ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет информатики, математики и компьютерных наук

Программа подготовки бакалавров по направлению
01.03.02 Прикладная математика и информатика

О Т Ч Е Т
по проектной работе

Разработка многомодальной системы распознавания эмоциональных состояний человека

Выполнил студент гр. 20ПМИ-1
Дудникова Екатерина Олеговна

(подпись)

Руководитель проекта:
специалист по анализу данных:
Абросимов Кирилл Игоревич

_____*7*_____ _____ _____ _____
(оценка) *(подпись)*_____
21.03.2023

Оглавление

1. Общее описание проекта

Описание компании:

Заказчик - Национальный исследовательский университет «Высшая школа экономики».

Инициатор - Factory5 Dev (ООО «M5»).

Руководитель проекта - специалист по анализу данных Абросимов Кирилл Игоревич.

Место работы по проекту - дистанционно.

Описание технического задания:

Цель: система, распознающая эмоциональное состояние человека по видеомодальности, аудиомодальности и текстовой модальности.

Задачи:

1. Изучить современные (state-of-the-art) статьи по задаче распознавания эмоций;

2. Изучить и произвести статистический анализ открытого набора данных MELD;

3. Изучить и применить методы классического обучения;

4. Изучить и применить современные архитектуры нейронных сетей;

5. Изучить и применить методы извлечения признаков из видеомадальности (сверточные нейронные сети);

6. Изучить и применить методы извлечения признаков из аудиомодальности (mel-спектрограммы, toolkit openSMILE);

7. Изучить и применить методы извлечения признаков из текстовой модальности (Вектора, на основе тональных словарей, BERT);

8. Для каждой модальности построить классификатор и оценить качество;

9. Объединить модальности на уровне векторов и построить общий классификатор, оценить качество;

10. Объединить модальности на уровне принятия решений, т.е. Построить классификатор, принимающий решение на основе предсказаний локальных классификаторов, построенных на определенной модальности.

2. Содержательная часть

2.1 Описание хода выполнения проектного задания

1) **В первую очередь было решено заняться изучением самого датасета MELD и теорией по классификации эмоций**

Было изучено количество и вид данных, имеющихся в датасете, в частности длина текста в словах, длина аудио/видео в секундах (и фреймах для видео), и самое главное – распределение эмоций (классов) по тренировочной, валидационной и тестовой выборкам. Спойлер, датасет с явным и весьма большим дисбалансом классов:

Распределение эмоций в датасете:

Эмоции	Тренировочная	Валидационная	Тестовая
Злость	1109	153	345
Отвращение	271	22	68
Страх	268	40	50
Радость	1743	163	402
Нейтральное	4710	469	1256
Грусть	683	111	208
Удивление	1205	150	281

Недостатки датасета:

- Датасет несбалансированный. Класс нейтрального настроения составляет 48% от всей выборки.
- Опираясь на описание из файлов .csv, некоторые видео не соответствуют фразе/диалогу/высказыванию, что будет путать модели в будущем.
- В одном диалоге могут присутствовать много актеров.
- Много посторонних звуков

2) **Далее было изучение имеющихся на просторах интернета методов и результатов работы с этим датасетом.**

Результаты будут представлены в последующих разделах.

3) **Следующим шагом была предобработка данных и извлечение различных признаков для будущих моделей.**

Видео:

Предобработка заключалась в выделении фреймов из видео (каждого 10-го), затем с помощью предобученной сети MTCNN из фреймов выделялось лицо, а после оно вырезалось из картинки и сохранялось. Лицо выбиралось с наибольшей вероятностью (сеть выдавала вероятности на каждый квадрат изображения, который считала лицом), и если самая большая из возможных вероятностей меньше 85%, то считалось, что лица на кадре нет.

Текст:

Для данной задачи были опробованы классические методы предобработки текста, в частности:

1. Токенизация (Разделение предложения на части (униграммы, биграммы, триграммы…))
2. Приведение текста к нижнему регистру;
3. Удаление стоп-слов и знаков препинания (Словари с разными стоп-словами можно найти в интернете и свободно использовать);
4. Стемминг/Лемматизация (Стемминг – обрезание слова до его основы, лемматизация – приведение слов к начальной форме);
5. Регулярные выражения;

Забавно, что конкретно в нашей задаче все сложные предобработки (1, 3, 4) плохо работали и давали результат хуже, чем просто удаление знаков препинания и приведение к нижнему регистру.

