**Классификация эмоций в речевом сигнале с использованием алгоритма k ближайших соседей, визуализация распределения данных.**

1. Использованные датасеты:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name | Basic emotion categories | N samples | Language | Acted/Improvized |
| CREMA-D[1] | ANG, HAP, SAD, NEU, DIS, FEA | 7442 | Eng | Acted |
| IEMOCAP[2] | ANG, HAP, SAD, NEU, EXC, FRU | 7304 | Eng | Both |
| Emo-DB[3] | ANG, HAP, SAD, NEU, DIS, FEA, BOR | 535 | Deu | Acted |
| RAVDESS[4] | ANG, HAP, SAD, NEU, DIS, FEA, SUR, CAL | 1440 | Eng | Acted |
| SAVEE[5] | ANG, HAP, SAD, NEU, DIS, FEA, SUR | 480 | Eng | Acted |
| TESS[6] | ANG, HAP, SAD, NEU, DIS, FEA, SUR | 2798 | Eng | Acted |

Таблица 1. Наборы данных, использованные в исследовании. Здесь и далее: ANG - Anger, Hap - Happiness, SAD - Sadness, NEU - Neutral, DIS - Disgust, FEA - Fear, EXC - Excitement, FRU - Frustration, BOR - Boredom, SUR - Surprise, CAL - Calm

Также был использован набор данных, содержащий 15781 образец из шести базовых эмоциональных категорий: ANG, HAP, SAD, NEU, DIS, FEA. Далее этот датасет указан как *English-Assembly-Six*. Все наборы данных расположены в директории root/datasets/

1. Подготовка датасетов:
   1. Извлечение признаков:

С помощью *OpenSMILE*[7] извлечен набор низкоуровневых признаков *Extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS)*[8]. Результат – 88-размерный признаковый вектор для каждого .wav файла. Итоговый набор признаков и разметки для каждого датасета представлен таблицей 88\*N, где N – количество файлов. Таблицы сохранены в файлы features.csv, расположенные в директории root/datasets/<dataset-name>/features/

* 1. Разметка .csv файлов:

Для каждого датасета в файл features.csv добавлена колонка ‘Label’, заполненная путём парсинга строки с названием соответствующего файла. Получившиеся таблицы сохранены в файл features\_with\_labels.csv, расположенные в директории root/datasets/<dataset-name>/features/

1. Классификация методом k ближайших соседей (k-NN). Результаты.

Проводилась в 3-х вариантах: на необработанных, нормализованных и стандартизованных данных. Использован *sklearn.neighbors.KNeighborsClassifier* с варьирующимся от 1 до 100 значением k количества ближайших соседей. Данные были предварительно разбиты на тренировочный и тестовый наборы в пропорции 70:30, random\_state=15, с сохранением распределения классов по тренировочному и тестовому наборам. В качестве метрики выбрана точность предсказания на тестовом наборе данных. Результаты представлены на рисунках 1-7.

