

Универзитет у нишу

Електронски факултет

- пројекат -

**АНАЛИЗА СЕНТИМЕНАТА**

**ПРЕДМЕТ**: ОБРАДА ПРИРОДНИХ ЈЕЗИКА

**Ментор**: Сузана Стојковић **Студент:** Филип Трајковић 1574

Ниш, 2024.

Садржај

[1. Увод 3](#_Toc159517377)

[2. Анализа сентимената у обради природних језика 4](#_Toc159517378)

[2.1. Теоријске основе анализе сентимената 4](#_Toc159517379)

[2.2. Примена анализе сентимената 4](#_Toc159517380)

[2.3. Технике обраде природног језика 5](#_Toc159517381)

[2.3.1. Препроцесирање података 5](#_Toc159517382)

[2.3.2. Припрема података за тренирање 7](#_Toc159517383)

[3. Практични део 11](#_Toc159517384)

[3.1. Учитавање библиотека 11](#_Toc159517385)

[3.2. Скуп података 12](#_Toc159517386)

[3.3. Препроцесирање скупа података 13](#_Toc159517387)

[3.4. Експлоративна анализа података 14](#_Toc159517388)

[3.5. Припрема података за тренирање модела 19](#_Toc159517389)

[4. Закључак 25](#_Toc159517390)

[5. Литература 26](#_Toc159517391)

# Увод

У савременом добу где се вештачка интелигенција широм примењује у различитим областима софтверског развоја, кључно је разумети њен начин функционисања и препознати потенцијалне могућности за њену примену. Да бисмо идентификовали те могућности, прво је битно стећи увид у различите технологије, савладати бројне концепте који су развијени у последње време, а такође и разумети како применити одређене технологије у конкретним сценаријима. Кроз ова разумевања, могуће је креирати систем вештачке интелигенције способан за решавање различитих проблема.

Системи за детекцију, разумевање и контекстуализацију текстуалних записа представљају широку област која се веома брзо развија почетком треће деценије 21. века. Убрзани развој и помама за оваквим системима креће крајем 2022. године са појавом до сада невиђеног система за обраду текстова под називом *eng*. *ChatGPT*. Широке корисничке масе које су се уз помоћ овог система упознале са могућности коју нуде системи за обраду текстова, додатно су убрзале даљи развој и напредак у обради природних језика стварајући нове случајеве коришћења овакве технологије.

Анализа сентимената представља један од основних случајева коришћења вештачке интелигенције у области обраде природних језика. Користе се процеси класичног машинског учења и процеси обраде природних језика како би се креирала једна симбиоза различитих области са циљем решавања специфичних проблема попут анализе сентименталних особина текстуалних записа. Анализа сентимената текстуалних записа представља област надгледаног учења у машинском учењу при чему је неопходно да се за сваки запис у скупу података наведе и припадност категорији којој припада наведени текст. Сви записи у скупу података требају припадати коначном скупу категорија. Решавање проблема анализе сентимената се своди на проблем класификације у машинском учењу.

Даљи текст овог рада је подељен на теоријски и практични део. У теоријском делу рада су обрађене технике обраде текстуалних података, док су у практичном делу рада приказани резултати ових обрада.

Теоријски део рада обухвата основну теоријску позадину иза техника обраде текстуалних података, док практични део рада обухвата примену наведених техника над специфичним скупом података са применом класификационих алгоритама машинског учења, као и уз помоћ претренираних језичких модела.

# Анализа сентимената у обради природних језика

Анализа сентимената представља област у обради природних језика која се бави проучавањем означених скупова података при чему се скуп података састоји од текстуалног записа и категоричног атрибута који представља категорију припадности одређеној категорији у скупу коначних категорија.

## Теоријске основе анализе сентимената

Концепт анализе сентимената као специфичног случаја обраде текстуалних података представља процес утврђивања квалитативних својстава текстуалних записа како би се из датог скупа података извеле одређене законитости које важе између контекстуалних репрезентација посматраних записа и категоричког атрибута који представља припадност одређеном скупу.

