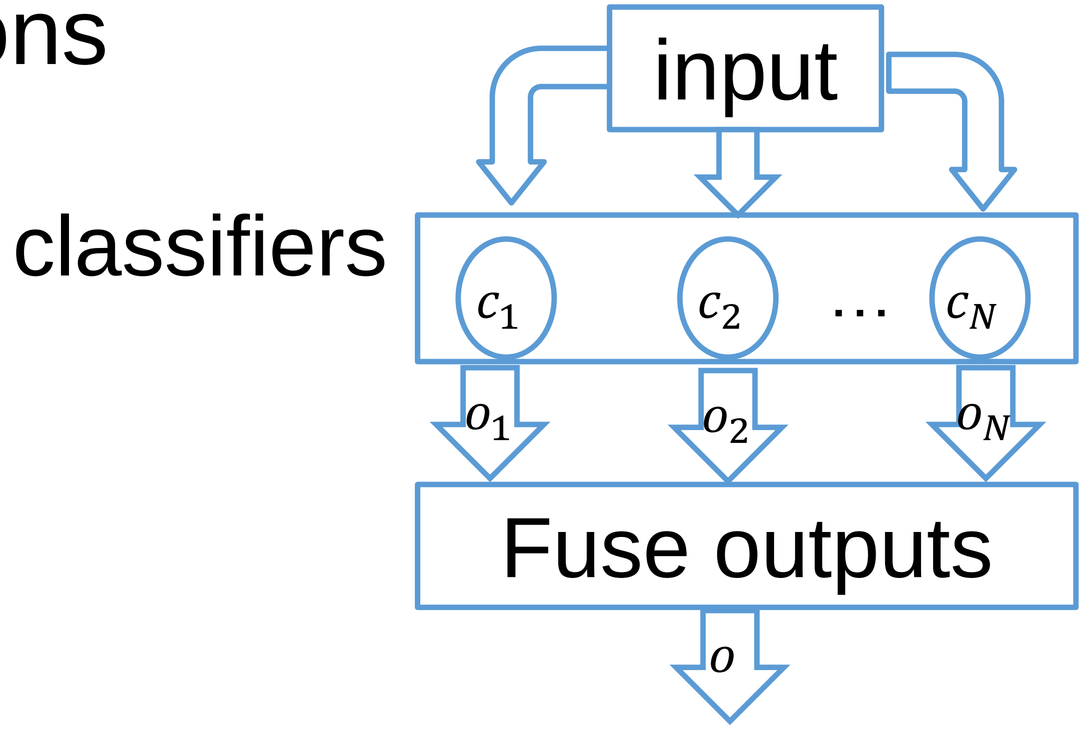
- The performance of a classifier degrades if the there is a mismatch between training and testing conditions



- Train system that recognize emotional categories using limited labeled data

Why Ensembles?

- Ensembles perform well in extreme scenarios with large or limited amounts of data
- The ensemble performance is better than its best classifier, under certain conditions
- Ensembles diversity
  - Using different data partitions
  - Using different sets of features
  - Using different classifier models
- Ensembles may mitigate the performance degradation



Related Work

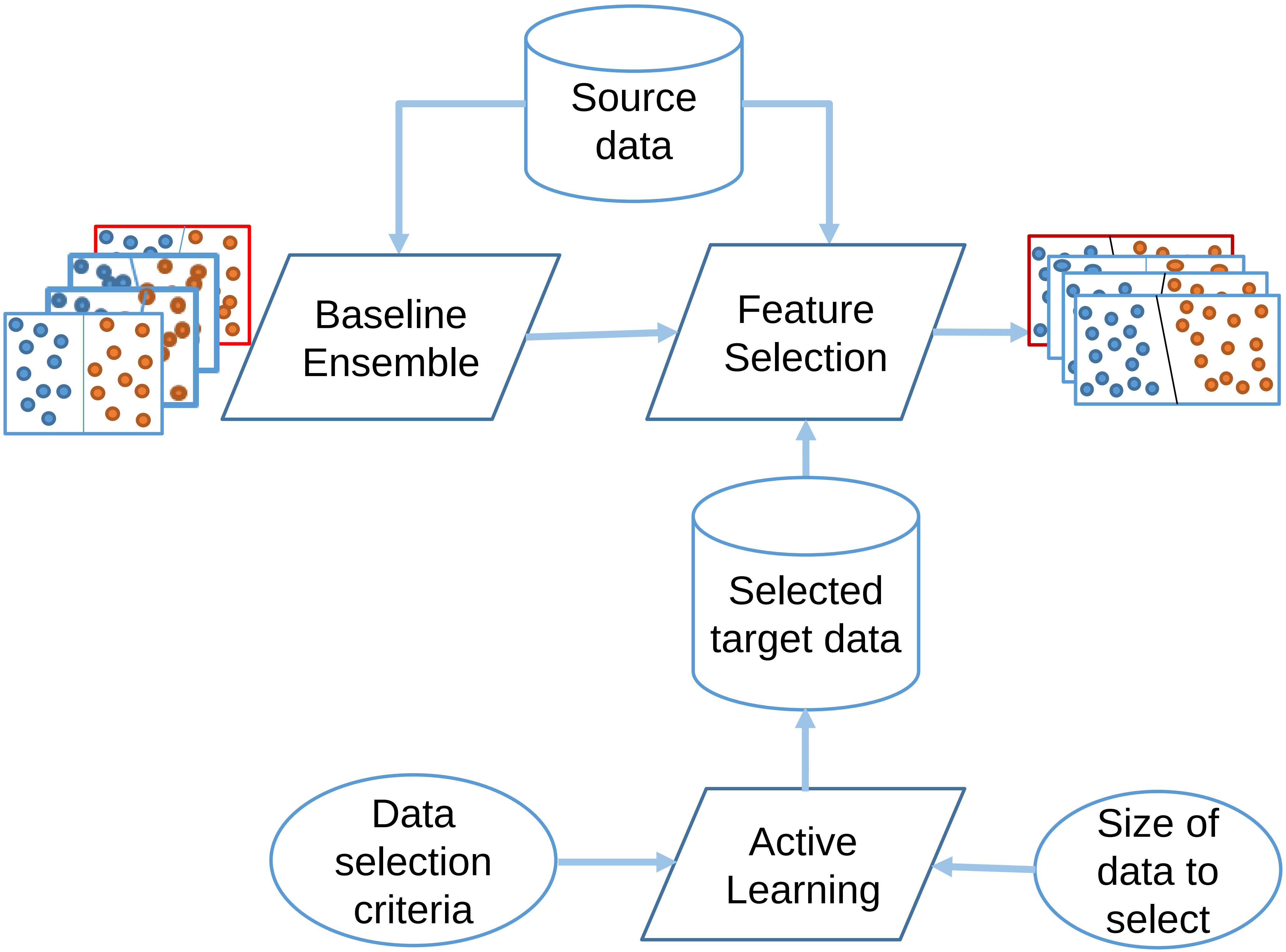
The main approaches to modify ensembles tested on new data

