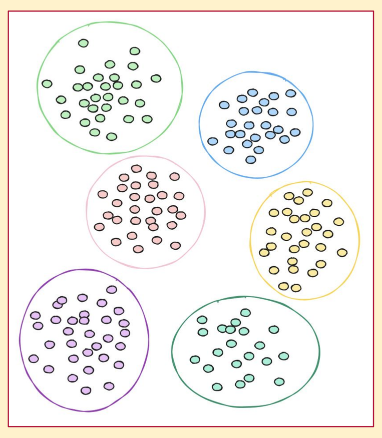
Clustering BERT Embeddings for Classification in SA via Dot Product (CBERTdp)

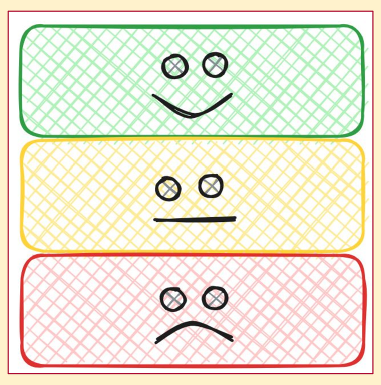
[1)]Thomas Vecchiato, [1)]Riccardo Zuliani,
[2)]Isabel Marie Ritter, [2)]Alice Schirrmeister

[1)]Ca' Foscari University of Venice
[2)]Osnabrück University
{880038, 875532, 1000371, 1000095}@stud.unive.it

Goals



Neural Networks are very costly, so we investigate the use of K-Means clustering in combination with the dot product to simplify classification tasks.



We use Sentiment Analysis as exemplary task and cluster BERT [1] embeddings. The dot product of a new sentence's embedding and the cluster centroids determines the corresponding class label.

Related Works

Previous works:
- Clustering BERT embedding is not a new idea [2]
- Many researches focusing on the topic modeling aspects and prototype selection

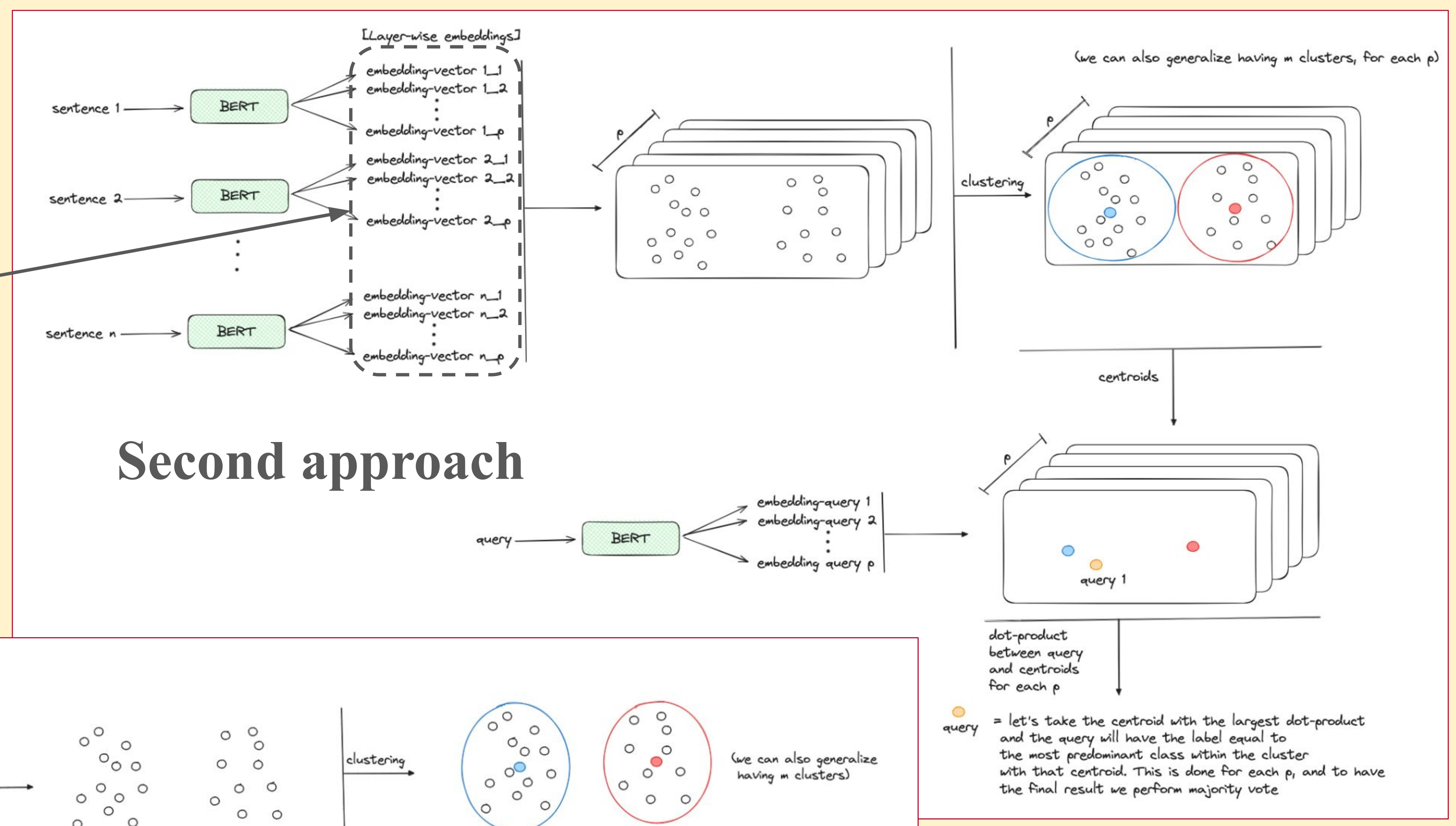
Baseline:
- *Non ML:* Naive baseline, Random choice
- *ML:* SVM, Naive Bayes, Logistic Regression, via TF-IDF word-embedding

