СОФИЙСКИ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ”

СТОПАНСКИ ФАКУЛТЕТ

Курсов проект по дисциплината
„Основи на текстовия анализ и обработката на естествен език”:

Анализ на настроението върху данни в сферата на административните услуги

Изготвил:

Слав Йолов
№ 5EB3100235

Георги Стоянов
№ 6EB3100244

Любен Зарев
№ 0EB3100293

Преподавател:

ас. д-р Глория Христова

София
Януари 2023

I. Въведение в казуса

Настоящият курсов проект представлява анализ на настроенията върху данни събрани от онлайн форуми и медии, обсъждащи дигитализацията на държавни и административни услуги. Целта на това изследване е да се разбере всеобщото мнение на гражданите към тези услуги, както и да се идентифицират някои основни болезнени точки или области, нуждаещи се от подобрение.

За тази цел използваме техники за обработка на естествен език. Текстовите данни биват анализирани и класифицирани в категории „положително“, „отрицателно“ и „неутрално“ настроение. Впрягаме в действие три различни алгоритъма за тази задача: baseline алгоритъм, VADER и TextBlob.

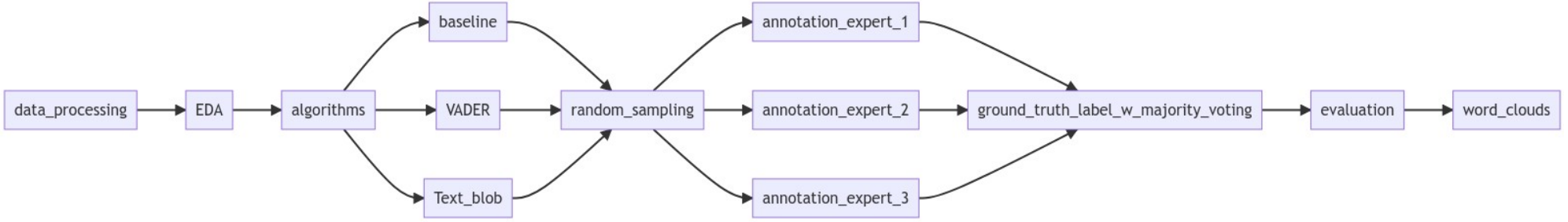
В допълнение към класифицирането на настроението на текста, също използваме инструменти за визуализация на данни, за да представим резултатите и да подчертаем всички модели или тенденции, които се появяват. Това ни позволява лесно да идентифицираме всички основни въпроси или проблемни области, които често се споменават в данните.

В изследването си, сравняваме резултатите от трите използвани алгоритъма и оценяваме ефективността на всеки по отношение на прецизност и F1-score. Съпоставката на тези алгоритми ще помогне да се определи кой е най-подходящ за задачата върху този конкретен набор от данни, и кой ще ни предостави най-достоверни резултати.

Това проучване има потенциала да послужи като ценен източник на информация за държавни служители и администратори, които искат да подобрят предлаганите административни услуги. Разбирайки настроенията на обществеността, служителите могат да идентифицират области, които се нуждаят от подобрение, и да предприемат стъпки за справяне с тях.

II. Методология

Графиката по-долу показва стъпките, през които екипът премина, за да достигне до предложения анализ на административните данни:



1. Запознаване с данните
2. Начална обработка на данните
3. Селектиране на алгоритми, които ще използваме в sentiment анализа.
4. Използване на следните три алгоритъма: baseline, VADER, text_blob.
5. Прилагане на random sampling и селектиране на определени редове от анализираната извадка.
6. Анотиране на данните от трима от членовете на екипа.
7. Анализ на получените резултати
8. Оценка на резултатите
9. Създаване на облаци от думи (word clouds), които визуализират честотата и важността на използваните думи.

Започваме работа, като импортираме серия от библиотеки, които сме идентифицирали, че ще използваме :

- pyhocon==0.3.59
- pandas==1.5.3
- dill==0.3.6
- tabulate==0.9.0
- pandas-profiling==3.6.2
- wordcloud==1.8.2.2
- nltk==3.8.1
- seaborn==0.12.2
- openpyxl==3.0.10
- vaderSentiment==3.3.2
- textblob==0.17.1
- scikit-learn==1.2.1

В конфигурационен файл сме поставили пътят към административните данни.