- Weighting the ensemble classifiers
- Weighting or resampling the source data to match the target distribution

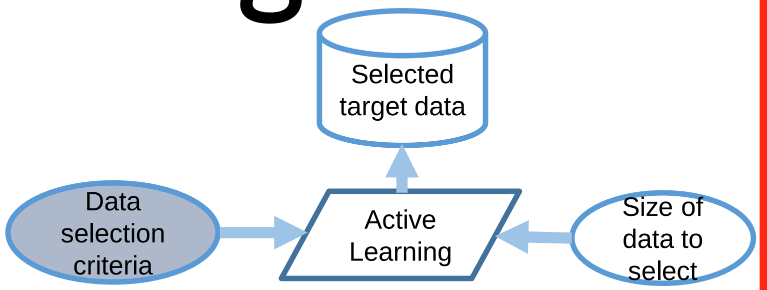
Our approach focuses on selecting a feature space that maximizes performance on the new data

Active Learning → Feature Selection → Train on Source → Test on Target

Proposed Approach



Data Selection Active Learning



- Vote Entropy

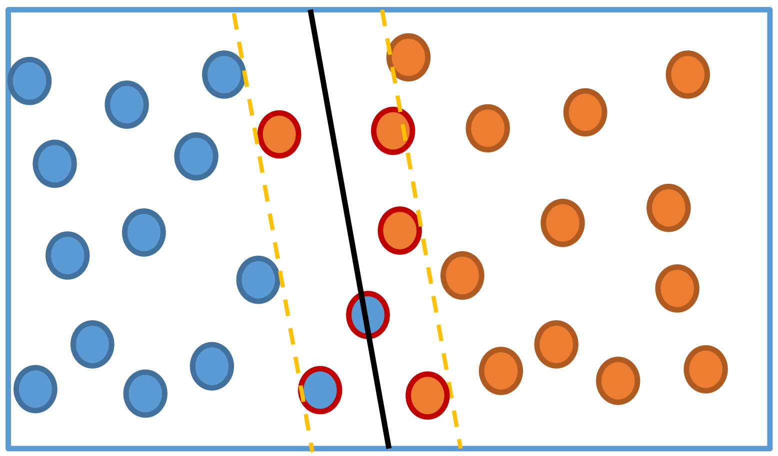
$$D(x) = -\sum_{c} \frac{V(c,x)}{k} \log(\frac{V(c,x)}{k})$$

k is the number of classifiers in ensemble

$V(c,x)$ is the number of classifiers assigning class c to sample x.

- Uncertainty Sampling

  Select the samples the classifier has the least confidence in.



- Random Sampling (passive learning)

Proposed Approach

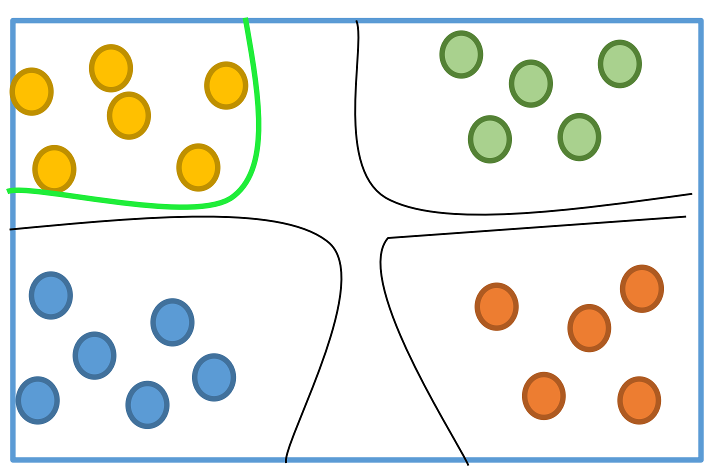
- The goal of the proposed approach
  - Minimize the mismatch between training and testing
  - Preserve the diversity of the ensemble
- This is achieved by:
  - Biasing different classifiers towards different classes
  - Eliminating overlap between feature sets used by the classifiers
  - Feature selection is conducted by maximizing the performance over the newly annotated data from the target domain