Competitors:
- Naive BERT
- BERT + (GRU, LSTM, Bi-LSTM)

What are the fundamental pillars:
- Saving computational time by not considering an additional NN for classification
- Memory saving
- Retaining the same performance or improving it with respect to the SOTA competitors

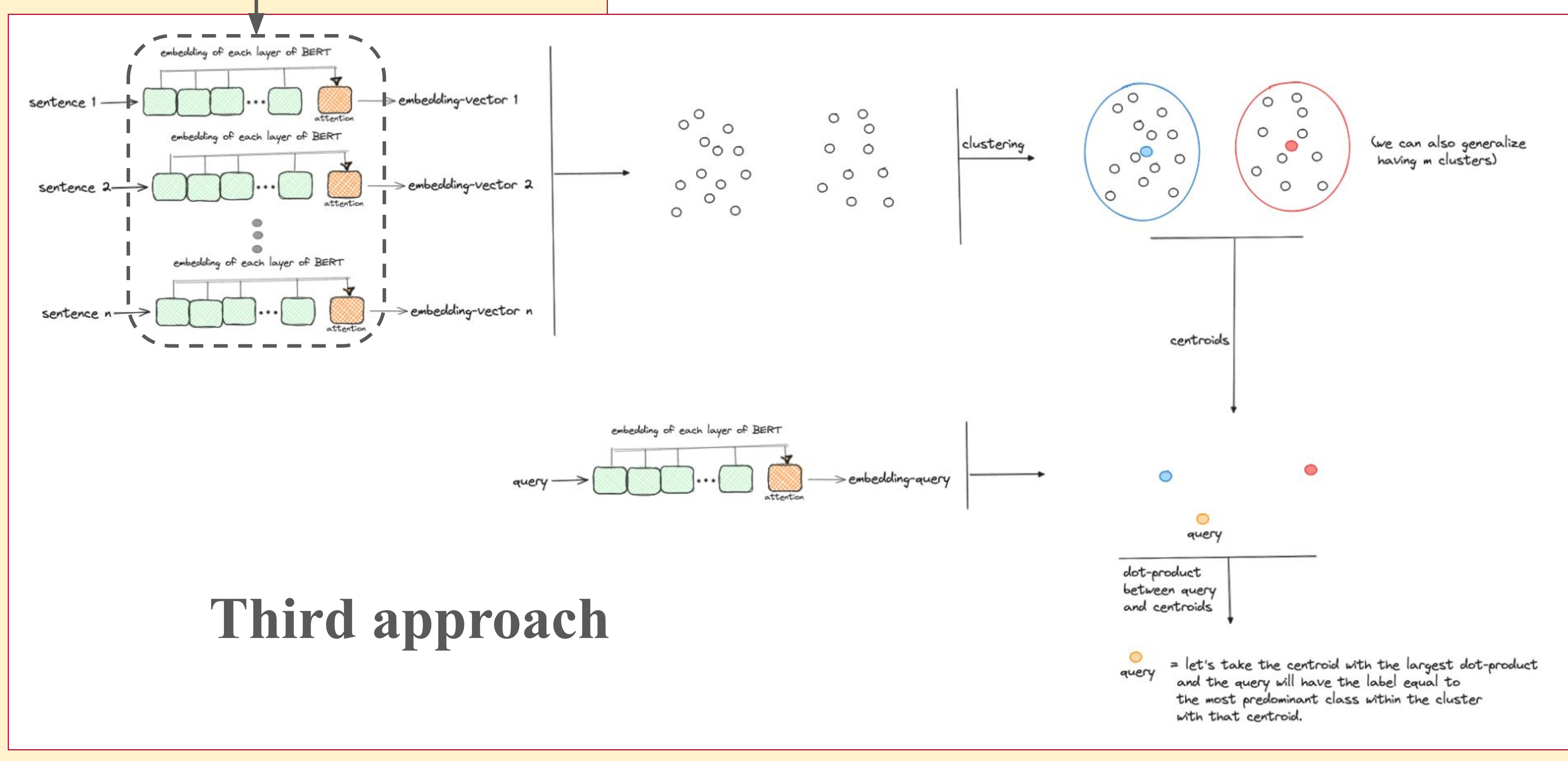
Architectures Pipeline

Layer-wise Embeddings



Second approach

Layer Aggregation

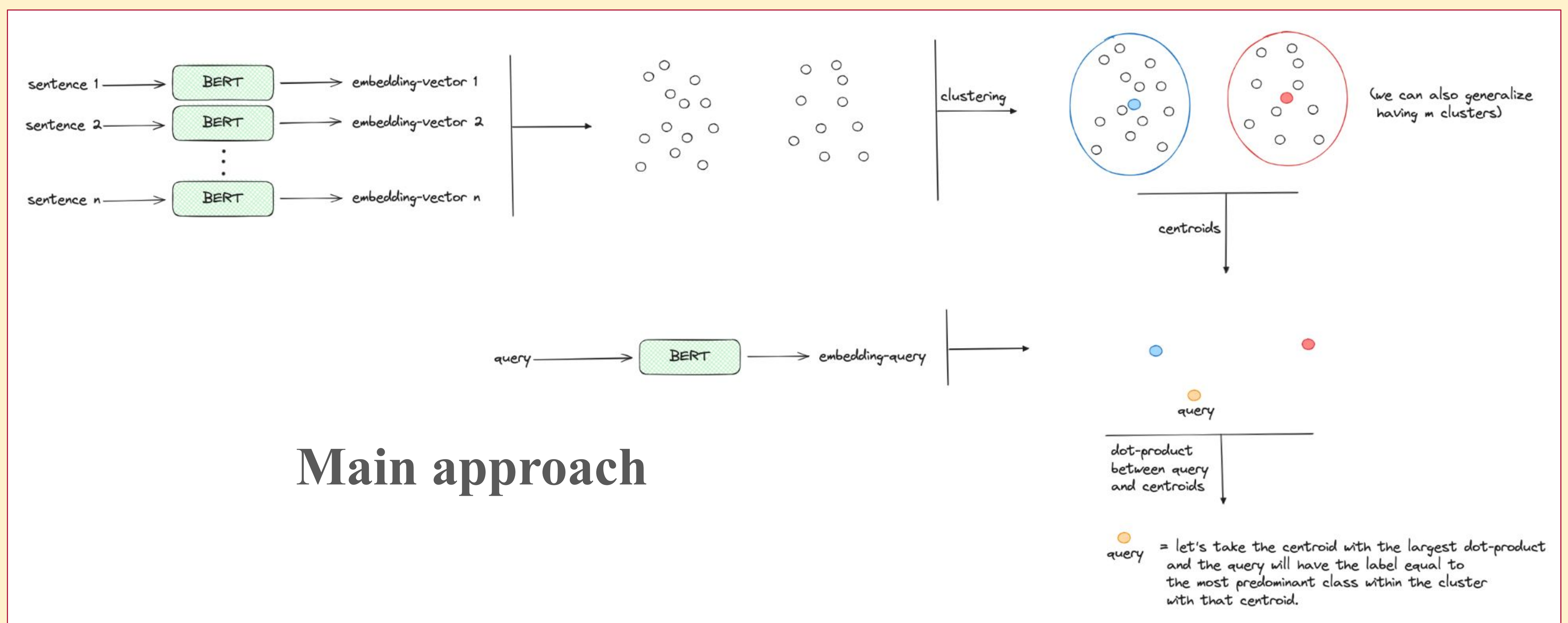


Third approach

The Proposed Method

We propose a novel method to perform sentiment analysis and not only, where the main ingredients are: a **large language model** (BERT), a **clustering algorithm** and a **simple dot-product**.