След като прочитаме данните, филтрираме единствено тези от тях, които имат основен текст на английски език. След филтрирането остават около 1 700 записа.

По категории филтрираните данни изглеждат така:

- Government website – 873 записа
- Forum – 501 записа
- Blog – 165 записа
- Company website – 107 записа
- Radio – 92 записа
- Video site – 2 записа.

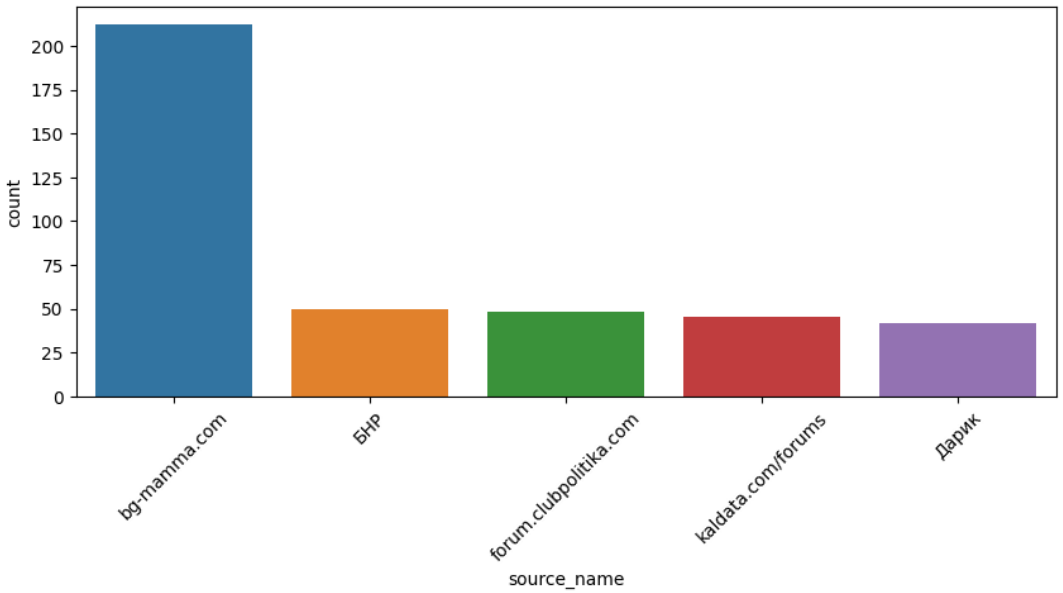
След преглед на данните, решаваме с допълнително филтриране да премахнем редовете от категория Government website и company website с оглед на това, че те не съдържат мнения, а по-скоро текстовете там имат информативен характер. След второто филтриране сетът ни остава със следните данни:

- Forum – 501 записа
- Blog – 165 записа
- Radio – 92 записа
- Video site – 2 записа.