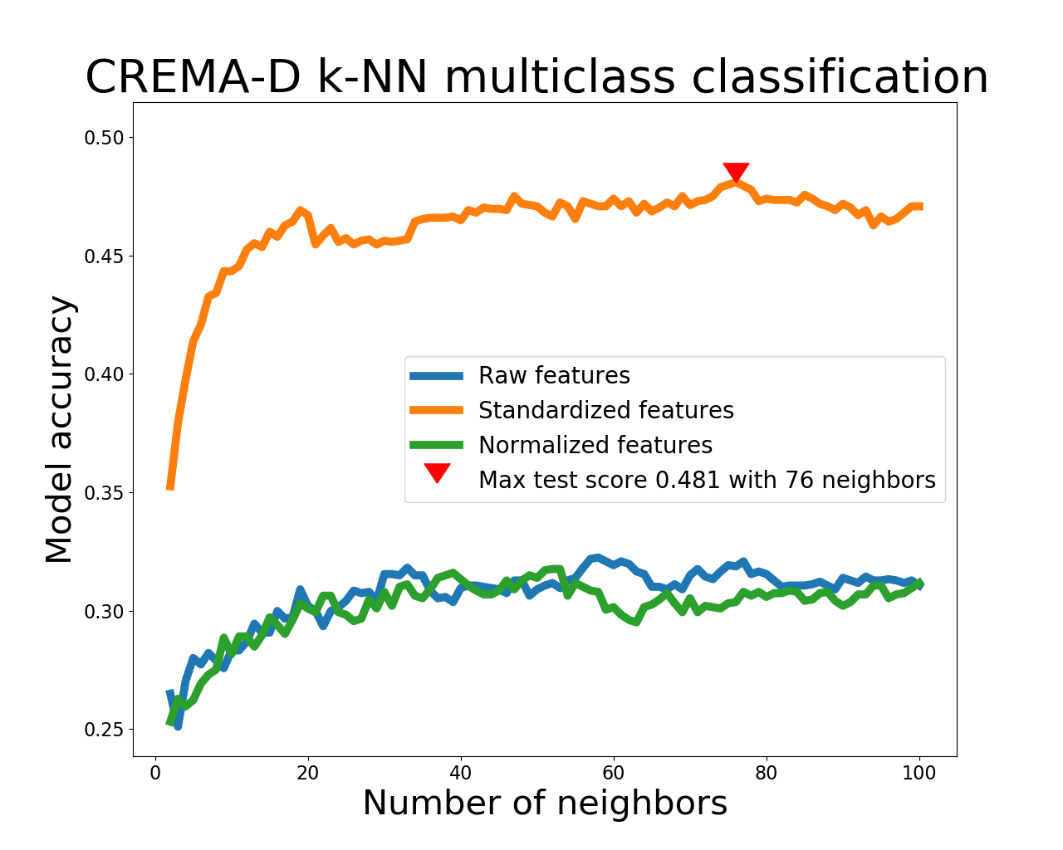


Рисунок 1

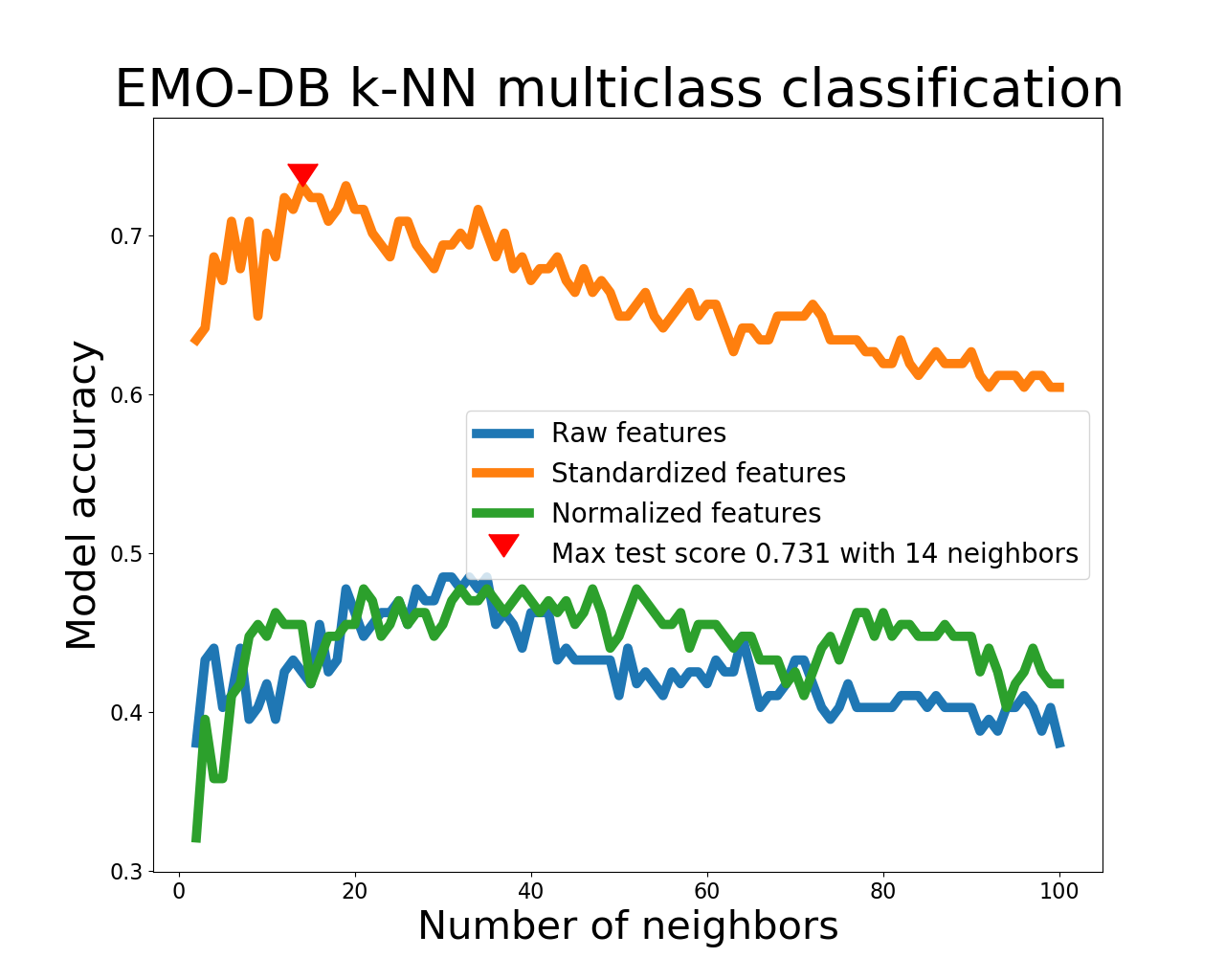


Рисунок 2