Following our methodology we encode each sentence into a vector using an LLM, run a clustering algorithm over the vector space, get the centroids of the clusters identified and assign to each centroid a label that can be **positive** or **negative**. In this way it is only necessary to save the **centroids** and the related labels. When a new sentence or query is given, the classification, a simple **dot product between the centroids and the vector sentence** given by an LLM, is performed. The final label for the query will be given according to the **most similar centroids and the prevailing sentiment in that cluster.**



Main approach

Experiments

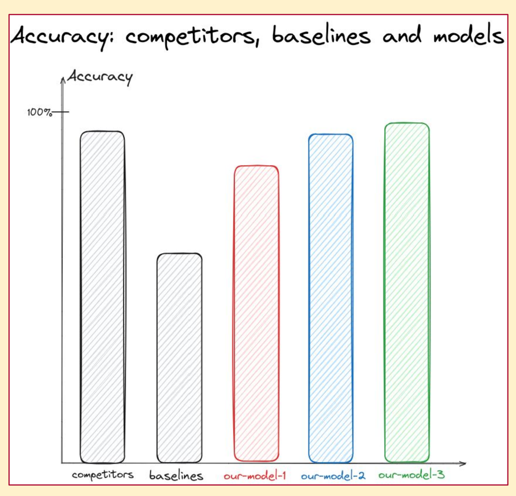
We evaluate our methods alongside the test set of 3 different **datasets**: *IMDb [3], Stanford Sentiment Treebank [4]* and *Yelp Review Dataset [5].*
Compare our methods with and without dimensionality reduction.

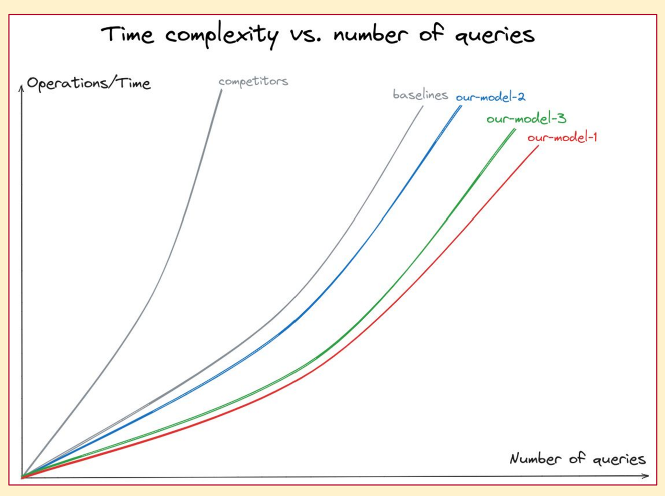
Measures:
- *Cluster Goodness/Purity Assessment:* Confidence measurement [6]
- *Performance Metrics:* Accuracy Score, F1-Score, Precision, Recall
- *Model effectiveness:* Compare with Baseline-accuracy
- *Complexity:* Comparison of computational complexity/costs