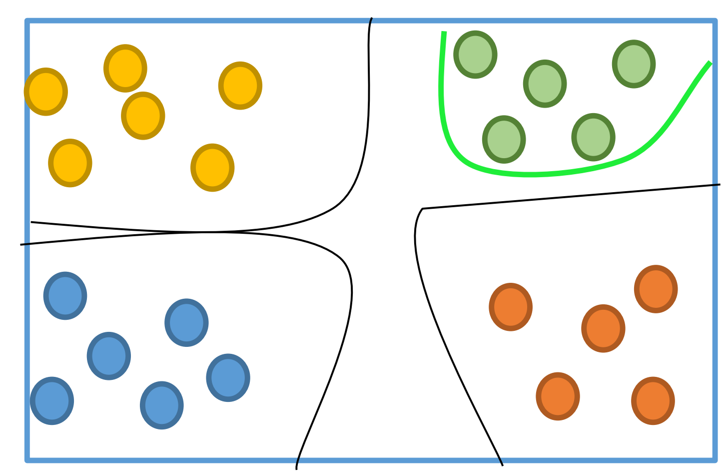
Proposed Approach

- We used F_2 score to bias the classifier towards a class

$$F_\beta = (1+\beta^2)\frac{precision * recall}{\beta^2 * precision + recall}$$

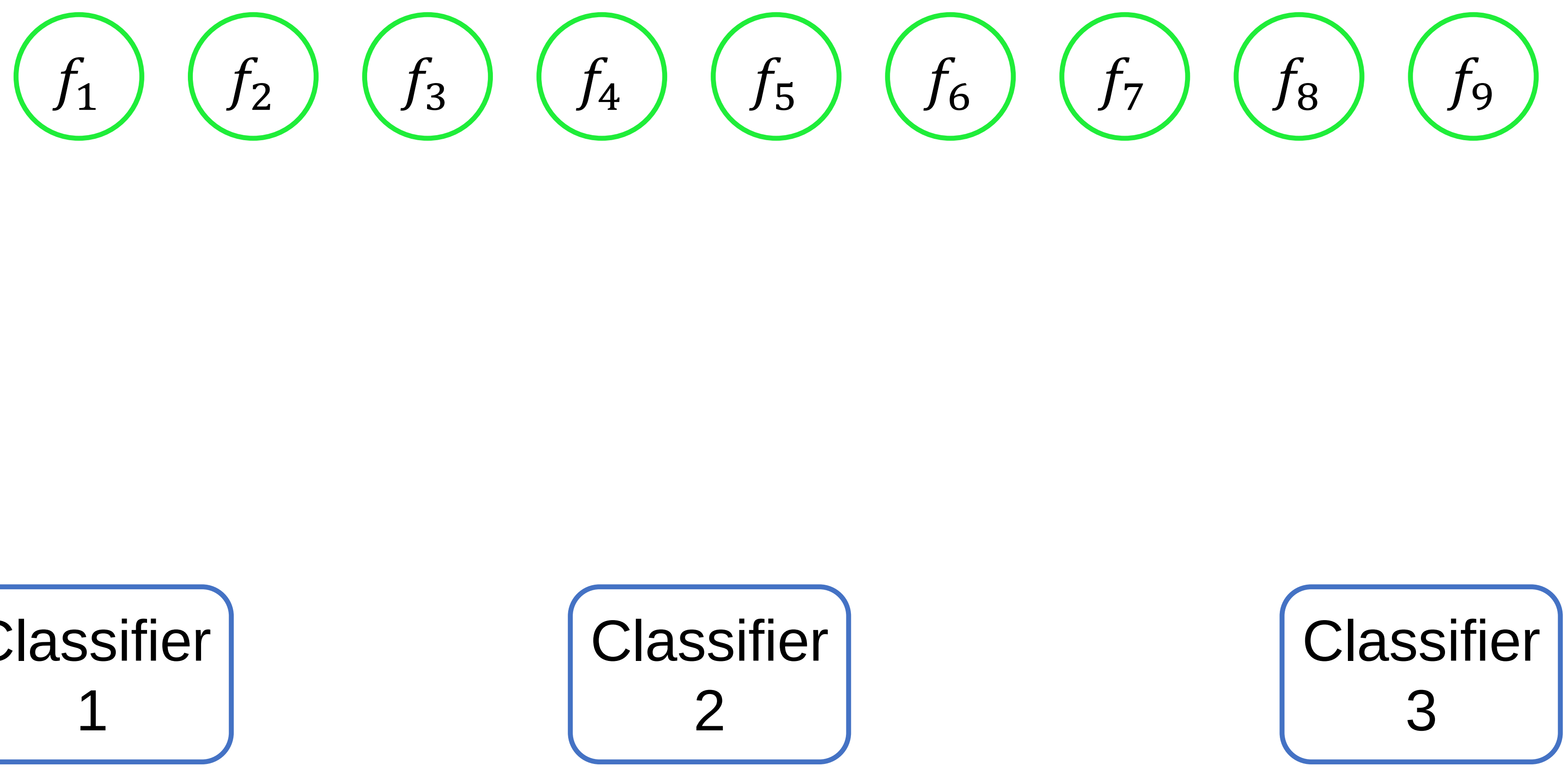
The classifier tries to maximize the F_2 of the selected class





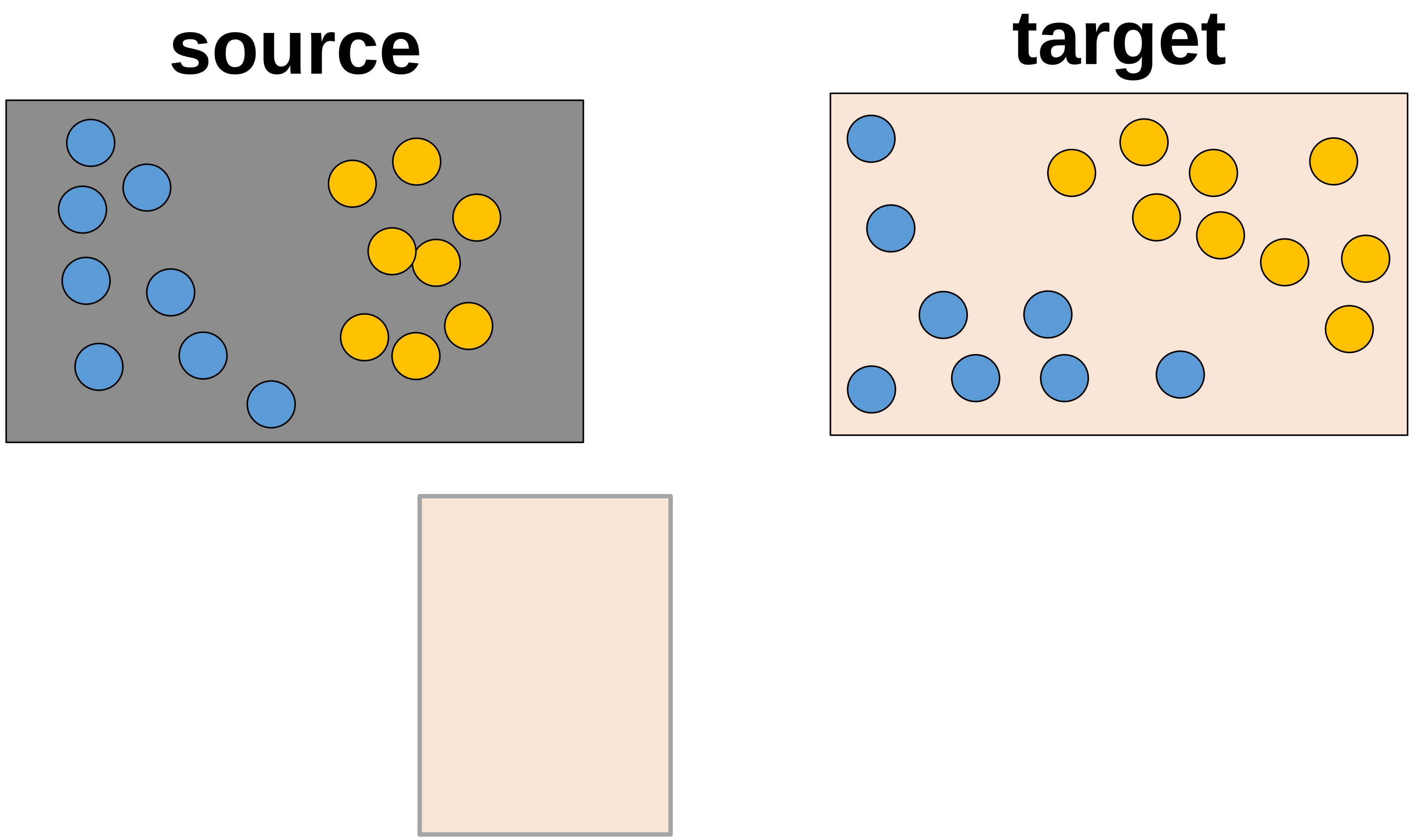
Proposed Approach

- Classifiers take turns selecting features
- Once a feature is selected, it is no longer available for the remaining classifiers