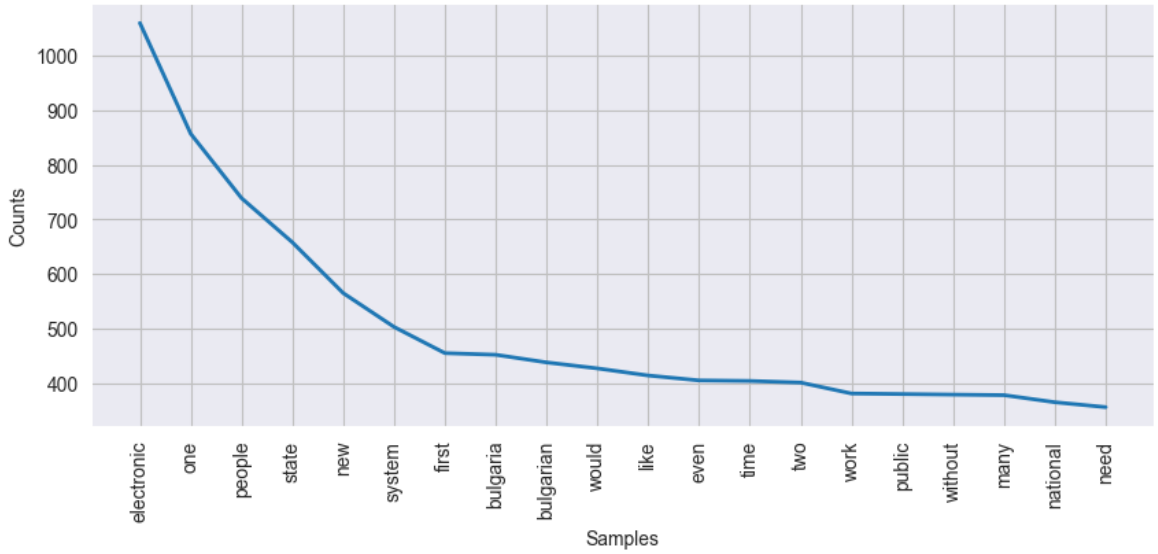
Общо записи за анализ: 760

Както е видно от графиката по-долу, най-много източници на информация имаме от следните сайтове:

- Bg-mamma.com
- БНР
- Forum.clubpolitika.com
- Kaldata.com/Forums
- Дарик



От анализираните 760 реда изваждаме следния frequency distribution на най-често срещаните думи в извадката:



Често срещани думи са:

- Electronic
- One
- People
- State
- New
- System

Всяка от тези думи има е използвана между 1100 и 500 пъти в извадката, която анализираме.

Следващата стъпка е да стартираме реалния sentiment analysis на избраните данни. Използваме три различни алгоритъма за sentiment анализа - baseline, VADER, text-blob.

Повече за селектираните модели:

Baseline

Моделът избира за всеки текст на произволен принцип една от опциите (positive, negative, neutral). Идеята за него е да ни даде базова и неточна прогноза, която би трябвало да бием с по-напреднали технологии. Ако алгоритъм не е в състояние да бие базовия модел, то той спокойно може да не бъде отхвърлен като неподходящ за текущия казус.

VADER-Sentiment-Analysis

VADER (Valence Aware Dictionary and sEntiment Reasoner) често се използва за sentiment анализ на мнения, изразени в социалните мрежи. VADER е подход базиран на правила, който използва лексикони, за да оцени дадени думи и изречения (positive, negative, neutral)

Характерно за него е да борави с обединени стойности (compund). Те се получават чрез събиране/съмъризиране на валентността (valence) на отделни думи в лексикона и след това нормализирани между [-1;1], където -1 е екстремно негативно изречения, а +1 е екстремно позитивно.

Text-blob

Бибилиотека за обработка на текстова информация. В библиотеката има компонент за анализ на настроенията (sentiment). Методът измерва полярността [-1, 1], където -1 е екстремно негативно изречения, а +1 е екстремно позитивно. Също така измерва и субективността [0.0, 1.0], измерител за фактуална или субективна информация. За целта на казуса ще използваме само измерителят за полярност като ще допуснем, че неутралните стойности варират в диапазона [-0.05, 0.05].

III. Резултати

Съобразно документацията на избраните модели, залагаме следната логика:

- Score >= 0.05 - positive sentiment
- Score > -0.05 and score < 0.05 - neutral sentiment
- Score <= -0.05 - negative sentiment