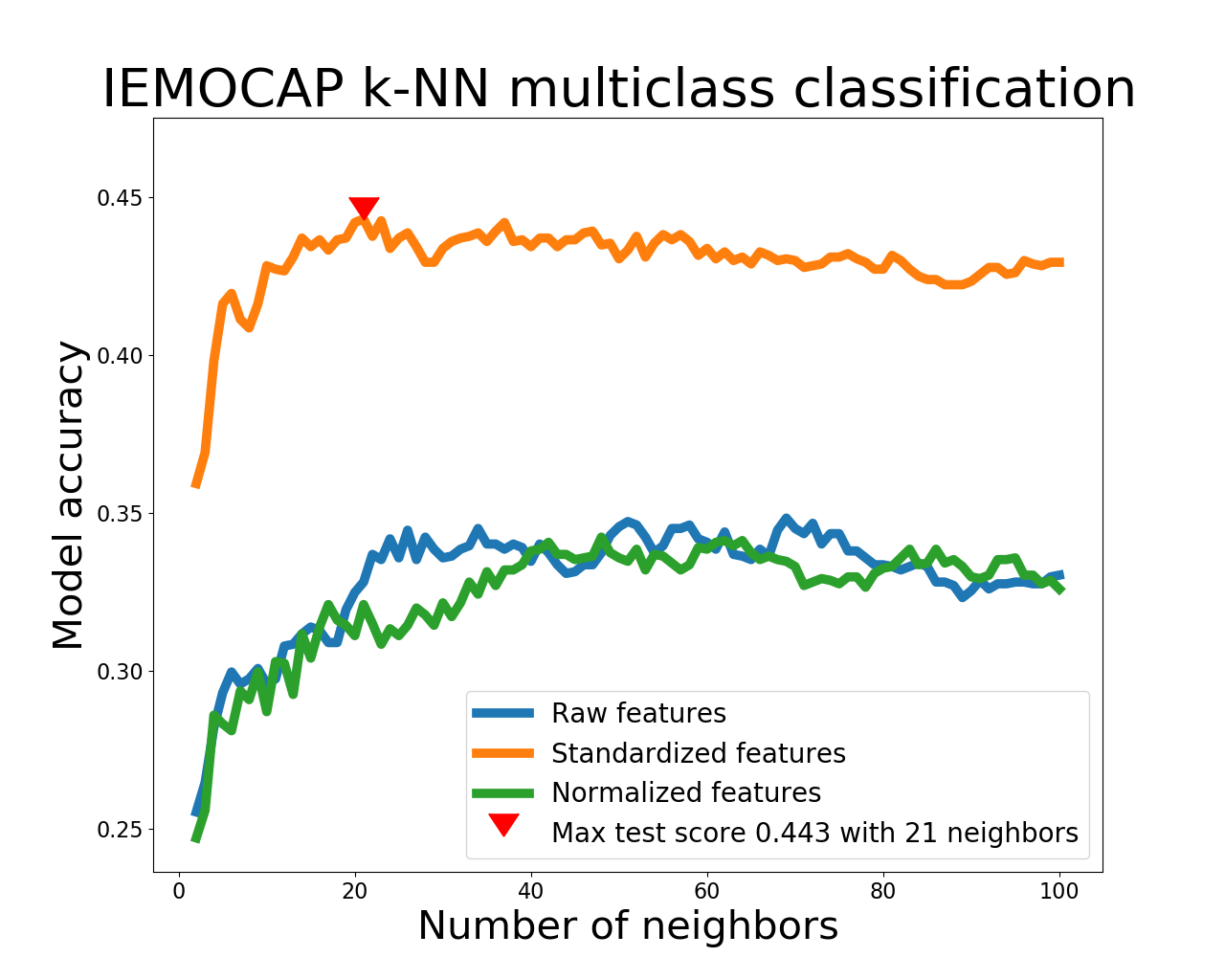


Рисунок 3

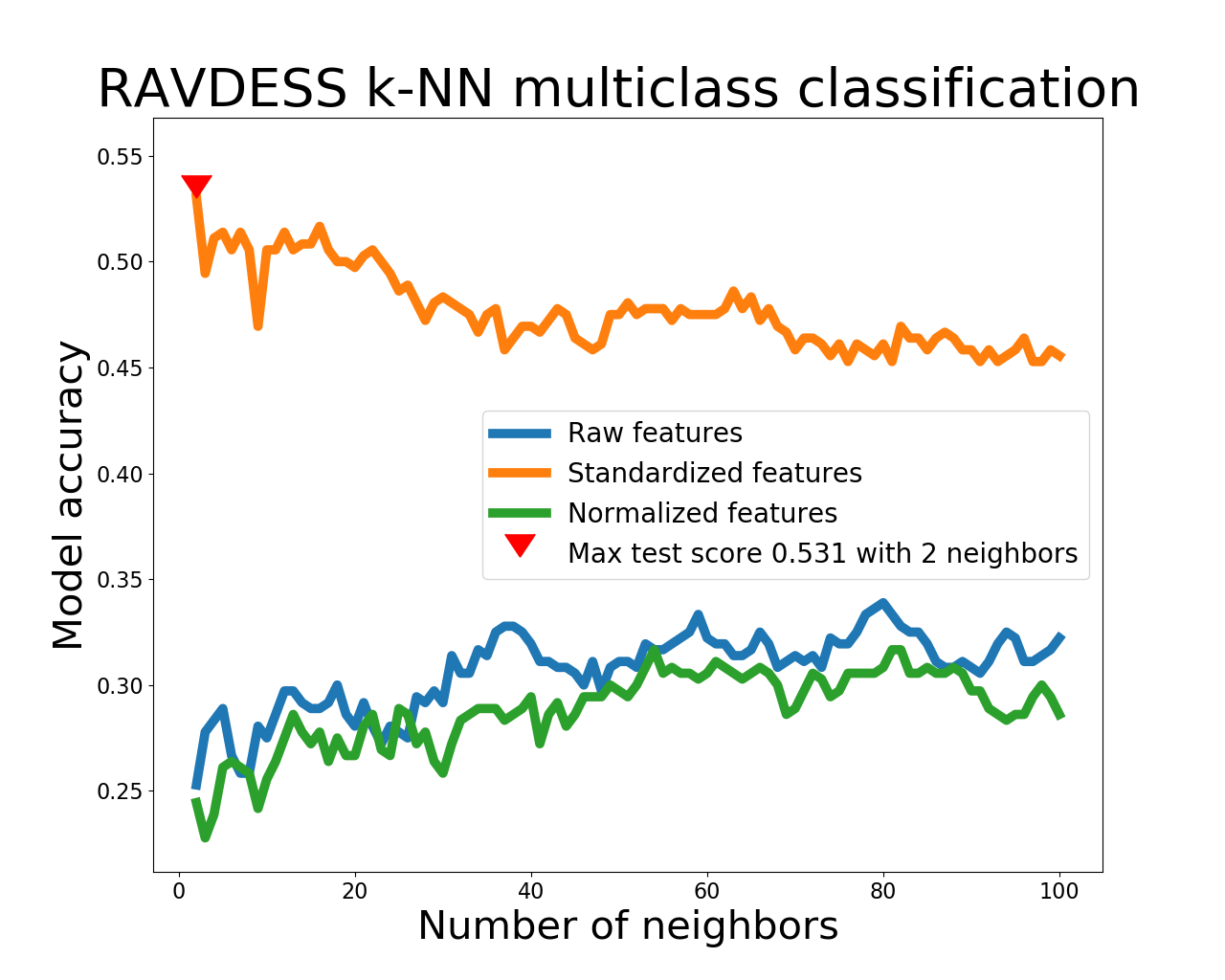


Рисунок 4