Циљ овог процеса јесте утврђивање емоционалног контекста датих текстова и тежину коју речи природног језика имају у односу на стање сентимента који описују. Ставови тј. контекст који се преноси кроз текст се обично означавају у неколико категорија попут:

* Позитивно
* Негативно
* Неутрално

## Примена анализе сентимената

Процес утврђивања сентемената у тексту са собом доноси нове случајеве примене оваквог типа технологија у различитим сферама. Примена анализе сентимената има највећих бенефита у сферама које се баве обрадом и анализом мишљења корисника, оцену одређених ентитета (брендова, производа, итд.), нивоима одушевљења или разочарења посматраног корисника и сл. Овакви случајеви коришћења могу додатно допринети анализи одређених појава јер до сада није било могуће аутоматизовати анализу и процесирање текстова на овај начин.

Овакав аутоматизован начин обраде великих текстуалних скупова података са собом доноси убрзани начин обраде и анализе различитих текстова.

Коначни резултати утврђивања контекста текстова имају за циљ презентовање јасније слике о стеченим мишљењима различитих корисника над датим објектима посматрања при чему се у анализу могу укључити и додатни параметри који могу додатно категоризовати типове личности корисника.

## Технике обраде природног језика

Технике обраде природног језика представљају технике којима се обрађују подаци који се у изворном облику налазе у текстуалном формату. Оваквим техникама се врши обрада и припрема података за даље кораке у развоју процеса анализе сентимената попут примене математичких метода машинског учења или тренирања модела.

Обзиром да над подацима који се изворно налазе у текстуалном формату није могуће применити споменуте даље кораке, такве податке је најпре неопходно претворити у погодан формат који се може обрађивати на хардверу који се користи у машинском учењу. Погодан формат јесте векторски/тензорски простор у којем је неопходно представити текстуални контекст посматраног скупа података.

Обрађени текстуални подаци постају вектори који се у векторском/тензорском простору представљају као низови бројева одређене дужине. Вредности које ови вектори носе са собом требају на високо-квалитативан начин описати дати текстуални податак при чему се требају одржати својства текстова попут редоследа речи у реченици, контекстуалне зависности између речи, дужина текстова, контекстуална сличности и сл.

### Препроцесирање података

Препроцесирање података се бави основним чишћењем улазних текстуалних података и подразумева неке од следећи техника:

* Чишћење текста
* Нормализација
* Токенизација
* Лематизација и стемизација
* Уклањање стоп-речи

Чишћење текста

Чишћење текста је први корак у препроцесирању текстуалних података и подразумева уклањање специјалних карактера из текста попут HTML тагова, знакова интерпункције, бројева и осталих не-текстуалних карактера. На овај начин се врши елиминација сувишних симбола из текста и крера једноставан и чистији текст.

Нормализација

Нормализација представља корак у препроцесирању текстуалних података при чему се врши конверзија текста у јединствени тј. стандардни формат за све карактере текста. Основни процес у номализацији јесте *case* конверзија при чему се текст претвара у *lower* или *upper case.*

Токенизација

Токенизација представља корак у препроцесирању текстуалних података при чему се улазна реченица дели на листу краћих текстуалних јединица које могу бити појединачне речи, реченице или параграфи. Токени представљају смислену целину која представља заједнички контекст одређене целине токенизоване реченице. У највећем броју случајева се токенизација врши поделом улазне реченице на листу речи. Оваква листа је погодна за даље кораке у препроцесирању података.

Лематизација и стемизација

Ове технике представљају технике обраде текста које се користе за конверзију и скраћивање речи на њихов основни облик или корен речи. Технике се примењују на идентичан начин при чему се примењују различите трансформације.

Стемизација представља једноставнији процес обраде речи при чему се посматрана реч своди на корен речи изведен из посматране. Додатна битна карактеристика јесте и време извршавања самог процеса обраде података. Такође, стемизација се обавља над сваком речју појединачно без додатног познавања контекста осталих речи у околини.

Лематизација представља сложенији процес обраде текста у односу на стемизацију јер у процесу обраде података узима шири контекст и већи број речи како би креирао „лему“. Лема представља основну форму речи, тј. облик речи који представља њено лексично значење или корен. Лематизација је процес редукције речи при чему се све варијације одређене речи своде на идентичан лексички облик.

Уклањање стоп-речи

Стоп-речи представљају специфичан корпус речи сваког језика које са собом не носе високи значај у обради и анализи текста. Овакве речи је приликом обраде текста могуће извацити из скупа улазних реченица јер не доприносе квалитативним својствима посматраних текстова, док повећавају време процеса обраде текста. Ове речи се сматрају често коришћеним у језику и не доприносе контекстуалној репрезентацији улазних реченица.