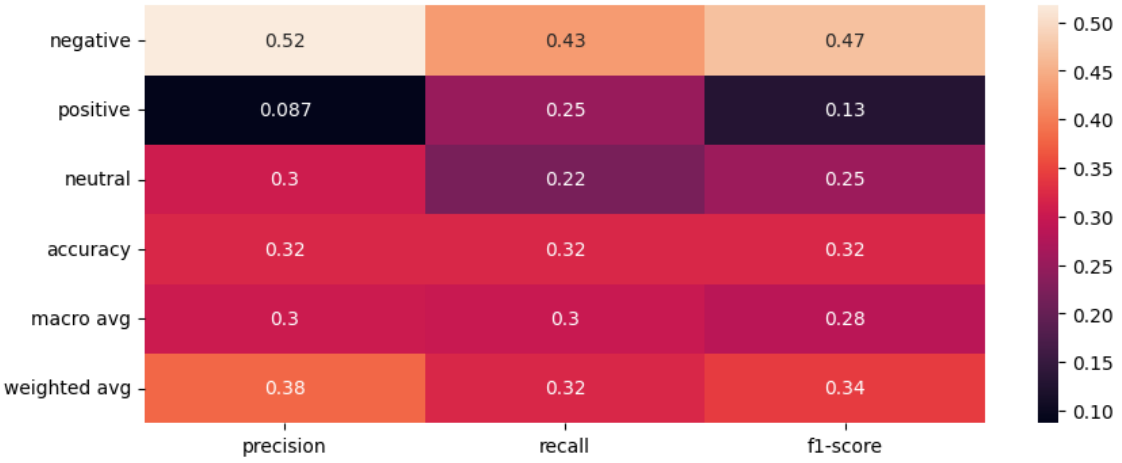
Текущата задача е от тип Unsupervised Learning, няма целева колона, която да се стараем да предвиди. Поради това, за да оценим решенията екипът реши да направи cross анотация на извадка от данните. Чрез random sampling избираме 75 записа (по 25 от всеки вид). Записите са взети на база категоризацията направена с VADER. Всеки от членовете на екипа анотира съответната извадка с една от трите стойности (positive, negative, neutral) като получените резултати се агрегират посредством най-популярно решение (majority voting), за да се избере таргет стойност за всеки ред. Тези таргет стойности биват използвани, за да оценим кой от моделите се представя най-добре. Оценката е направена посредством macro f1-score:

- Baseline: 0.28
- VADER: 0.49
- Text_blob: 0.3

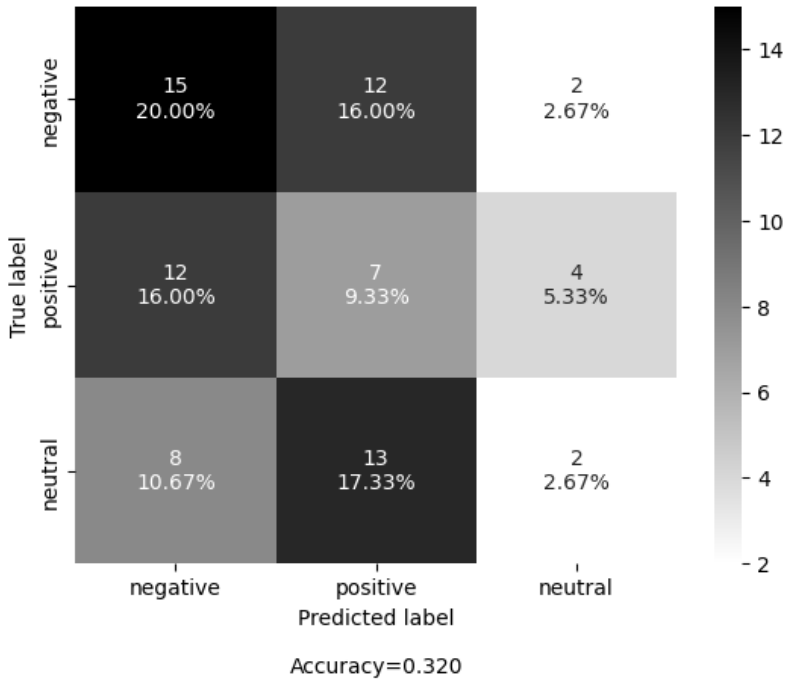
Въпреки че f1-score и на трите модела е сравнително нисък, VADER успява да постигне по-добър резултат спрямо останалите два подхода. VADER е в състояние да се справи сравнително добре в сравнително сложни текстове от различен характер.

Сентимент анализът посредством text_blob показва повече грешки спрямо VADER. Алгоритъмът е по-несигурен да оцени дали един текст е положителен или отрицателен и често алгоритъмът ги класифицира като неутрални. Резултатът на алгоритъм е близък до случаен избор, което е достатъчно да го отхвърлим като опция за текущия дейтасет.