4) **Построение унимодальных систем для своих модальностей (видео, текст)**

Видео:

Для обработки выделенных признаков из видео, был выбран предобученный ResNet18, но позже были получены результаты лучше с помощью ResNet50 – большая сверточная нейронная сеть с глубиной 50 слоев, к ней добавлялся полносвязный слой-классификатор. Результаты:

	precision	recall	f1-score	support
0	0.18	0.06	0.09	345
1	0.00	0.00	0.00	68
2	0.00	0.00	0.00	50
3	0.18	0.40	0.25	402
4	0.47	0.60	0.52	1256
5	0.16	0.02	0.03	208
6	0.18	0.02	0.04	281
accuracy			0.35	2610
macro avg	0.17	0.16	0.13	2610
weighted avg	0.30	0.35	0.30	2610

Текст:

В проекте было несколько вариантов обработки текста: с помощью предобученного трансформера «RoBERTa», которым занималась не я, а также классические методы обработки текста. Было опробовано несколько разных моделей, в частности простые TF-IDF, Bag of words, а также Bag of words + Sentiment (SentiwordNet).

Результаты для TF-IDF:

CLASSIFICATIION METRICS

	precision	recall	f1-score	support
anger	0.33	0.07	0.12	345
disgust	0.80	0.06	0.11	68
fear	0.17	0.02	0.04	50
joy	0.48	0.21	0.30	402
neutral	0.54	0.90	0.67	1256
sadness	0.52	0.07	0.13	208
surprise	0.53	0.41	0.46	281
accuracy			0.53	2610
macro avg	0.48	0.25	0.26	2610
weighted avg	0.50	0.53	0.45	2610

Результаты для Bag Of Words:

CLASSIFICATIION METRICS

	precision	recall	f1-score	support
0	0.68	0.89	0.77	1256
1	0.48	0.26	0.34	345
2	0.50	0.06	0.11	68
3	0.14	0.02	0.04	50
4	0.48	0.53	0.50	402
5	0.39	0.12	0.19	208
6	0.55	0.51	0.53	281
accuracy			0.61	2610
macro avg	0.46	0.34	0.35	2610
weighted avg	0.57	0.61	0.57	2610

Результаты выше, чем для TF-IDF, но скорее всего можно лучше. Далее попробовала использовать доп. Столбик Sentiment при таком же способе выделения эмбеддингов – BoW.

CLASSIFICATIION METRICS

	precision	recall	f1-score	support
0	0.70	0.85	0.77	1256
1	0.40	0.37	0.39	345
2	1.00	0.03	0.06	68
3	0.00	0.00	0.00	50
4	0.49	0.51	0.50	402
5	0.34	0.20	0.25	208
6	0.52	0.42	0.46	281
accuracy			0.60	2610
macro avg	0.49	0.34	0.35	2610
weighted avg	0.57	0.60	0.57	2610

BoW + вектора из Sentiwordnet (тонального словаря), построенные по логике: [максимальное значение позитивного класса, среднее по позитивным оценкам слов в предложении, макс. значение негативное, среднее по негативным оценкам, количество слов с позитивной окраской, количество слов с негативной окраской, количество слов с нейтральным окрасом в предложении]

	precision	recall	f1-score	support
0	0.72	0.81	0.76	1256
1	0.46	0.33	0.38	345
2	0.26	0.07	0.11	68
3	0.11	0.08	0.09	50
4	0.48	0.51	0.50	402
5	0.31	0.19	0.24	208
6	0.49	0.59	0.53	281
accuracy			0.59	2610
macro avg	0.40	0.37	0.37	2610
weighted avg	0.57	0.59	0.57	2610

Класс	Номер
neutral	0
anger	1
disgust	2
fear	3
joy	4
sadness	5
surprise	6

Результат на тесте для этой системы. Да, общий скор не очень высок, но примечательно, что все классы предсказываются с каким-то значением. (disgust и fear игнорируются большинством моделей, так как их количество мало в тренировочной выборке). Интересующая нас метрика balanced accuracy = 0.37, что больше, чем в когда использовался чистый Bag of Words для выделения признаков.

5) Построение мультимодальной системы.

Мультимодальная система строилась в стратегиях *Late Fusion* – в качестве входных данных использовались полученные предсказания из унимодальных систем – и *Early Fusion* – для предсказаний использовались только эмбеддинги из унимодальных систем.

1) Войтинг

Метод позднего слияния, основанный на получении предсказаний с унимодальных моделей (текст, аудио, видео) и выбора среди них самого популярного варианта. Если все модели давали разный ответ, то выбиралось предсказание текстового классификатора, так как он имеет большую точность.

	precision	recall	f1-score	support
anger	0.50	0.26	0.35	345
disgust	0.00	0.00	0.00	68
fear	0.00	0.00	0.00	50
joy	0.56	0.46	0.51	402
neutral	0.63	0.92	0.75	1256
sadness	0.47	0.11	0.17	208
surprise	0.56	0.43	0.49	281
accuracy			0.60	2610
macro avg	0.39	0.31	0.32	2610
weighted avg	0.55	0.60	0.55	2610

Модель не считаем хорошей, так как полностью игнорируется 2 класса.