Стоп-речи су углавном често коришћене речи, везници, предлози, речце и сл. Примери оваквих речи у енглеском речнику су следеће: “the”, “a”, “and”, “is”, “in”, “into”…

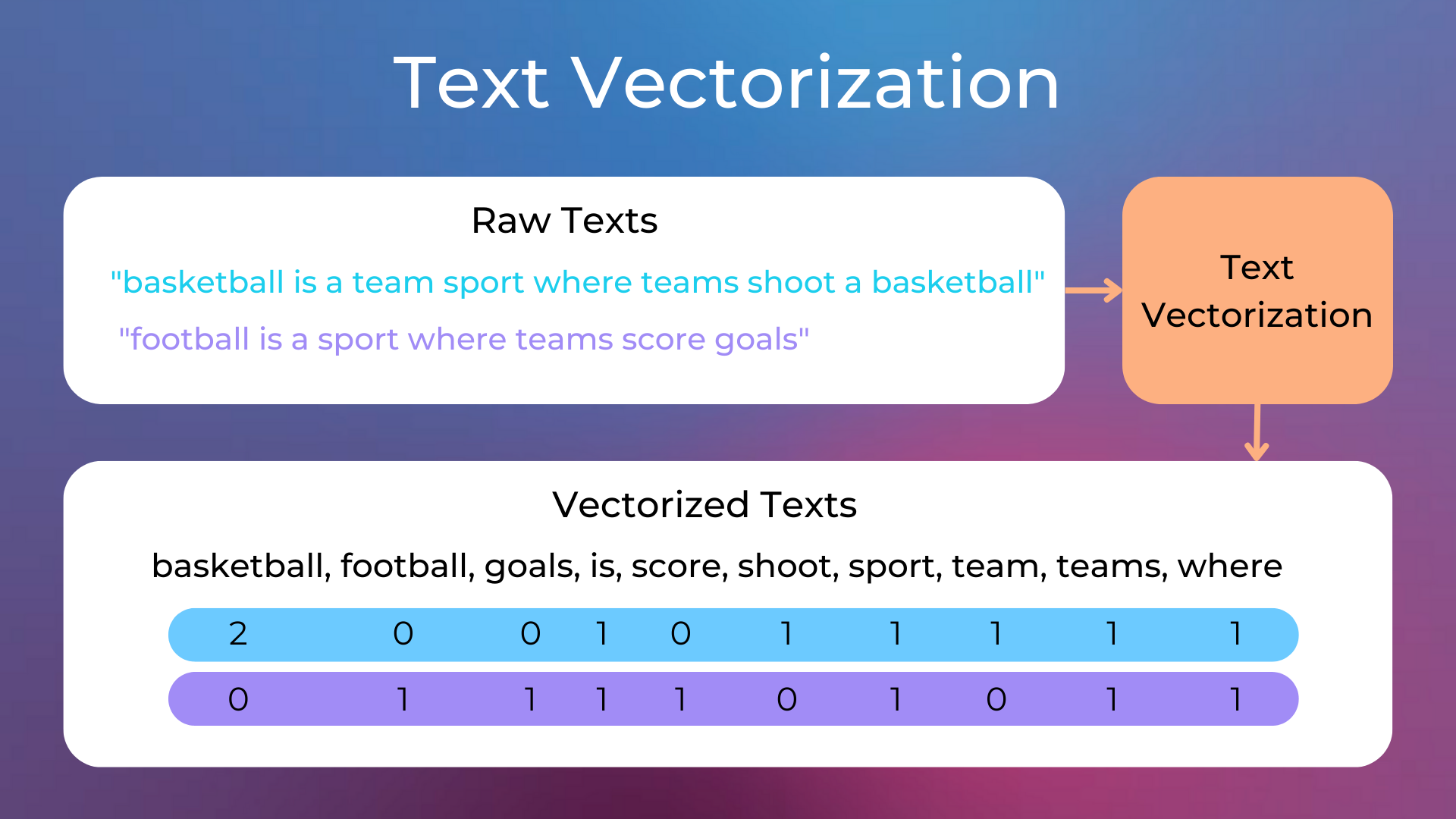
У овом одељку су наведене само неке од техника за препроцесирање улазног текста како би се прилагодио за даљу анализу и припрему података за тренирање.