1. Baseline

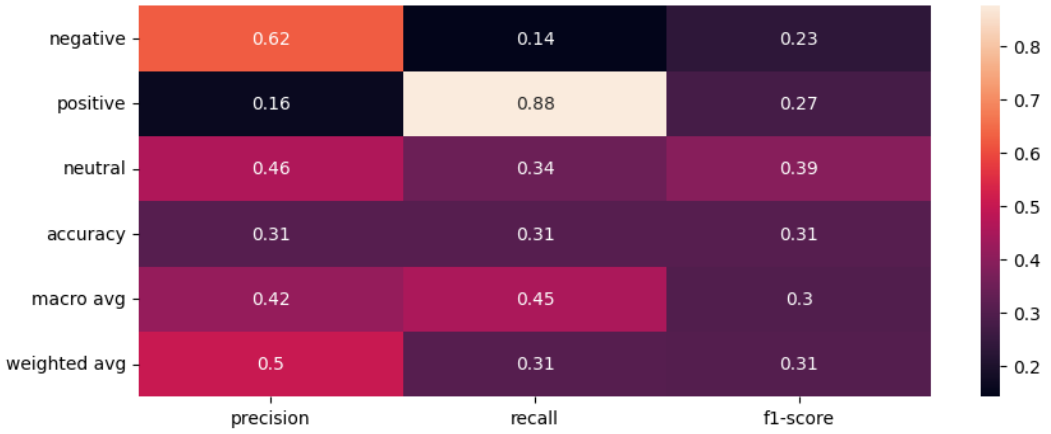


Фиг. 3а: Classification report -baseline

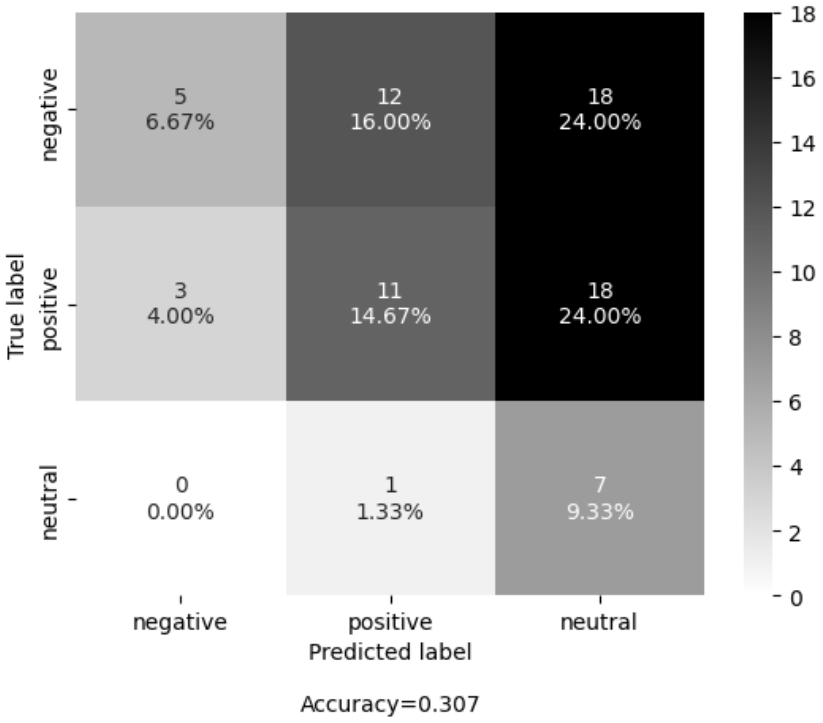


Фиг. 3б: Confusion matrix - baseline

2. Text-blob

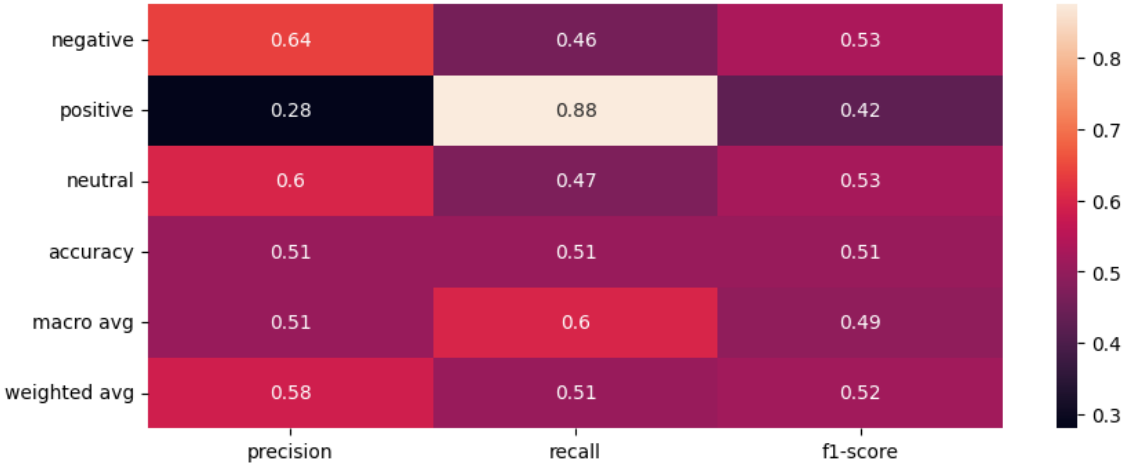


Фиг. 4а: Classification report - text-blob

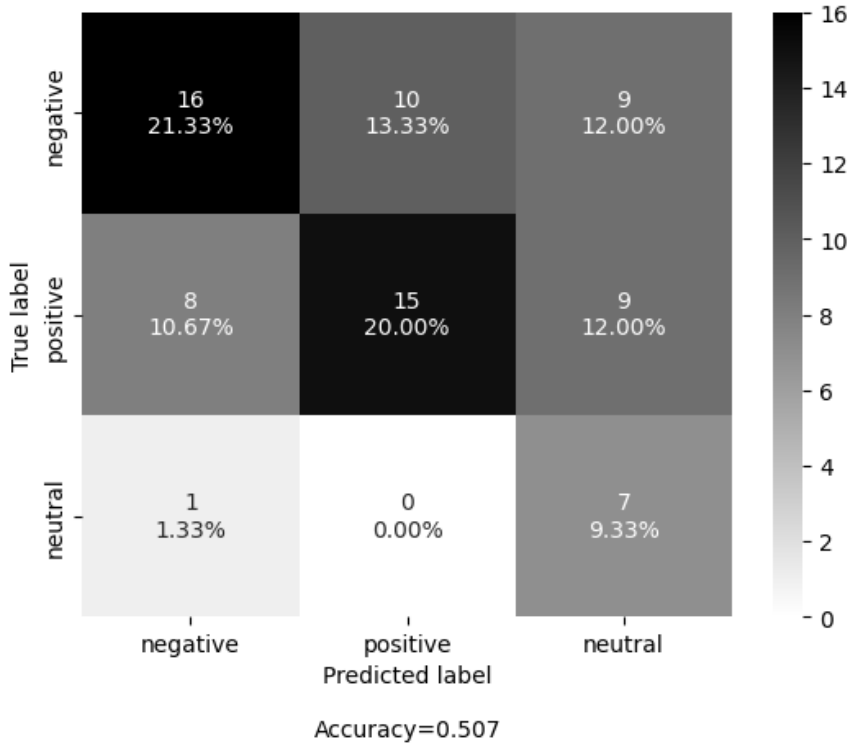


Фиг. 4б: Confusion matrix - text-blob

3. VADER

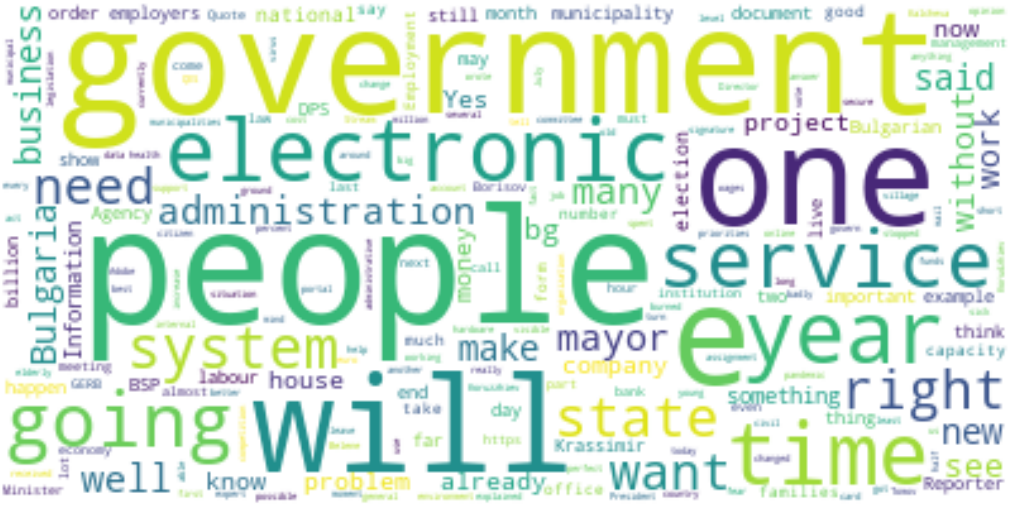


Фиг. 5а: Classification report - VADER



Фиг. 5б: Confusion matrix - VADER

- **Words with a negative sentiment**



- **Words with a neutral sentiment**

- **Words with a positive sentiment**

IV. Заключение

Това проучване предоставя ценна информация за настроенията на обществеността към държавните и административни услуги. Чрез използване на техники за обработка на естествен език и алгоритми за анализ на настроението, успяхме да класифицираме част от предоставения набор от данни. Резултатите от този анализ могат да послужат при евентуално модернизиране и цифровизация на публичните услуги в България.

Сравнявайки резултатите от трите използвани алгоритъма - baseline, VADER и TextBlob, успяхме да оценим техния потенциал в справянето със задачата. На базата на това сравнение стигнахме до заключението, че алгоритъмът VADER се представя по-добре от другите два.