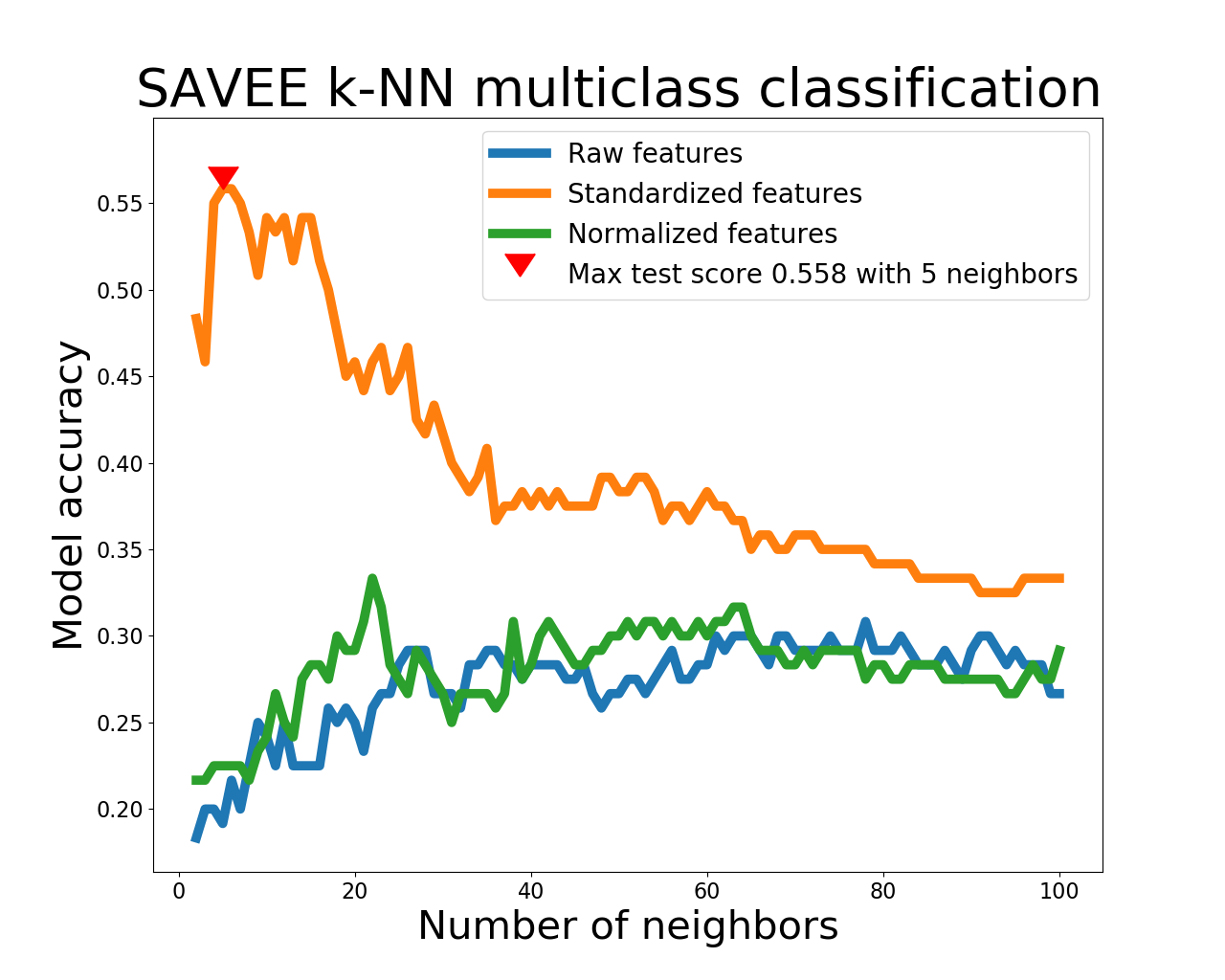
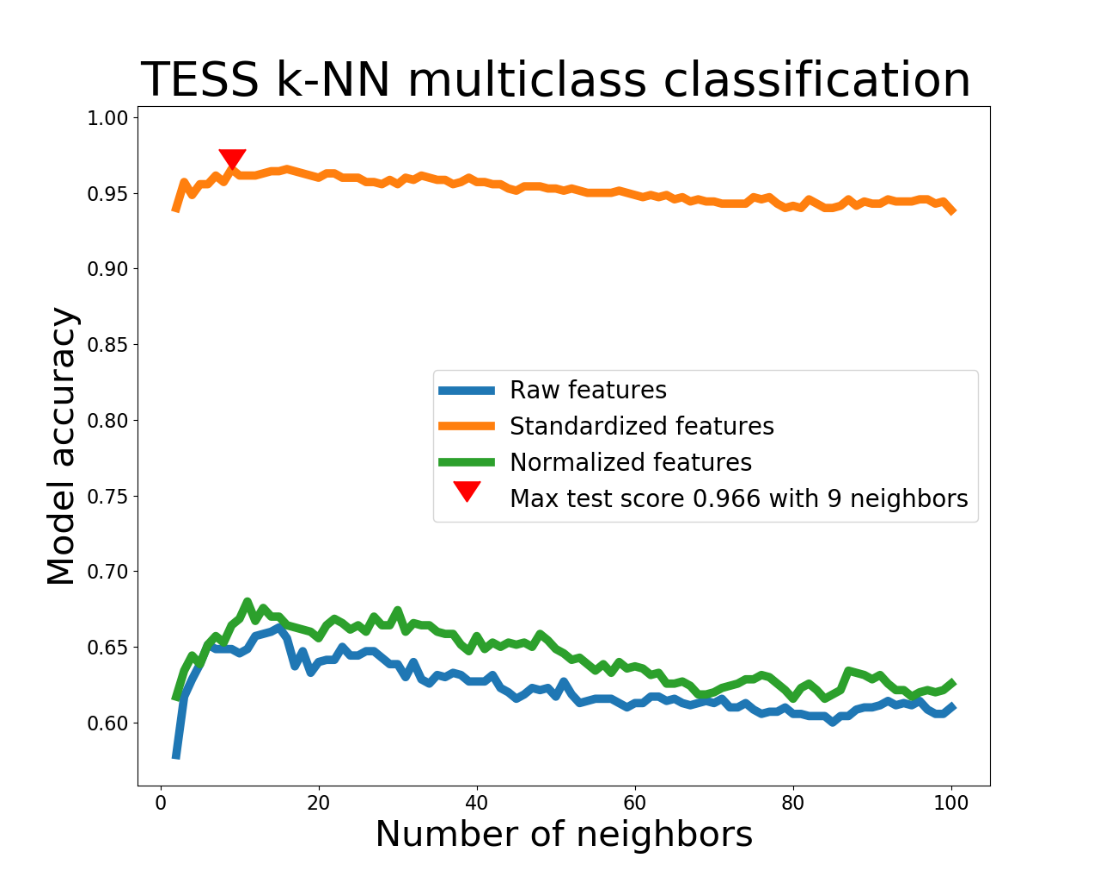
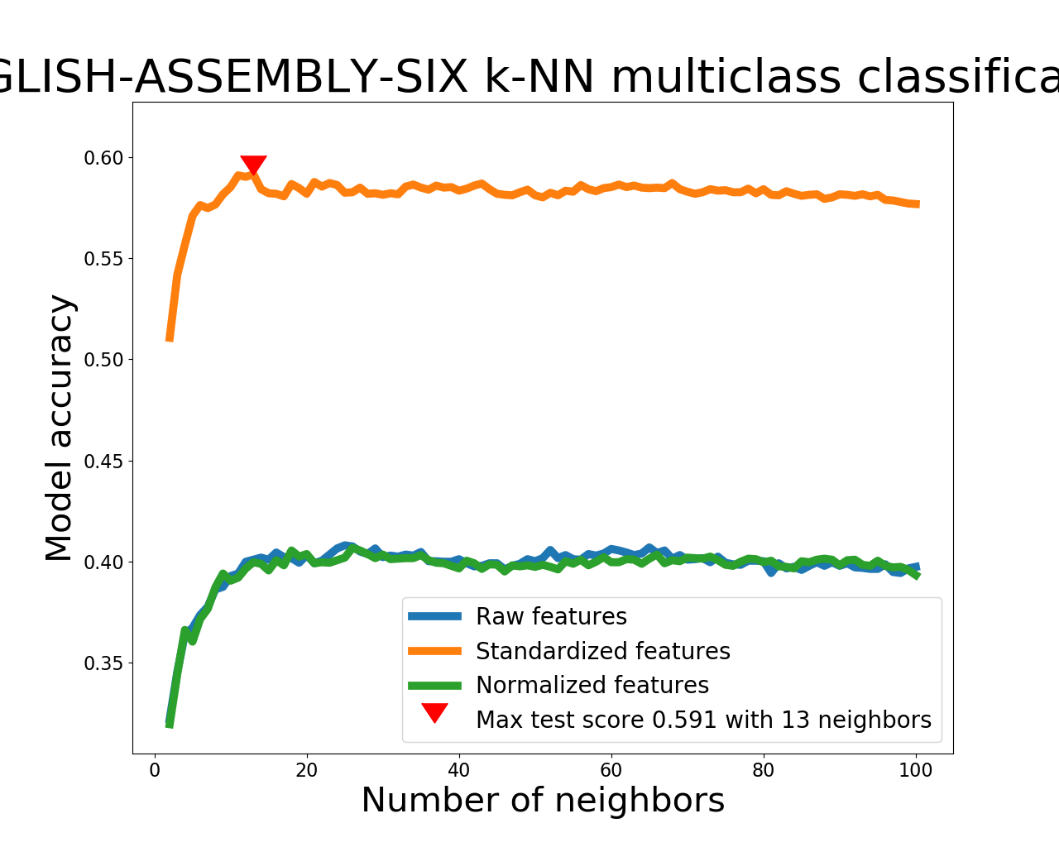