### Припрема података за тренирање

Процес припреме података за тренирање представља поступак креирања статистичких параметара о скупу реченица које се обрађују. Улазни подаци овог процеса представљају препроцесирани подаци из претходног поглавља. На улазу се добијају обрађени подаци у текстуалном формату који се даље прослеђују у процесу векторизације припреме података.

#### Векторизација текстова

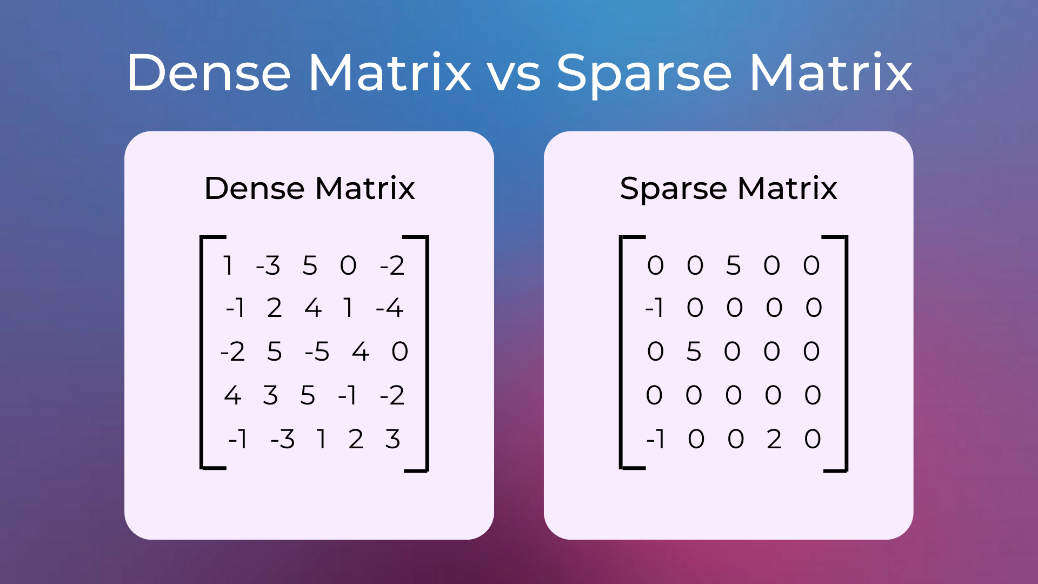
Овај процес подразумева конверзију улазних текстуалних података у нумеричке податке које може рачунар да разуме како би се даље вршило прилагођење (eng. *fitting)* . Нумерички подаци се представљају у облику векторизованих елемената матрице при чему сваки текст представља један ред матрице. Овакве матрице се називају ретко поседнутим (*eng. sparse matrix*).



Сл. 1 – Процес векторизације реченице

На слици изнад је приказан процес векторизације при чему свака реч у унији речи улазних текстова представља једну колону матрице. Сваки ред ове матрице представља појединачне улазне реченице, док вредности матрице по колонама представљају број понављања датих речи у свакој реченици понаособ. У случају када је корпус реченица изразито велики и различитог контекста, вокабулар корпуса ће бити већи што ће утицати на повећани број колона ове матрице. Уколико би текстови били различитих контекста, добили бисмо веома ретко поседнуту матрицу код које би велика већина вредности матрице имао вредност 0.

На слици испод се може видети како изгледа репрезентација ретко поседнуте матрице (*eng. sparse matrix*) у односу на густо поседнуту матрицу (*eng. dense matrix*).



Сл. 2 – Приказ репрезентације густо (лево) и ретко (десно) поседнутих матрица

Два основна начина за векторизацију текстова су:

* Bag of words
* Term frequency-Inverse document frequency

Bag of words

Bag of words (врећа речи) представља приступ формирања ретко поседнуте матрице описан у примеру претходног поглавља где се у матрици за сваку реч означава број понављања у свакој реченици појединачно. Овај начин обраде улазних текстова поседује вокабулар речи који се добија као скуп свих могућих речи из целокупног улазног скупа података. На основу вокабулара речи се затим креира мера понављања одређених речи из вокабулара.