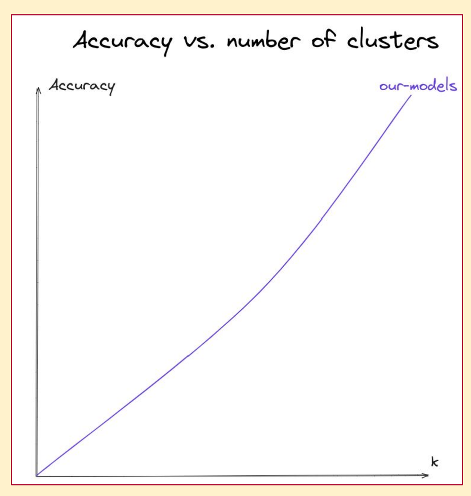
Images from: https://www.shutterstock.com/de/image-vector/key-performance-indicator-concept-icons-efficiency-2167691037

Results

Our results are focused on two key aspects: the **accuracy of the models** and their **computational complexity**. Concerning the accuracy, we expect to see a behavior that is **almost equal to the state of the art approaches** present in the literature. Instead, as regards the **time complexity, we expect to see that our approaches outperform all existing competitors.** Our resulting models should represent a reasonable compromise in terms of accuracy and complexity. In addition, an important parameter to take into account is **K** (the number of clusters), whose value can change the final results.







References

1. BERT: Pre-training of deep bidirectional transformers for language understanding. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019.
2. Classification and clustering of arguments with contextualized word embeddings. Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019.
3. IMDd - Dataset at Hugging Face.
4. Stanford Sentiment Treebank - Dataset at Hugging Face.
5. Yelp Review Dataset - Dataset at Hugging Face.
6. Using dominant sets for K-NN prototype selection. Sebastiano Vascon, Marco Cristani, Marcello Pelillo, Vittorio Murino. 2013.

Takeaways

Our proposed approaches offer a **compelling solution for real-world sentiment analysis tasks.**

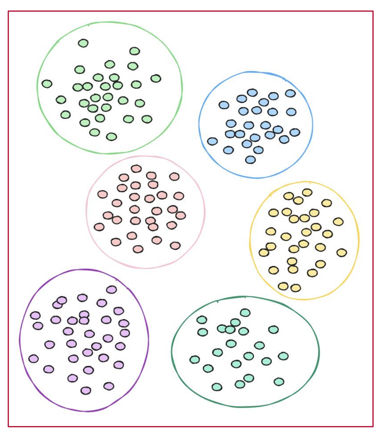
By combining BERT embeddings and K-Means clustering we achieve **competitive accuracy while minimizing computational complexity.**

Clustering BERT Embeddings for Classification in SA via Dot Product (CBERTdp)

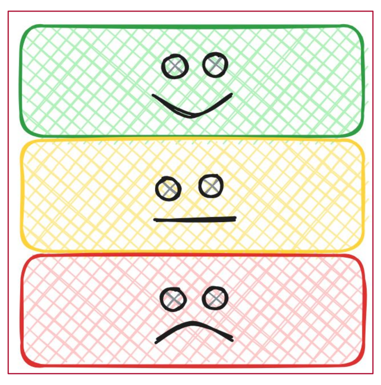
[1)]Thomas Vecchiato, [1)]Riccardo Zuliani,
[2)]Isabel Marie Ritter, [2)]Alice Schirrmeister

[1)]Ca' Foscari University of Venice
[2)]Osnabrück University
{880038, 875532, 1000371, 1000095}@stud.unive.it

Goals



Neural Networks are very costly, so we investigate the use of K-Means clustering in combination with the dot product to simplify classification tasks.



We use Sentiment Analysis as exemplary task and cluster BERT [1] embeddings. The dot product of a new sentence's embedding and the cluster centroids determines the corresponding class label.