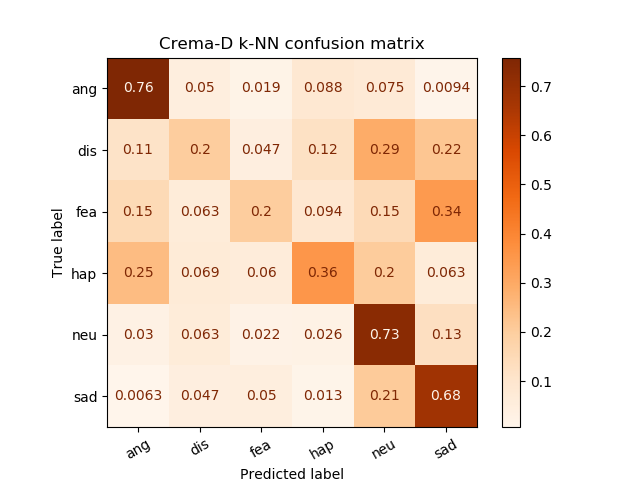
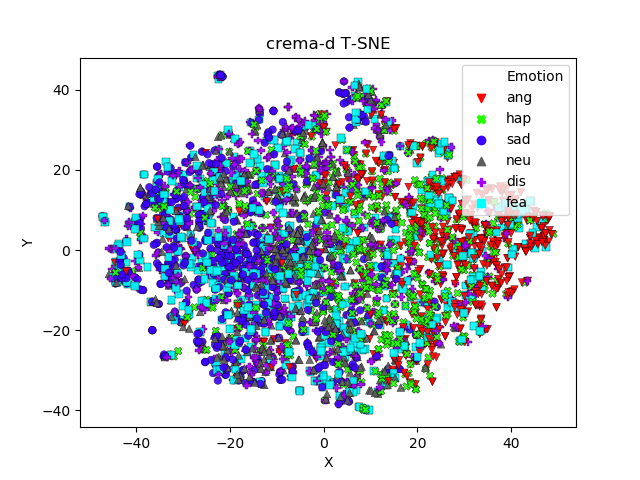
Рисунок 5