Proposed Approach

- Feature selection is conducted by maximizing the performance over the newly annotated data from the target domain



Databases

- <u>Train:</u> USC-IEMOCAP
  - 12 hours of conversational recordings from 10 actors in dyadic sessions
  - Sessions consists of emotional scripts as well as improvised interactions
  - Turns are annotated by 3 evaluators into categorical emotions
- <u>Test:</u> MSP-IMPROV
  - Dyadic interaction sessions from 12 actors
  - Contains 8,438 turns including improvised natural interactions
  - Turns are labeled into four categorical emotions as well as dimensional attribute scores by at least 5 annotators

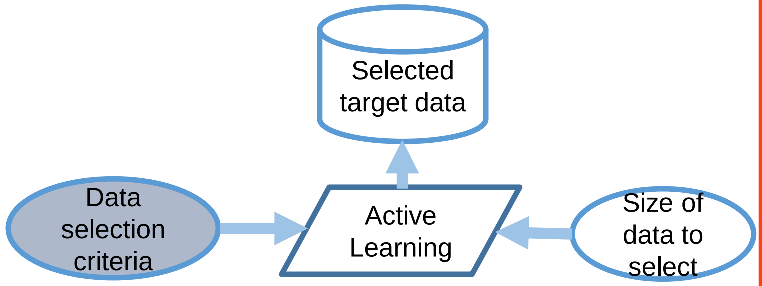
IEMOCAP

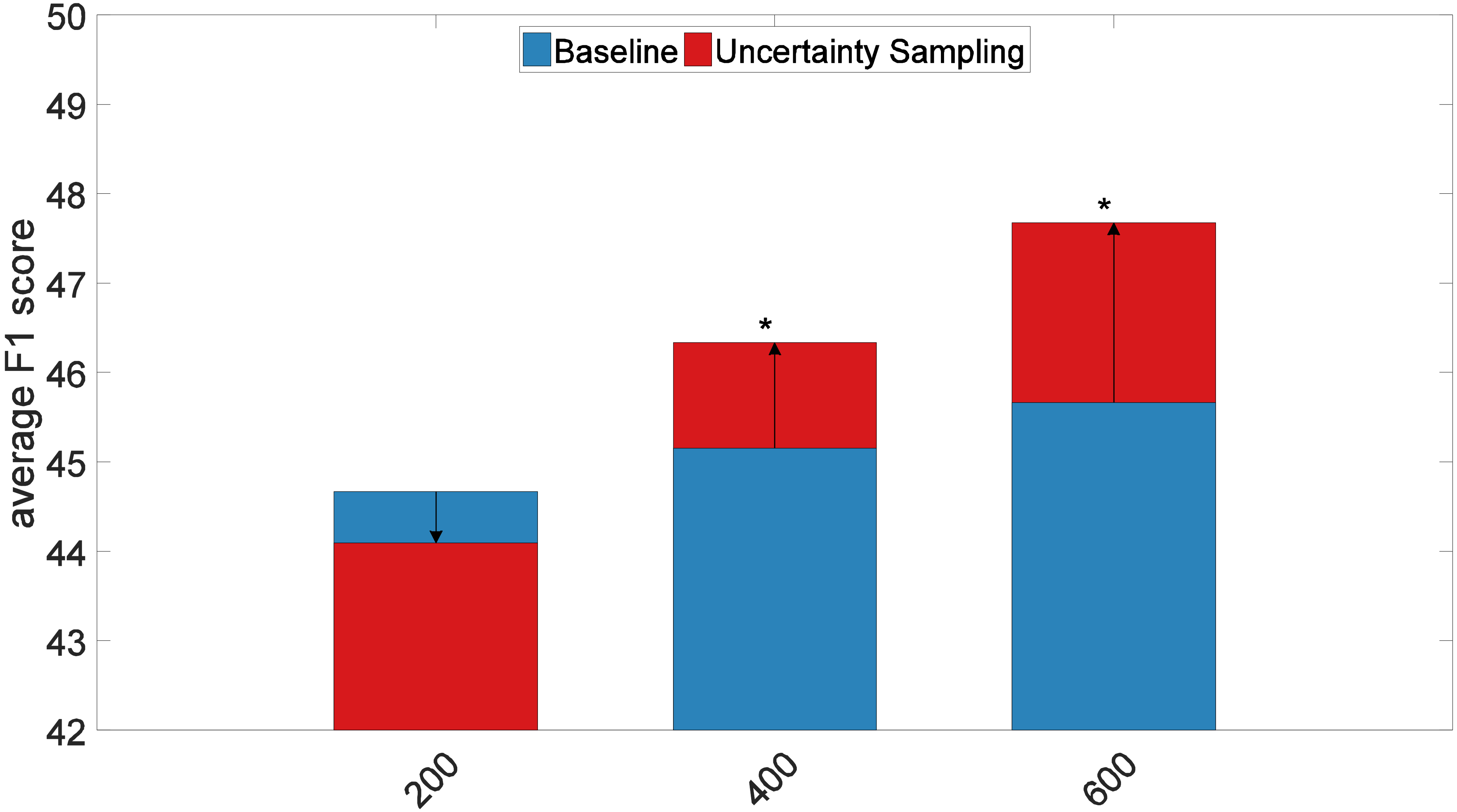
MSP-IMPROV

Experimental Settings

- Ensemble is composed of 40 SVM Classifiers
- Four class balanced classification problem
  - Angry, happy, sad, neutral
  - Random Under-sampling
- Classifier is trained on USC-IEMOCAP and tested on MSP-IMPROV
- Baseline is an ensemble of classifiers using features optimized on the source domain
- We used Interspeech 2013 feature set
  - Correlation Feature Selection 6,373 ⇨ 3,000
  - Each classifier selects 40 features