2) По предсказаниям – конкретному классу

3 предобученные унимодальные системы выдают предсказания для трейновых, валидационных и тестовых данных, затем на этих данных происходит обучение какого-нибудь классификатора, который выбирался на кросс-валидации:

model_name	Mean Accuracy	Standard deviation
CatBoostClassifier	0.770446	0.014717
LGBMClassifier	0.769745	0.014798
LinearSVC	0.579638	0.006039
LogisticRegressior	0.572332	0.014939
RandomForestCla	0.739011	0.012758

Catboost дал лучший результат, поэтому был обучен и итоговый результат на тесте:

	precision	recall	f1-score	support
anger	0.47	0.51	0.49	345
disgust	0.00	0.00	0.00	68
fear	0.00	0.00	0.00	50
joy	0.59	0.60	0.60	402
neutral	0.74	0.84	0.79	1256
sadness	0.48	0.23	0.31	208
surprise	0.53	0.60	0.57	281
accuracy			0.65	2610
macro avg	0.40	0.40	0.39	2610
weighted avg	0.61	0.65	0.62	2610

То же самое, очень плохо предсказывает fear и disgust (игнорирует), следовательно, эту модель тоже было решено не использовать в итоговом продукте.

3) По предсказаниям – вероятностям попадания элемента в конкретный класс.

Чтобы получить результаты в виде вероятностей, была использована функция softmax на предсказаниях из нейронных сетей – унимодальных моделей.

model_name	Mean Accuracy	Standard deviation
CatBoostClassifier	0.880167	0.011947
LGBMClassifier	0.877765	0.012297
LinearSVC	0.867954	0.014246
LogisticRegression	0.856842	0.013873
RandomForestClassi	0.839223	0.017254

На кросс-валидации выбирался классификатор. В данном случае лучше всех себя показали CatBoost и LGBM классификаторы, но на самом деле при подборе параметров метод опорных векторов (LinearCSV классификатор) показал себя лучше всего. Именно для этого варианта получился лучший результат по accuracy:

CLASSIFICATIION METRICS

	precision	recall	f1-score	support
anger	0.48	0.42	0.45	345
disgust	0.00	0.00	0.00	68
fear	0.00	0.00	0.00	50
joy	0.59	0.67	0.62	402
neutral	0.77	0.83	0.80	1256
sadness	0.43	0.30	0.35	208
surprise	0.54	0.67	0.60	281
accuracy			0.65	2610
macro avg	0.40	0.41	0.40	2610
weighted avg	0.62	0.65	0.63	2610

Но, как назло, снова не предсказываются 2 класса из 7, то есть модель не подходит. При дальнейшем подборе параметров выявился лучший результат, который после и отправился в конечный продукт проекта:

CLASSIFICATIION METRICS

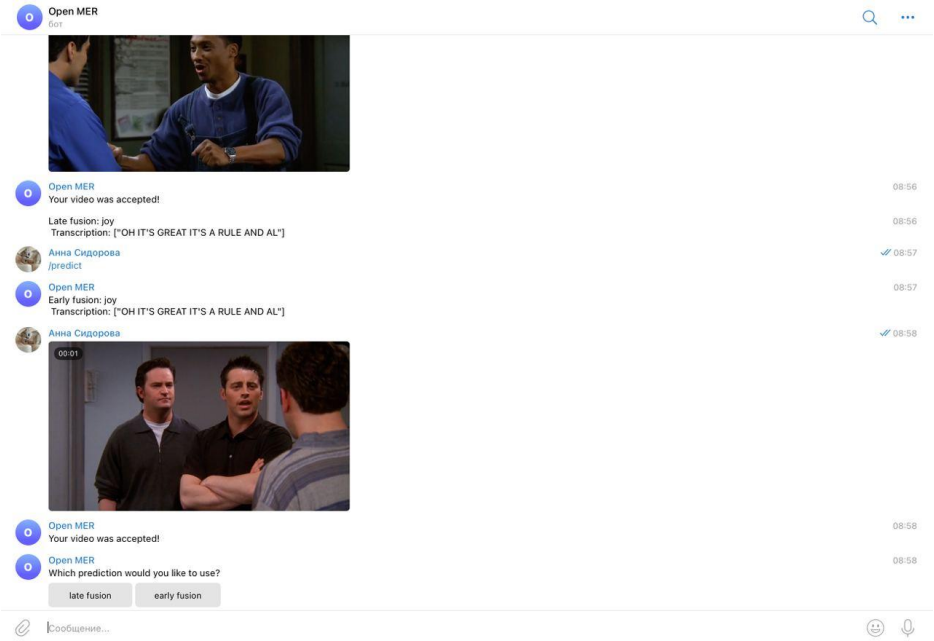
	precision	recall	f1-score	support
anger	0.47	0.51	0.49	345
disgust	0.15	0.04	0.07	68
fear	0.18	0.04	0.07	50
joy	0.56	0.61	0.58	402
neutral	0.78	0.79	0.78	1256
sadness	0.36	0.29	0.32	208
surprise	0.55	0.62	0.58	281
accuracy			0.63	2610
macro avg	0.43	0.41	0.41	2610
weighted avg	0.62	0.63	0.62	2610