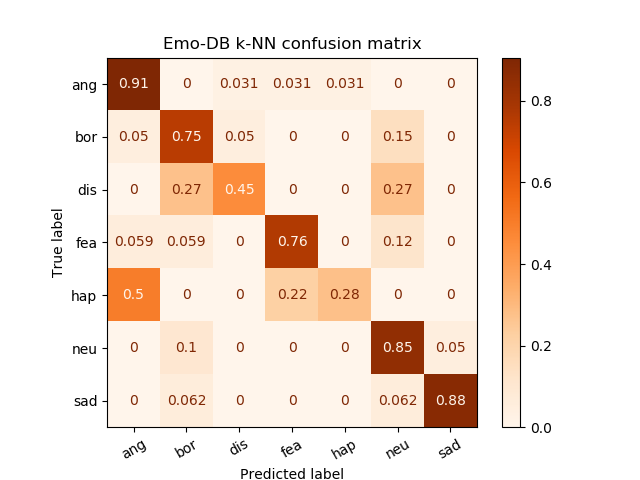
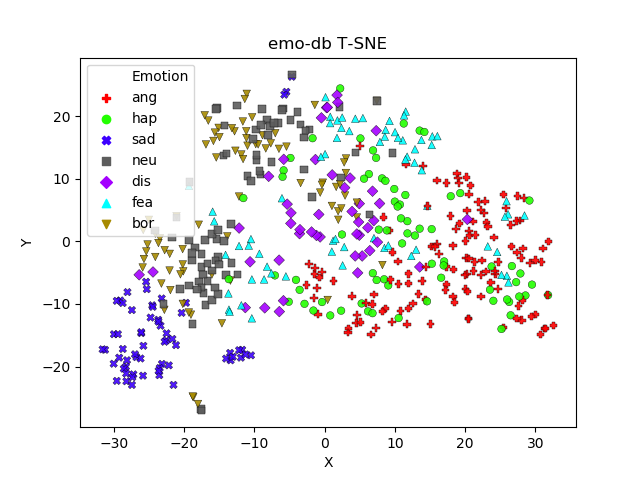


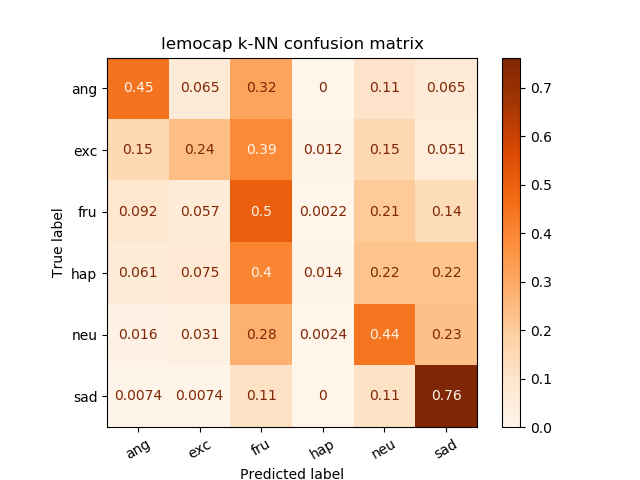
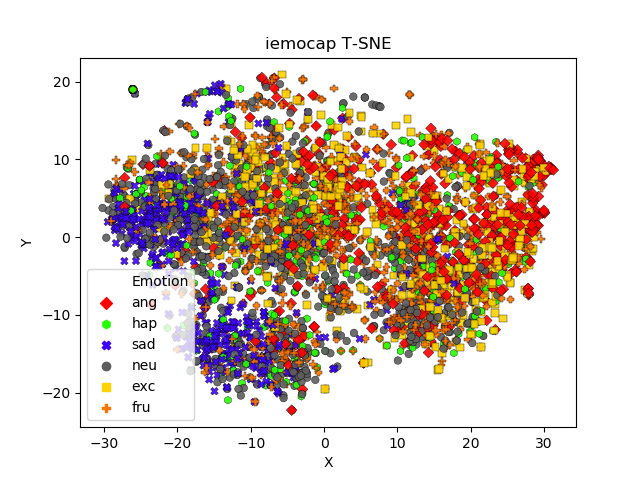
****Рисунок 6

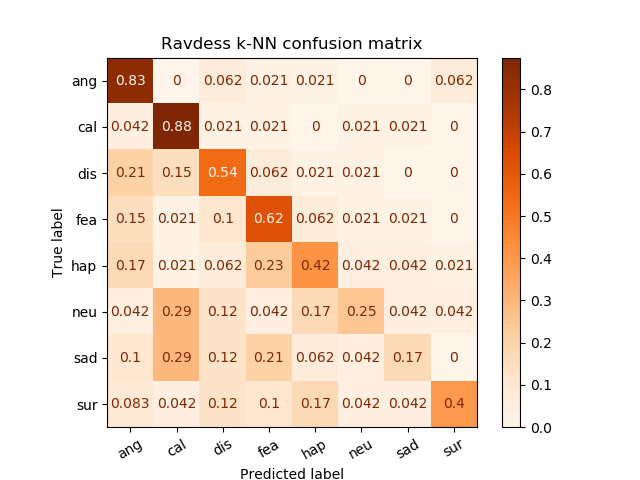
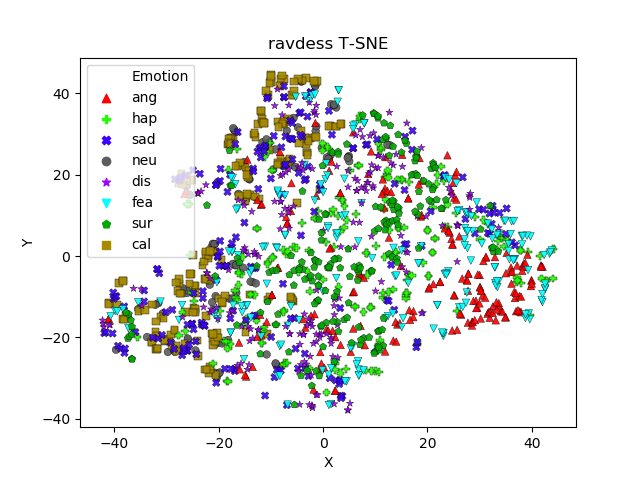
Видим, что стандартизация внесла значительные улучшения в качество классификации, тогда как нормализация лишь незначительно его улучшила, и не во всех случаях. Данные графики сохранены в директории root/plots/knn/evaluation