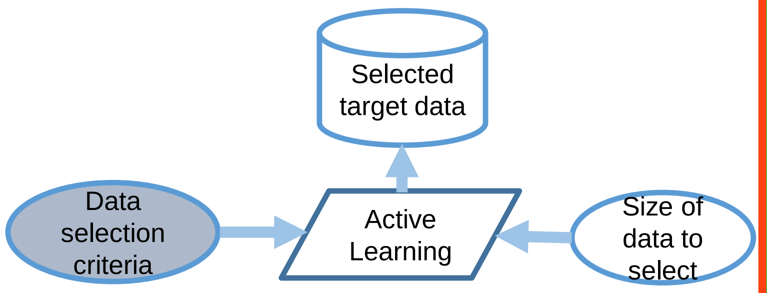
Comparing with the Baseline

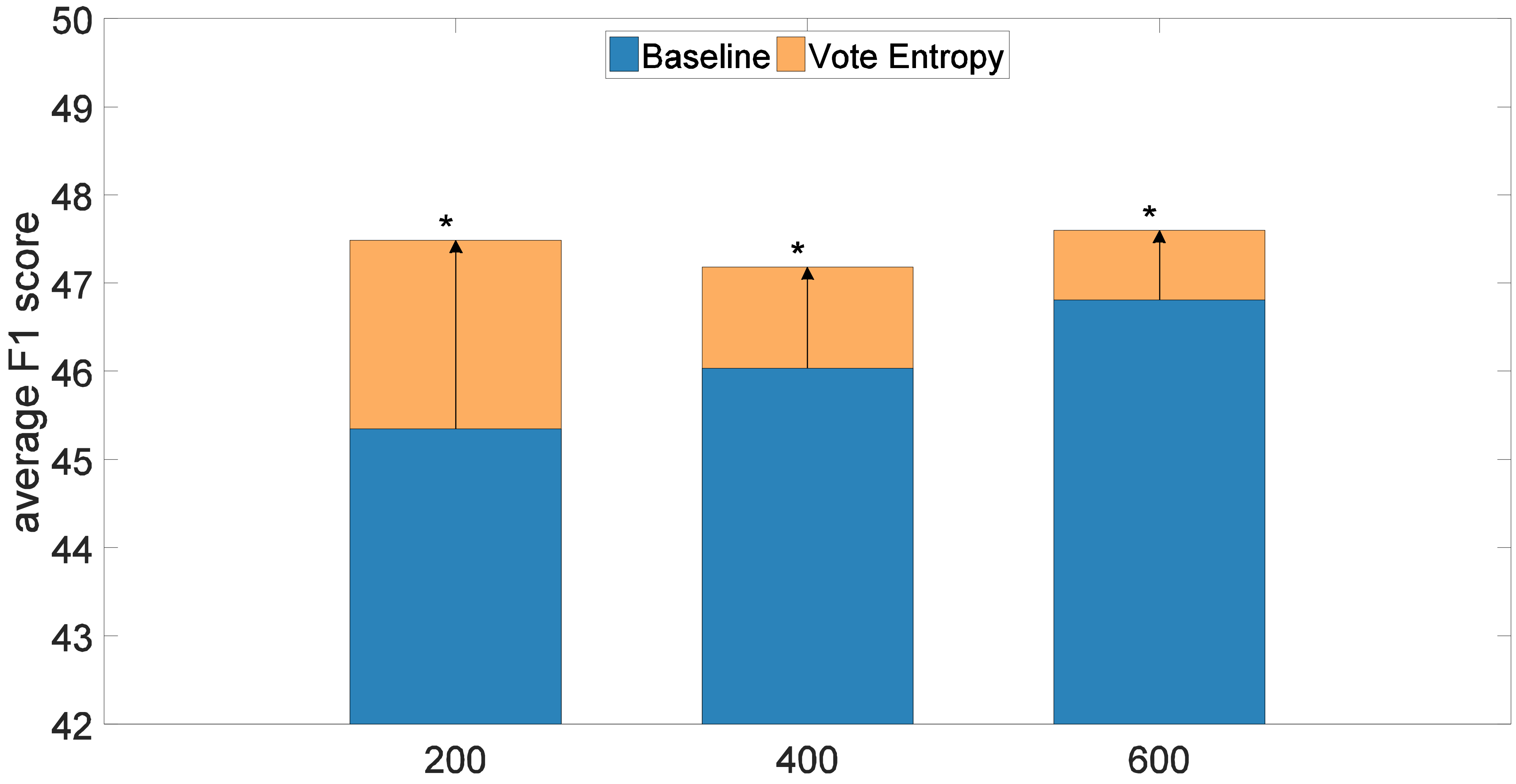




Uncertainty Sampling provides improvement over the baseline with more annotated data