Уз помоћ овог приступа се не могу запамтити комплексније зависности између речи у реченици попут структуре и редоследа, већ се памти само информација да ли одређена реч постоји или не у текстовима.

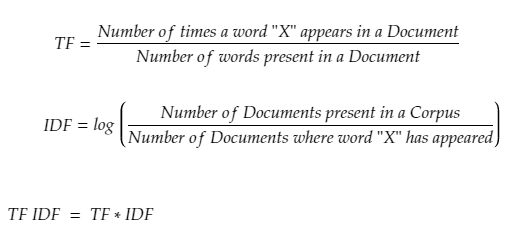
Главни проблем овог приступа је доминанта учесталост одређених речи унутар документа тј. улазног текста при чему такве речи не морају нужно да носе довољан информативни контекст. Контекст документа се заснива на речима које су специфичне домену који се обрађује, а такве речи могу имати мању учесталост у документима.

Term frequency-Inverse document frequency

Овај приступ представља унапређену верзију BOW приступа при чему се у првом делу обраде користи резултат добијен BOW методом. Термин *Еng. Term frequency* се односи на учесталост понављања токена унутар докумената и рачуна се као број понављања одређене речи подељено са бројем речи у посматраној реченици. Овај параметар предстљавља значајност дате речи у датом документу.

Термин *Eng. Inverse document frequency* се односи на учесталост појављивања датих речи у целокупном скупу података. Коначни резултат јесте производ две добијене вредности.

У следећим формулама је приказан детаљан начин израчунавања ових вредности.



TF-IDF (x, Document, Corpus)=TF (x, Document) × IDF (x, Corpus)

Коначни резултат овог израчунавања је ретко поседнута матрица са идентичним колонама и врстама као у претходном случају, по колонама су документи/реченице док су по колонама појединачне речи.

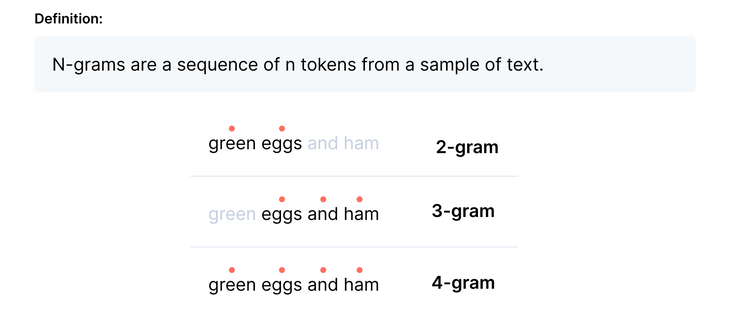
Вредност за сваку реч представља значајност речи у датом документу и у односу на целокупан корпус. Она омогућава да на тај начин дођу до изражаја речи специфичне за контекст посматраног документа.

#### Груписање речи помоћу N-грамова

Додатан напредак у домену контекстуализације докумената и хардверских захтева приликом припреме података за тренирање модела представља груписање речи у N-грамове. N-грамови представљају синтагме речи у реченици које се односе на сличан појам. У досадашњем тексту је због комплексности у обради тематике описан рад над 1-gram подацима, тј. код векторизације текстова се анализирала свака реч као засебна јединица (*eng. gram*).

Битно је напоменути да се могу креирати случајеви са 2-*gram* и 3-*gram* репрезентацијама при чему се скуп 2 или 3 речи посматра као засебна јединица у векторизацији текстова. Овакан корак у припреми података може додатно допринети контекстуализацији вектора, смањењу величине ретко поседнуте матрице и смањењу времена извршења самог процеса припреме података и тренирања.

Проблем са 2 или 3 речи у засебној јединици обраде јесте тај што се он примењује у случајевима када је корпус реченица изразито велики јер је тада учесталост таквих синтагми повећана у односу на мали корпус реченица.



Сл. 3 – Пример 2, 3 и 4-*gram* издвојених јединица за обраду

Наведене технике за обраду природног језика представљају основне кораке у припреми скупа података који се користи за даљи процес тренирања језичких модела. У случају анализе сентимената, неопходни формат је управо векторска матрица која се добија применом векторизације над препроцесираним скупом података. Скуп података припремљен на тај начин је спреман за процес тренирања. У даљем тексту извештаја у делу практичног дела је детаљније описан поступак тренирања модела.