Related Works

Previous works:

- Clustering BERT embedding is not a new idea [2]
- Many researches focusing on the topic modeling aspects and prototype selection

Baseline:

- *Non ML:* Naive baseline, Random choice
- *ML:* SVM, Naive Bayes, Logistic Regression, via TF-IDF word-embedding

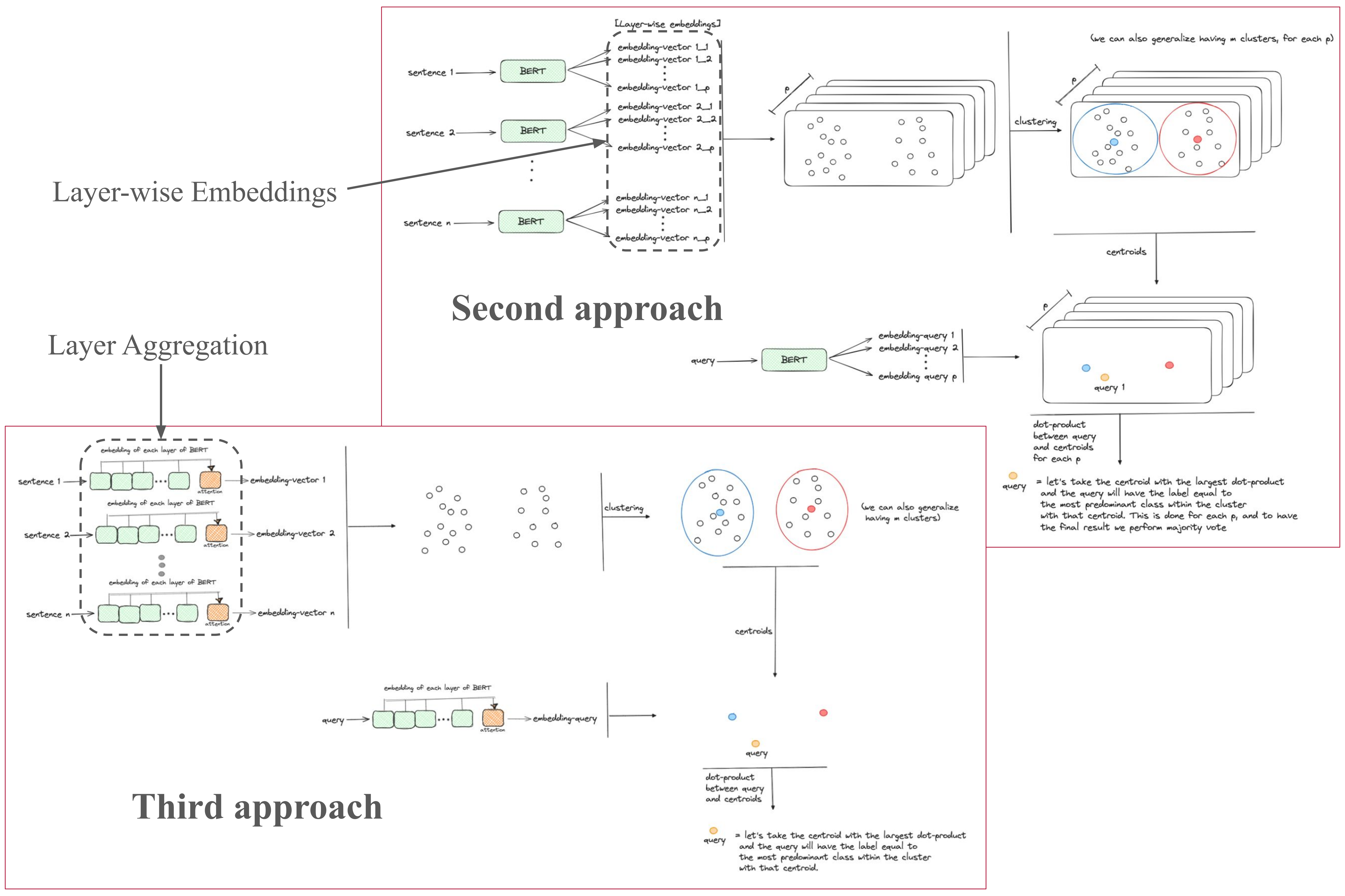
Competitors:

- Naive BERT
- BERT + (GRU, LSTM, Bi-LSTM)

What are the fundamental pillars:

- Saving computational time by not considering an additional NN for classification
- Memory saving
- Retaining the same performance or improving it with respect to the SOTA competitors