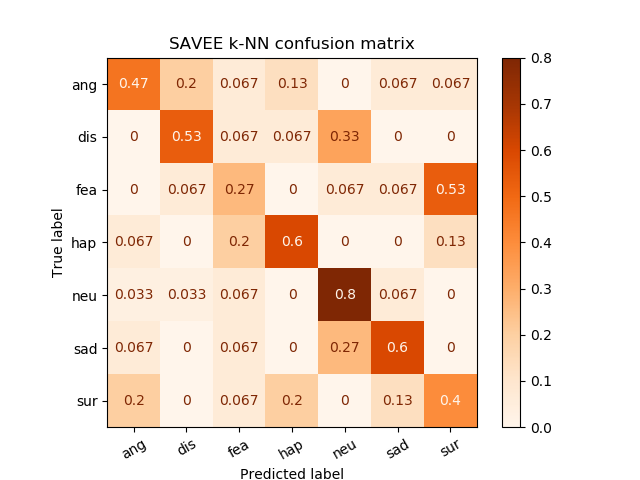
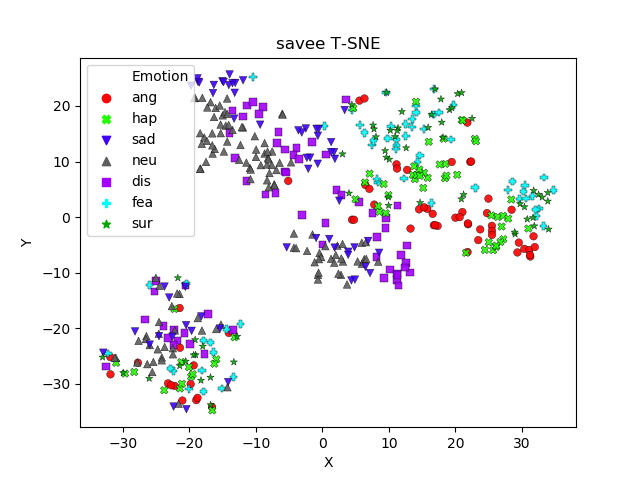
1. Сопоставление результатов классификации с визуализацией распределения данных в признаковом пространстве.

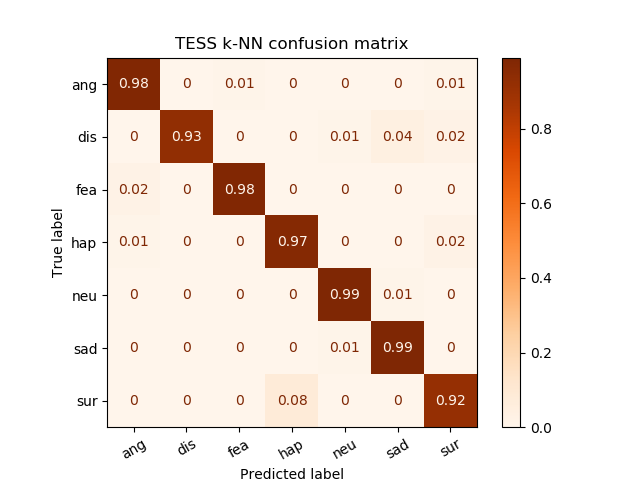
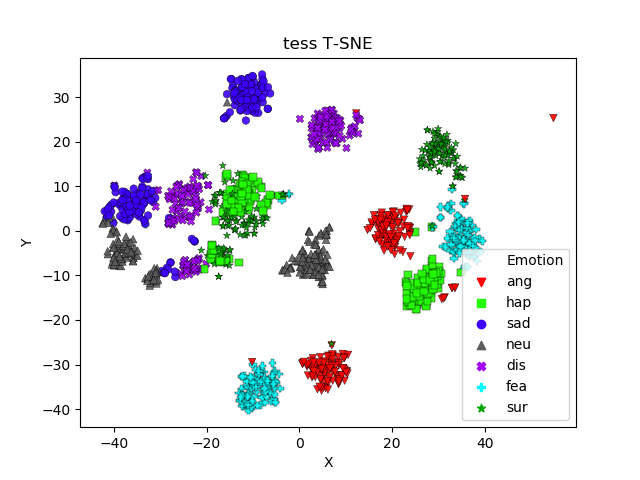
Чтобы получить представление о распределении данных в признаковом пространстве, использован алгоритм уменьшения размерности T-SNE[9]. Он позволяет получить визуализацию, по которой можно делать выводы о возможности и целесообразности классификации представленных данных. К каждому набору предварительно стандартизованных данных был применен *sklearn.manifold.TSNE, n\_iter=10000, learning\_rate=0.1, early\_exaggeration=2*, и варьирующимся от 5 до 75 параметром *perplexity[10]*, таким образом было получено 70 визуализаций для каждого набора данных. Ниже представлены самые удачные. Также, чтобы дополнить картину, графики распределения сопоставлены с матрицами ошибок классификатора, обученного с лучшим для данного набора данных параметром k.