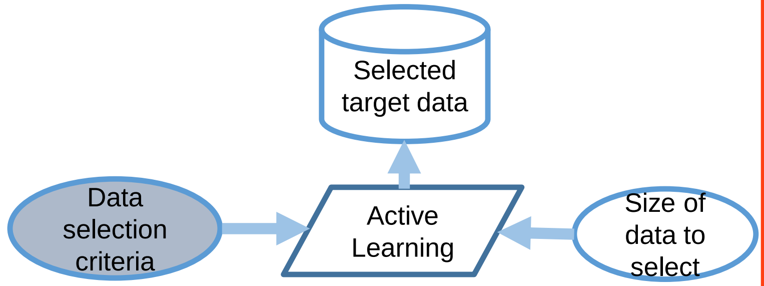
Comparing with the Baseline

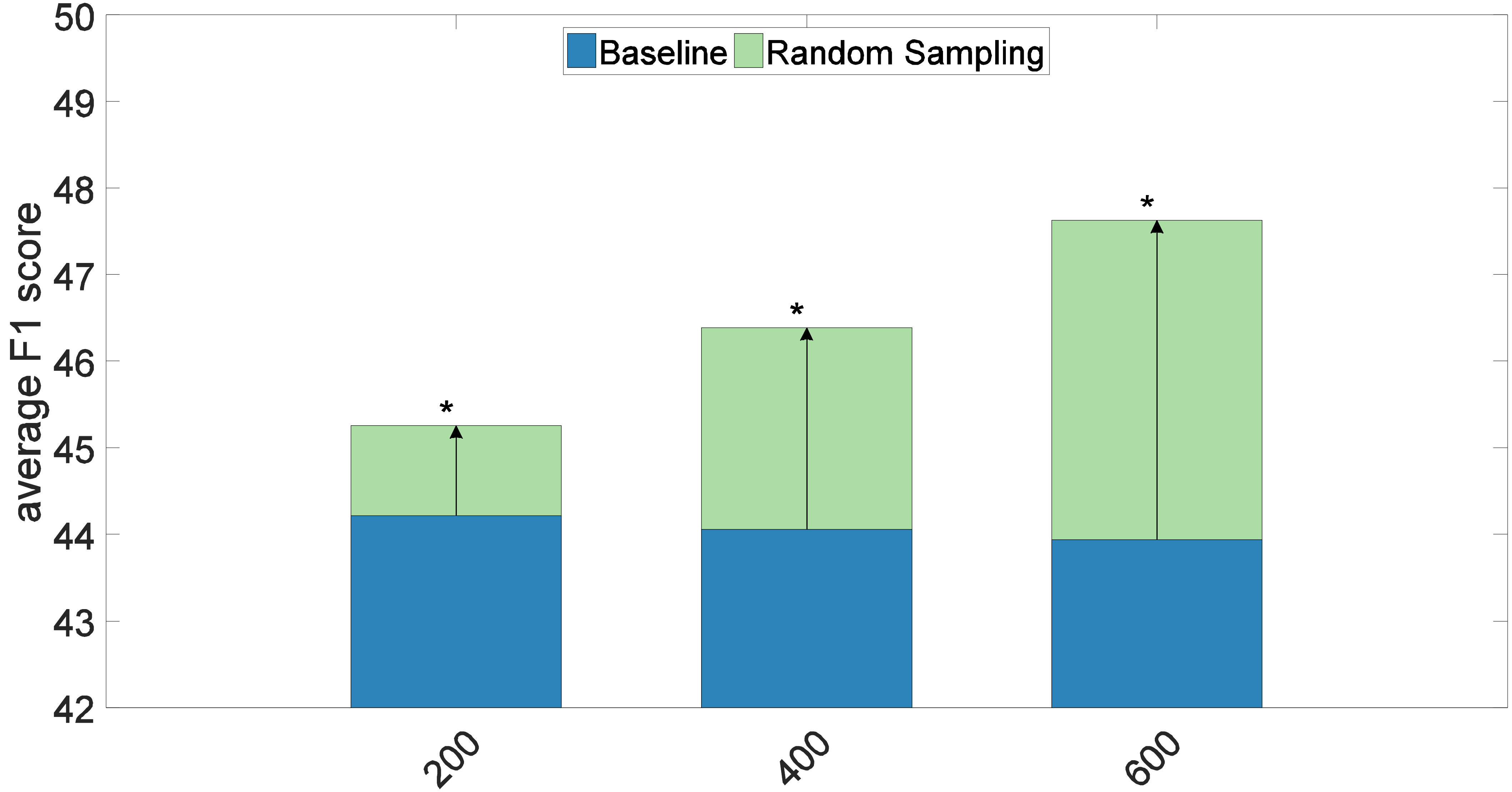




Vote Entropy performance gap drops as more data is selected

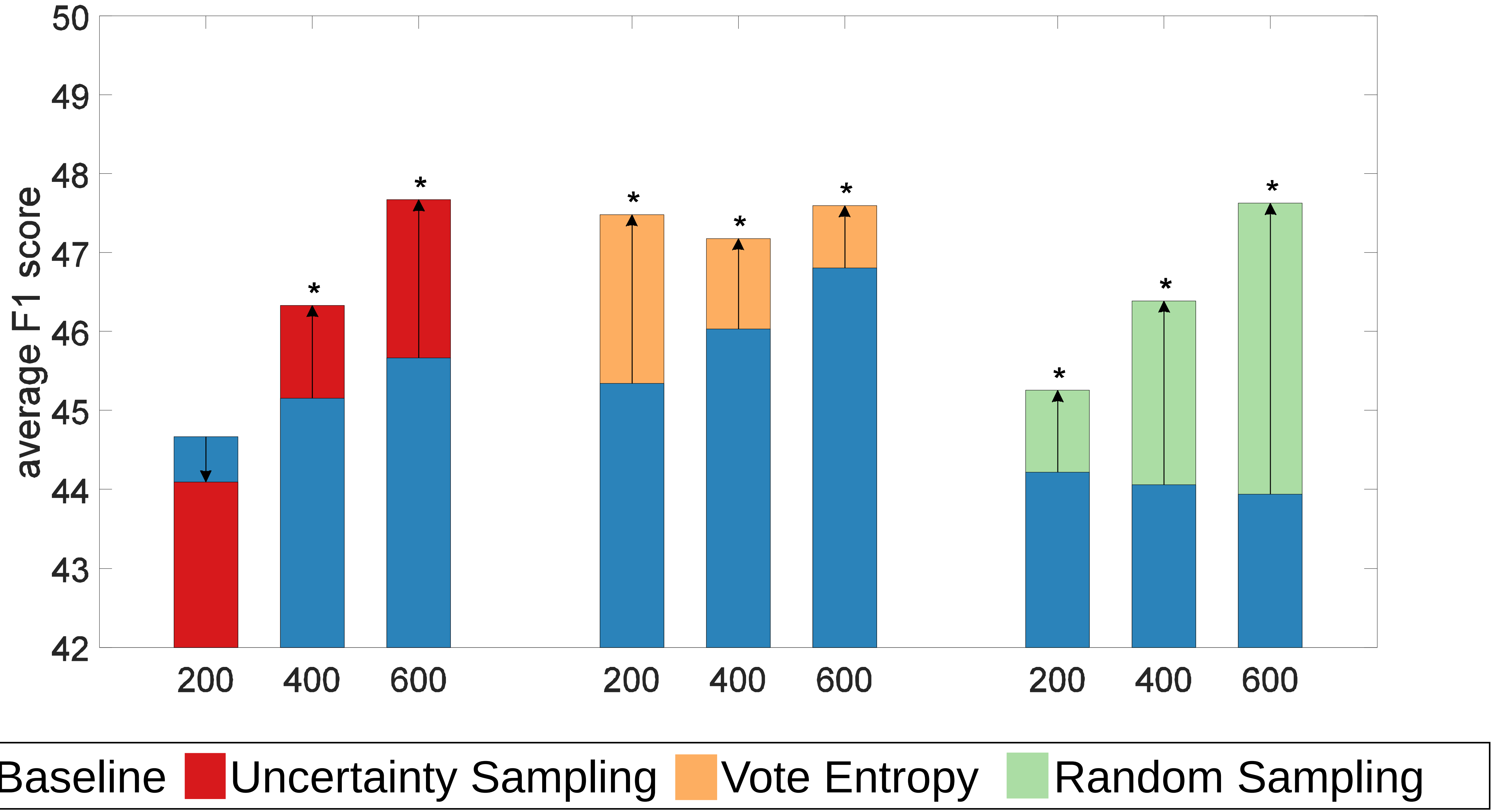
Comparing with the Baseline



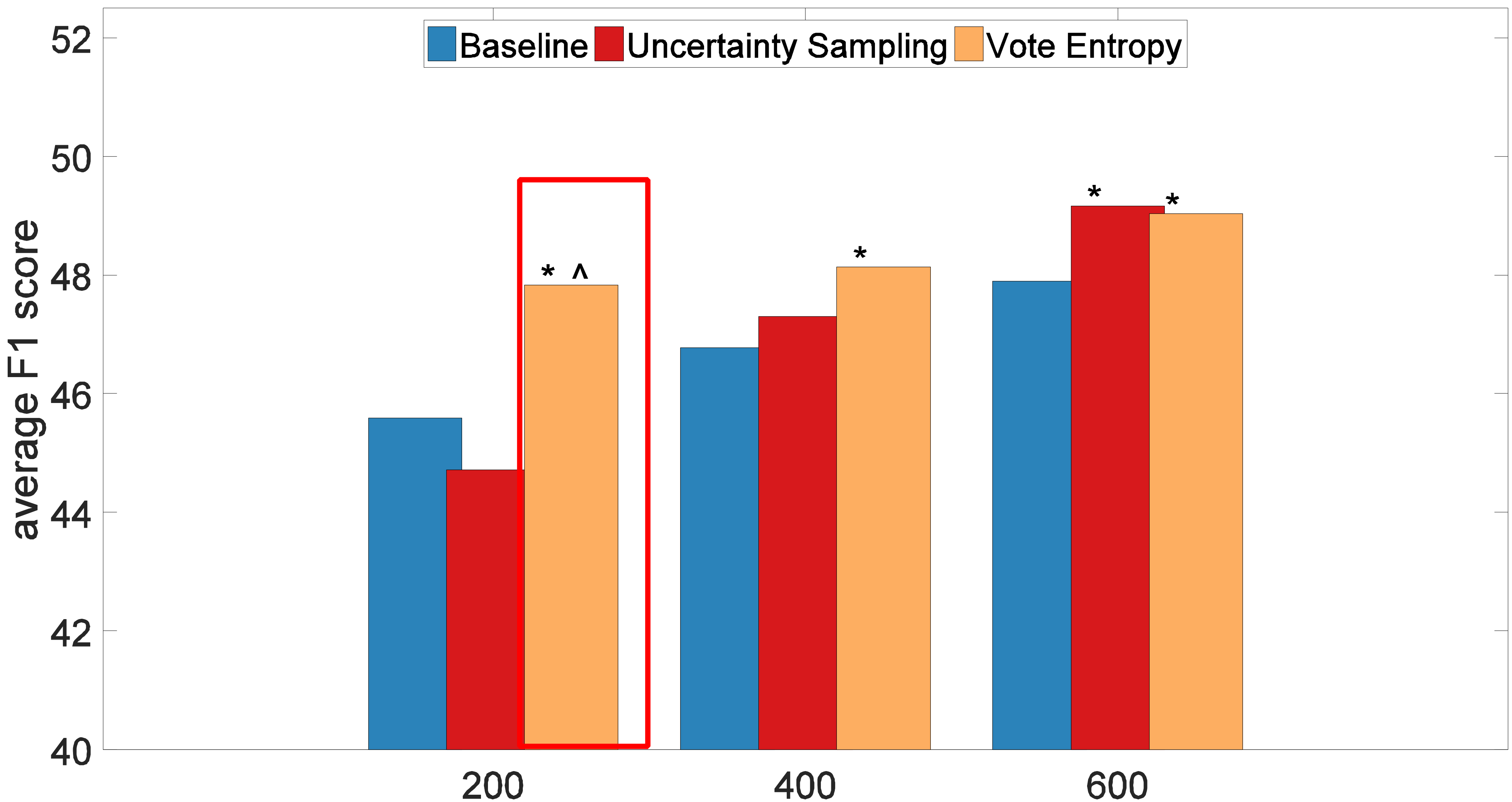


Random Sampling performance gap increases as more data is selected

Comparing with the Baseline

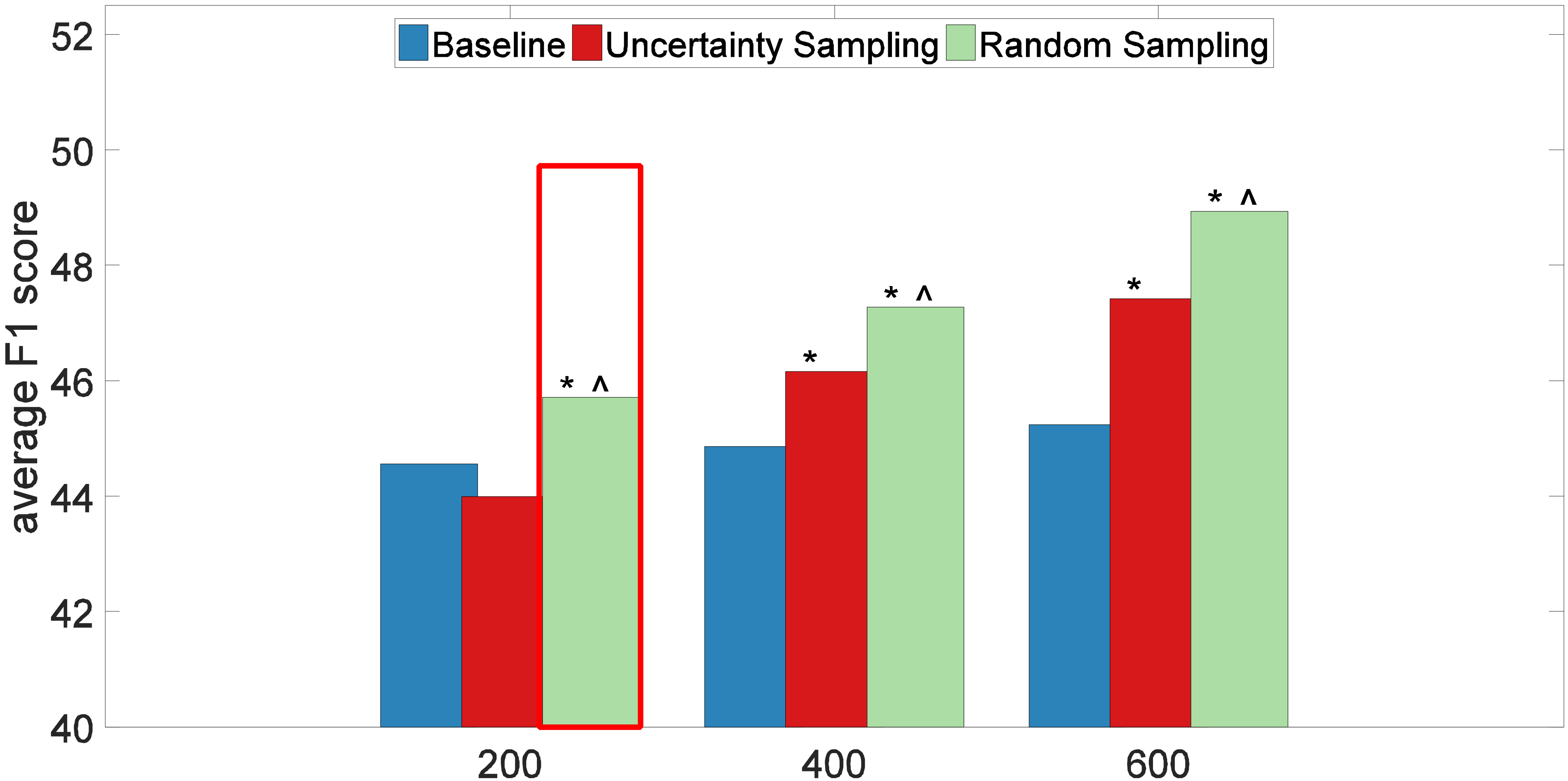


Comparing Data Selection Criteria



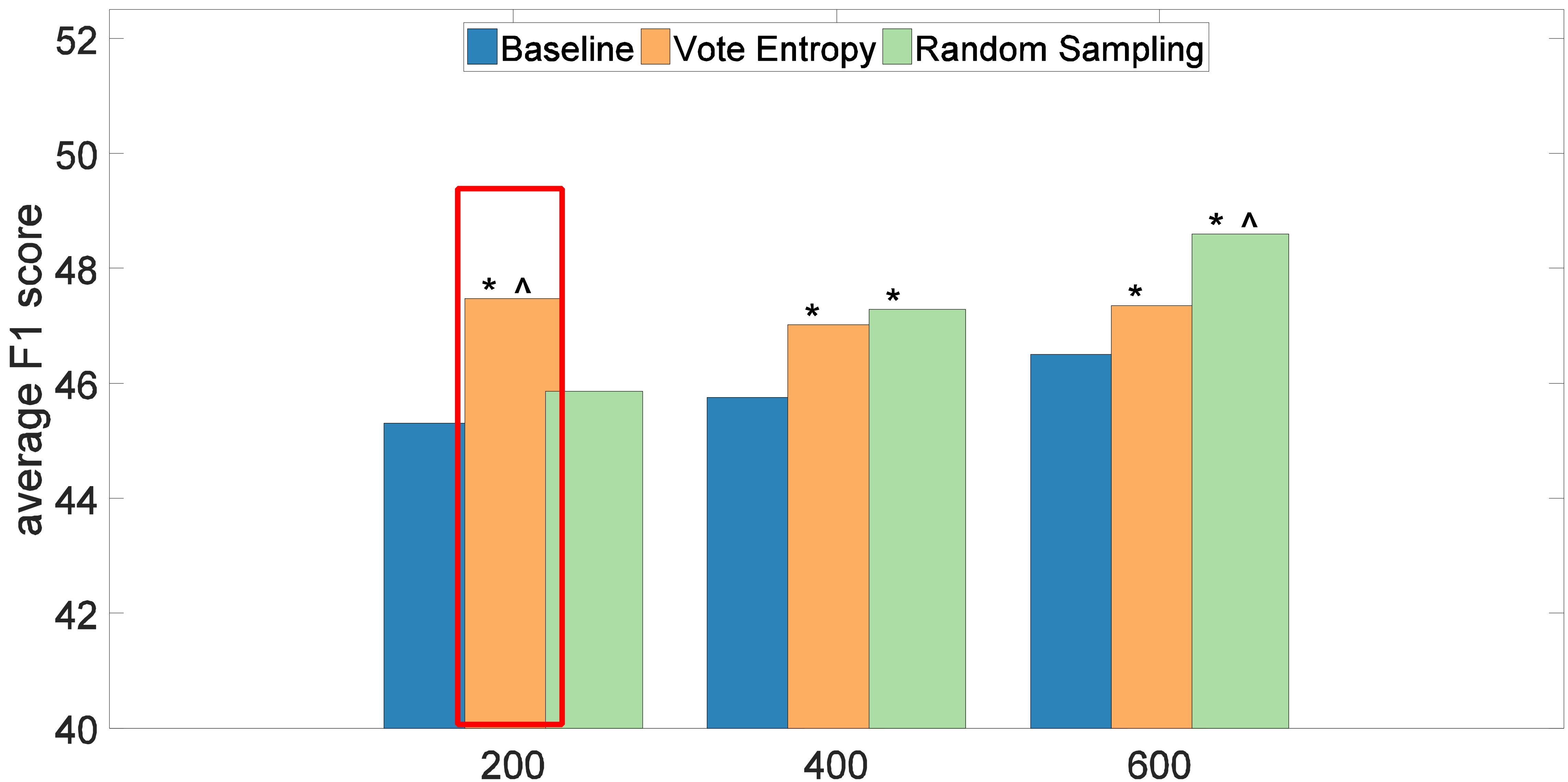
Vote Entropy outperforms for small data size
Uncertainty Sampling catches up as we increase the size

Comparing Data Selection Criteria



Random sampling outperforms in all data sizes

Comparing Data Selection Criteria



Vote Entropy outperforms for small data size