6) Написание бота, который используя написанную командой локальную библиотеку, использующую лучшие модели, предсказывает эмоции по видео.

С помощью библиотеки python-telegram-bot был написан бот, который может принимать на вход видео/документ в формате .mp4 и выдавать предсказания на основе построенных нами моделями.

2.2 Описание результатов проекта/продукта

В результате работы проекта получился Telegram бот, который получая на вход видео может предсказать эмоцию двумя разными способами: 1. Late fusion 2. Early fusion – это были два вида создания многомодальной системы.



Результаты	balanced accuracy	accuracy	accuracy Статья[1]	f1 weighted	f1 weighted Статья[1]
Only Text	0.399841	0.6533	0.6724	0.627947	0.6623
Only Audio	0.264384	0.3605	0.4904	0.376306	0.3963
Only Video	0.156249	0.3745	0.4563	0.30669	0.3244
Early Fusion	0.469572	0.6134	0.6728	0.622541	0.6681
Late Fusion	0.413862	0.6341	0.6785	0.622822	0.6671

Наши результаты выглядят хуже, чем в статье, но конкретно эти метрики не являются показательными в датасетах с дисбалансом классов, а других там предоставлено не было. Считаем, что наш результат очень хороший, потому

что в погоне за accuracy и f1 weighted можно забыть про действительно важные и имеющие роль метрики, а мы не забыли. (balanced accuracy)

2.3 Описание используемых в ходе проекта способов и технологий

В ходе проекта были изучены способы работы с видео/изображениями с помощью языка программирования python (Библиотеки PIL, OpenCV), способы работы с предобученными нейронными сетями (MTCNN, ResNet…), способ дообучения нейросетей (ResNet18, ResNet50). Классические методы обработки текста (которые пригодились для работы с тональным словарем), виды векторизации текста TF-IDF и Bag Of Words, основы работы с тональными словарями.

2.4 Описание своей роли в командной работе

В проекте я была ответственна за видео-модальность, а также частично за текст, работала над созданием многомодальной системы с помощью стратегии Late Fusion и в конце написала telegram бота, который и является итоговым продуктом нашего проекта.

2.5 Описание отклонений и трудностей, которые возникли в ходе проекта

1) В ходе проекта у нас команда уменьшилась на одного человека

2) Изначально были немного неверно рассчитаны сроки, так что времени на работу было на месяц меньше, чем рассчитывалось изначально.

3) Что хотелось бы еще сделать в данном проекте:

- Глобально улучшить результат по видео. Он относительно приемлем, но хотелось бы лучше (в частности, чтобы по видео предсказывались все 7 классов).

- Углубиться в работу с тональными словарями и в общем в сентимент анализ, мне кажется, что это может помочь улучшить модель, выдающую предсказания по тексту.

3. Заключение

В ходе проекта было достаточно много трудностей, так как анализ данных на момент начала проекта был темным лесом – манящим и неизведанным. К концу проекта я понимаю, что моя работа выглядит достаточно наивно и относительно просто, но в процессе я очень многому научилась и очень многое узнала. Если бы я начинала проект заново, имя багаж знаний, который имею сейчас, многое бы поменялось, и результат бы определенно улучшился. Считаю, что это показатель моего развития на проекте, поэтому он прошел не зря. Очень многого достигла:

Получилось:

- Поработать с видео/изображениями

- Использовать предобученные модели MTCNN и ResNet для выделения признаков и предсказания эмоций по видеоданным

- Поработать над задачей классификации эмоций, попробовать применить тональные словари, методы выделения признаков из текста TF-IDF и Bag Of Words

- Изучила разные варианты создания многомодальной системы по стратегии Late Fusion

- Написала telegram бота

4. Источники

[1] V. Chudasama, P. Kar, A. Gudmalwar, N. Shah, P. Wasnik, and N. Onoe. \emph{M2FNet: Multi-modal Fusion Network for Emotion Recognition in Conversation}. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pages 4652–4661, June 2022.