# Практични део

Практични део овог рада се састоји од практичне имплементације концепата које се баве прибаљањем, обрадом и анализом података при чему се у завршном делу рада врши класификација скупа података са циљем добијања коначних резултата о анализи сентимената одабраног скупа података. Практична имплементација се састоји од следећих поглавља:

* Учитавање библиотека
* Учитавање и преглед основног скупа података
* Препроцесирање скупа података
* Експлоративнаа анализа података (*eng. EDA - Explanatory data analysis*)
* Припрема података за тренирање модела
* Тренирање модела

## Учитавање библиотека

За рад на практичном делу пројекта неопходно је креирати виртуелно окружење у *Python* програмском језику и инсталирати све неопходне библиотеке за рад. Библиотеке које се користе су:

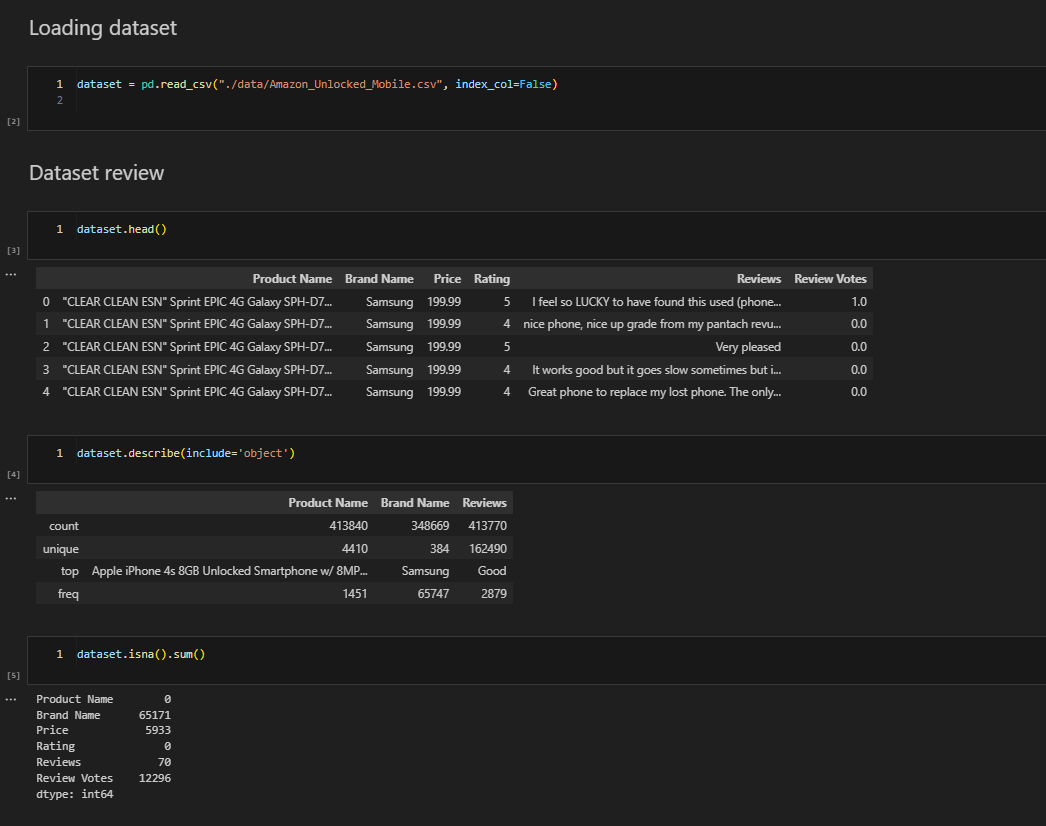
* ipykernel
* pandas
* matplotlib
* seaborn
* scikit-learn
* nltk
* transformers
* tensorflow

Наведене библиотеке је неопходо инсталирати командом pip install <назив библиотеке>, а затим све неопходне модуле ових библиотека укључити у пројекат на почетку *jupyter* радне свеске.