Random Sampling outperforms for larger data size

Conclusions

- Significant improvement by performing feature selection on a small set from the target domain
- Ensuring the ensemble's diversity yields better generalization
- It is important to carefully choose which data to use in the feature selection
  - If you are selecting a small sample Vote Entropy is the best option
  - If the sample size is large Random sampling better represents the target domain

Thanks for your attention!

[1] L. Duan, I.W. Tsang, D. Xu, and T.S. Chua, “Domain adaptation from multiple sources via auxiliary classifiers,” in *International Conference on Machine Learning (ICML 2009)*, Montreal, QC, Canada, June 2009, pp. 289–29.

[2] W. Dai, Q. Yang, G.R. Xue, and Y. Yu, “Boosting for transfer learning,” in *International conference on Machine learning (ICML 2007)*, Corvallis, OR, USA, June 2007, pp. 193–200

[3] J. Gao, W. Fan, J. Jiang, and J. Han, “Knowledge transfer via multiple model local structure mapping,” in *ACM SIGKDD international conference on Knowledge discovery and data mining*, Las Vegas, NV, USA, August 2008, pp. 283–291.

[4] R. Xia, C. Zong, X. Hu, and E. Cambria, “Feature ensemble plus sample selection: Domain adaptation for sentiment classification,” *IEEE Intelligent Systems*, vol. 28, no. 3, pp. 10–18, May-June 2013.

[5] B. Schuller, Z. Zhang, F. Weninger, and G. Rigoll, “Using multiple databases for training in emotion recognition: To unite or to vote?,” in *Interspeech 2011*, Florence, Italy, August 2011, pp. 3285–3288