Изследването може да бъде разширено и продължено по няколко начина в бъдеще. Един потенциален метод е да се разшири обхватът на изследването, да се включат допълнителни източници на данни, като социални мрежи и платформи. По този начин ще се увеличи размерът на извадката и ще се подобри представителността на данните. Друга възможност е да се включат по-напреднали NLP техники като моделиране на теми (topic modeling), за да се идентифицират основните теми, обсъждани в данните, и как те се отнасят към настроенията. Освен това, бъдещите анализи биха могли да експериментират с използването на модели за машинно обучение, което може да доведе до подобрена точност и производителност.

В обобщение, проектът представя информация за настроенията на обществеността към административните услуги и има потенциала да информира за подобрения в този сектор. Бъдещи изследвания могат да развият казуса, черпейки текстови данни от по-наситените от информация и емоция социални мрежи, да подобрят представителността на данните, включвайки по-напреднали NLP техники, и да изследват използването на ML модели.

V. Преглед на използваната литература

TextBlob documentation. (n.d.). Retrieved from https://textblob.readthedocs.io/en/dev/

Hutto, C.J., and Eric Gilbert. "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text." Eighth International Conference on Weblogs and Social Media (ICWSM-14), 2014, https://ojs.aaai.org/index.php/ICWSM/article/view/14550

"Sentiment Analysis: Baseline Algorithm Using NLP in Machine Learning." Analytics Vidhya, Medium, 21 Sept. 2019, https://medium.com/analytics-vidhya/sentiment-analysis-baseline-algorithm-using-nlp-in-machine-learning-7f9b3e52f8a.

Malde, Ravi. "A Short Introduction to VADER." Towards Data Science, Medium, 22 Dec. 2020, https://towardsdatascience.com/an-short-introduction-to-vader-3f3860208d53.

Shah, P. My absolute go-to for sentiment analysis - textblob., Medium. Towards Data Science., https://towardsdatascience.com/my-absolute-go-to-for-sentiment-analysis-textblob-3ac3a11d524 (Accessed: January 30, 2023).

Leung, K. (2022) *Micro, Macro & weighted averages of F1 score, clearly explained*, *Medium*. Towards Data Science., https://towardsdatascience.com/micro-macro-weighted-averages-of-f1-score-clearly-explained-b603420b292f#:~:text=The%20macro%2Daveraged%20F1%20score,regardless%20of%20their%20support%20values.&text=The%20value%20of%200.58%20we,score%20in%20our%20classification%20report. (Accessed: January 30, 2023).