Architectures Pipeline



Experiments

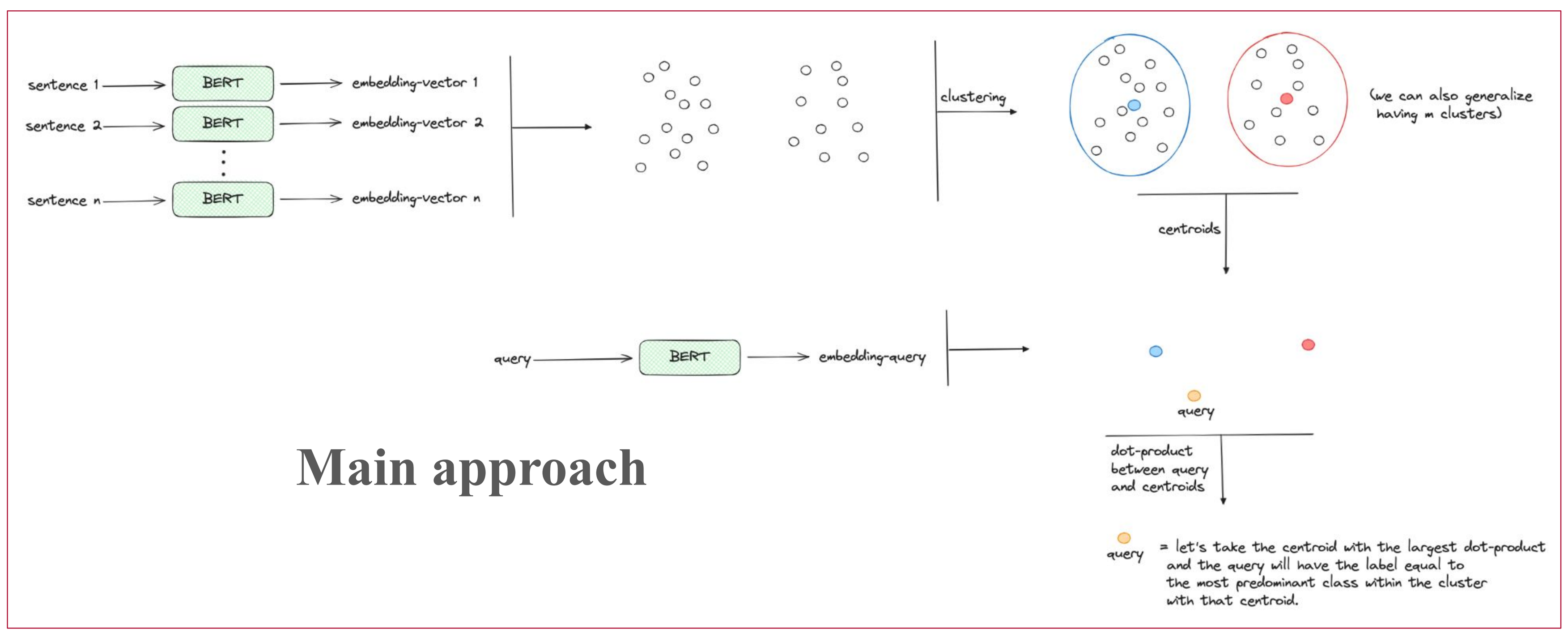
We evaluate our methods alongside the test set of 3 different **datasets**: *IMDb [3], Stanford Sentiment Treebank [4]* and *Yelp Review Dataset [5].*
Compare our methods with and without dimensionality reduction.
Measures:

- *Cluster Goodness/Purity Assessment:* Confidence measurement [6]
- *Performance Metrics:* Accuracy Score, F1-Score, Precision, Recall
- *Model effectiveness:* Compare with Baseline-accuracy
- *Complexity:* Comparison of computational complexity/costs

Images from: https://www.shutterstock.com/de/image-vector/key-performance-indicator-concept-icons-efficiency-2167691037