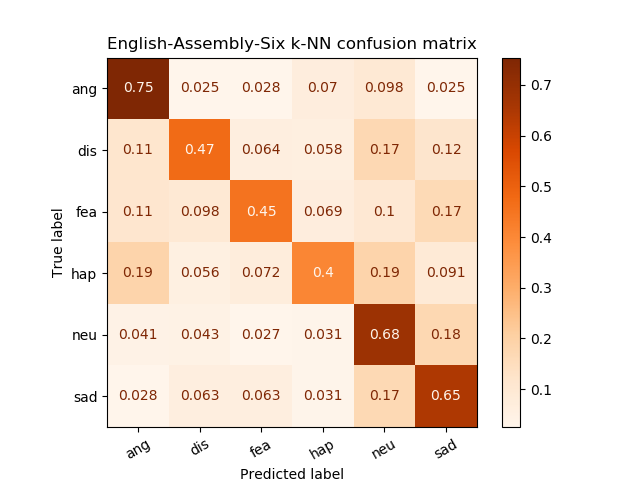
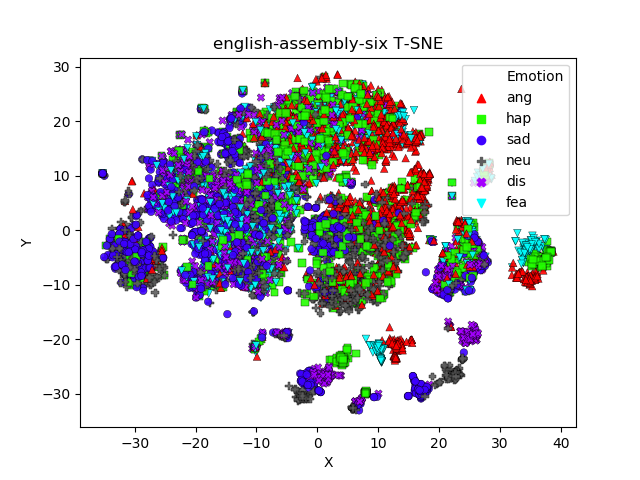


****

****

****

****

****

1. Выводы.

Глядя на эти визуализации и сопоставления их с матрицами ошибок, можно сделать массу интересных наблюдений и выводов, выделю главное:

1. В подавляющем большинстве случаев, у k-NN не возникает проблем с выявлением гнева, который является доминирующей базовой эмоцией, воспринимаемой из эмоционального спектра речи агрессора[11]. Также достаточно уверенно распознается печаль и нейтральная эмоциональная окраска. Нужно попробовать бинарную классификацию «Положительная эмоциональная окраска» vs. «Отрицательная эмоциональная окраска»
2. В некоторых наборах данных (savee, iemocap) T-SNE выявляет кластеры по отличному от эмоциональной окраски признаку. Это могут быть кластеры для говорящих разного пола. Применение разделения по полам может улучшить качество классификации.

Визуализации T-SNE сохранены в директории root/plots/tsne, матрицы ошибок – в директории root/plots/knn/confusion\_matrix

1. Дальнейший вектор

?????

Вместо многоклассового классификатора обучить бинарный «положительная эмоция» - «отрицательная эмоция»

Улучшить визуализации

Привести в порядок код

Попробовать SVM (не ожидаю значительной прибавки к точности)

Применить k-fold cross-validation, LOSA-cross-validation

Получить доступ к другим датасетам и провести те же исследования на них

Еще поиграться с данными (применить PCA и построить pairplot по первым четырем компонентам, например)

?????

1. Источники:

[1] Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., & Verma, R. (2014). CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset. IEEE Transactions on Affective Computing, 5(4), 377–390.

[2] Busso, Carlos & Bulut, Murtaza & Lee, Chi-Chun & Kazemzadeh, Abe & Mower Provost, Emily & Kim, Samuel & Chang, Jeannette & Lee, Sungbok & Narayanan, Shrikanth. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. Language Resources and Evaluation. 42. 335-359.

[3] Burkhardt, Felix & Paeschke, Astrid & Rolfes, M. & Sendlmeier, Walter & Weiss, Benjamin. (2005). A database of German emotional speech. 9th European Conference on Speech Communication and Technology. 5. 1517-1520.

[4] S. R. Livingstone и F. A. Russo, «The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English», PLOS ONE, т. 13, вып. 5, с. e0196391, май 2018.

[5] Jackson, Philip & ul haq, Sana. (2011). Surrey Audio-Visual Expressed Emotion (SAVEE) database.

[6] M. K. Pichora-Fuller и K. Dupuis, «Toronto emotional speech set (TESS)». Scholars Portal Dataverse, 2020

[7] Eyben, Florian & Wöllmer, Martin & Schuller, Björn. (2010). openSMILE -- The Munich Versatile and Fast Open-Source Audio Feature Extractor. MM'10 - Proceedings of the ACM Multimedia 2010 International Conference. 1459-1462.

[8] Eyben, Florian & Scherer, Klaus & Schuller, Björn & Sundberg, Johan & Andre, Elisabeth & Busso, Carlos & Devillers, Laurence & Epps, Julien & Laukka, Petri & Narayanan, Shrikanth & Truong, Khiet. (2015). The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. IEEE Transactions on Affective Computing. 7. 1-1.

[9] Maaten LV, Hinton G. Visualizing data using t-SNE. Journal of machine learning research. 2008;9(Nov):2579-605.

[10] Poličar P.G., Stražar M., Zupan B. (2019) Embedding to Reference t-SNE Space Addresses Batch Effects in Single-Cell Classification. In: Kralj Novak P., Šmuc T., Džeroski S. (eds) Discovery Science. DS 2019. Lecture Notes in Computer Science, vol 11828. Springer, Cham

[11] Комалова Л.Р.. "Перцептивно-слуховой профиль (образ) агрессора" Вестник Московского государственного лингвистического университета. Гуманитарные науки, no. 7 (746), 2016, pp. 116-126.