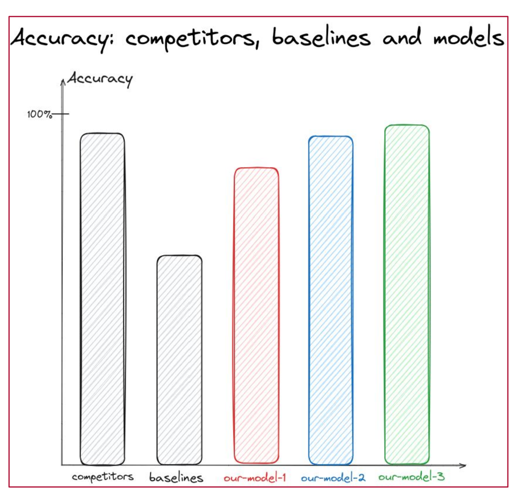
The Proposed Method

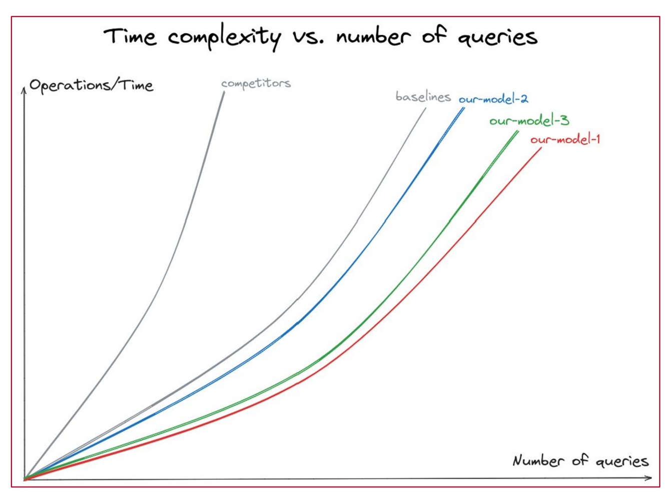
We propose a novel method to perform sentiment analysis and not only, where the main ingredients are: a **large language model** (BERT), a **clustering algorithm** and a **simple dot-product**.
Following our methodology we encode each sentence into a vector using an LLM, run a clustering algorithm over the vector space, get the centroids of the clusters identified and assign to each centroid a label that can be **positive** or **negative**. In this way it is only necessary to save the **centroids** and the related labels. When a new sentence or query is given, the classification, a simple **dot product between the centroids and the vector sentence** given by an LLM, is performed. The final label for the query will be given according to the **most similar centroids and the prevailing sentiment in that cluster.**

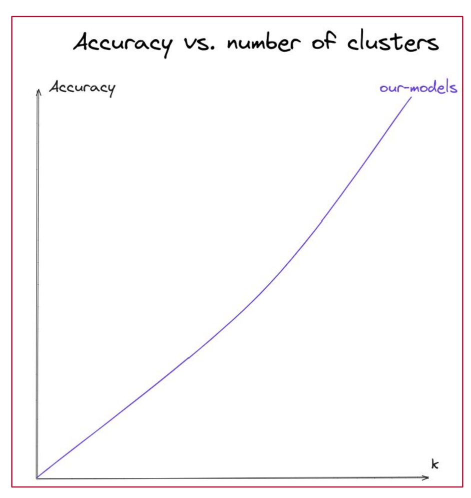


Results

Our results are focused on two key aspects: the **accuracy of the models** and their **computational complexity**. Concerning the accuracy, we expect to see a behavior that is **almost equal to the state of the art approaches** present in the literature. Instead, as regards the **time complexity, we expect to see that our approaches outperform all existing competitors.** Our resulting models should represent a reasonable compromise in terms of accuracy and complexity. In addition, an important parameter to take into account is **K** (the number of clusters), whose value can change the final results.







References

1. BERT: Pre-training of deep bidirectional transformers for language understanding. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019.
2. Classification and clustering of arguments with contextualized word embeddings. Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019.
3. IMDd - Dataset at Hugging Face.
4. Stanford Sentiment Treebank - Dataset at Hugging Face.
5. Yelp Review Dataset - Dataset at Hugging Face.
6. Using dominant sets for K-NN prototype selection. Sebastiano Vascon, Marco Cristani, Marcello Pelillo, Vittorio Murino. 2013.

Takeaways

Our proposed approaches offer a **compelling solution for real-world sentiment analysis tasks.**

By combining BERT embeddings and K-Means clustering we achieve **competitive accuracy while minimizing computational complexity.**