## Скуп података

Одабрани скуп података над којим ће се применити сви до сада обрађени концепти анализе сентимената јесте *Amazon Reviews: Unlocked Mobile Phones* који садржи више од 400 000 узорака података. Овај скуп података описује рецензије купљених откључаних телефона преко сајта *Amazon.com*. Додатне информације о скупу података се могу пронаћи на следећем линку:

<https://www.kaggle.com/datasets/PromptCloudHQ/amazon-reviews-unlocked-mobile-phones>



Сл. 4 – Учитавање и преглед скупа података

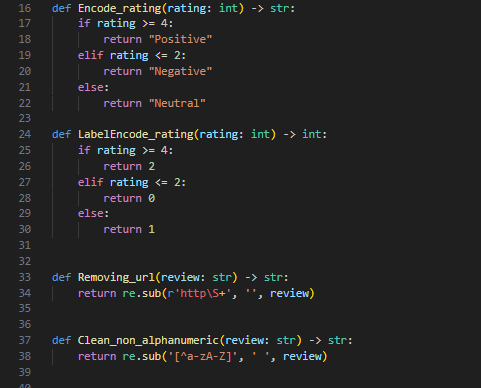
Скуп података се састоји атрибута *Product Name*, *Brand Name*, *Price*, *Rating*, *Reviews*, *Review Votes.* За процес анализе сентимената ће бити од користи атрибути *Reviews* и *Rating*. Из основног прегледа скупа података, могуће је утврдити да велики број података нема попуњене податке за одређене атрибуте и да треба обратити пажњу приликом препроцесирања и филтрирања скупа података како би се сачувао највећи број узорака скупа података који имају валидне вредности за два атрибута од значаја.

## Препроцесирање скупа података

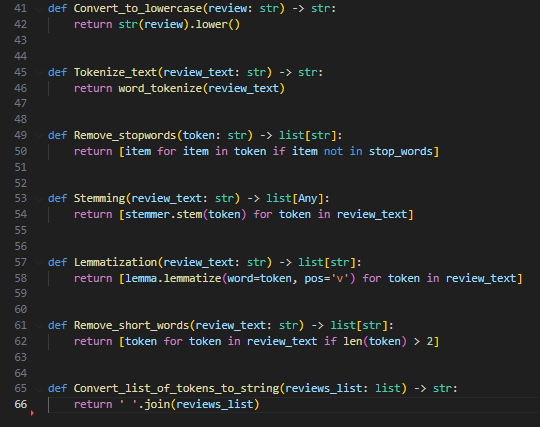
У практичном делу рада приликом препроцесирања података примењене су следеће технике:

* Издвајање значајних атрибута
* Енкодирање *Rating* атрибута
* Избацивање редова без попуњених вредности (око 70 редова)
* Избацивање *URL* тагова из улазних текстова
* Избацивање карактера који нису алфа-нумерички
* Пребацивање текстова у мала слова (*eng*. *lower case*)
* Токенизација текстова
* Избацивање стоп речи
* Лематизација (или стемизација)
* Избацивање редова са више од 2500 токена у текстуалној рецензији
* Скраћивање скупа података на мањи број погодан за тренирање модела

На следећим сликама је приказана практична имплементација појединих техника за препроцесирање текстова у формату појединачних функција које се касније учитавају и :



Сл. 5 – Први део функција за препроцесирање



Сл. 6 – Други део функција за препроцесирање

## Експлоративна анализа података

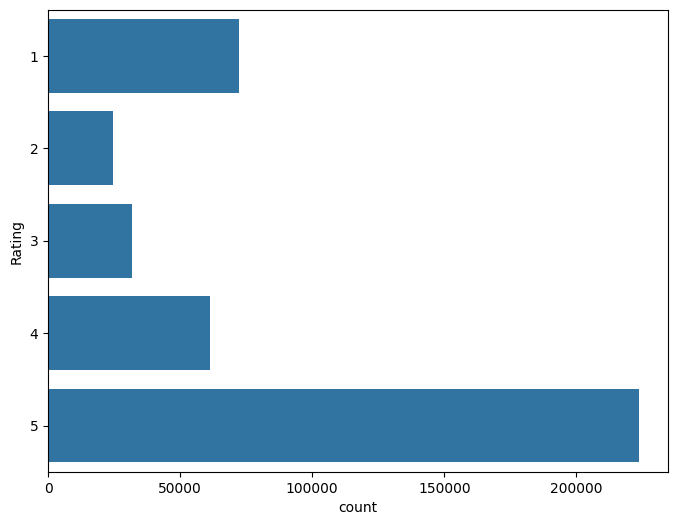
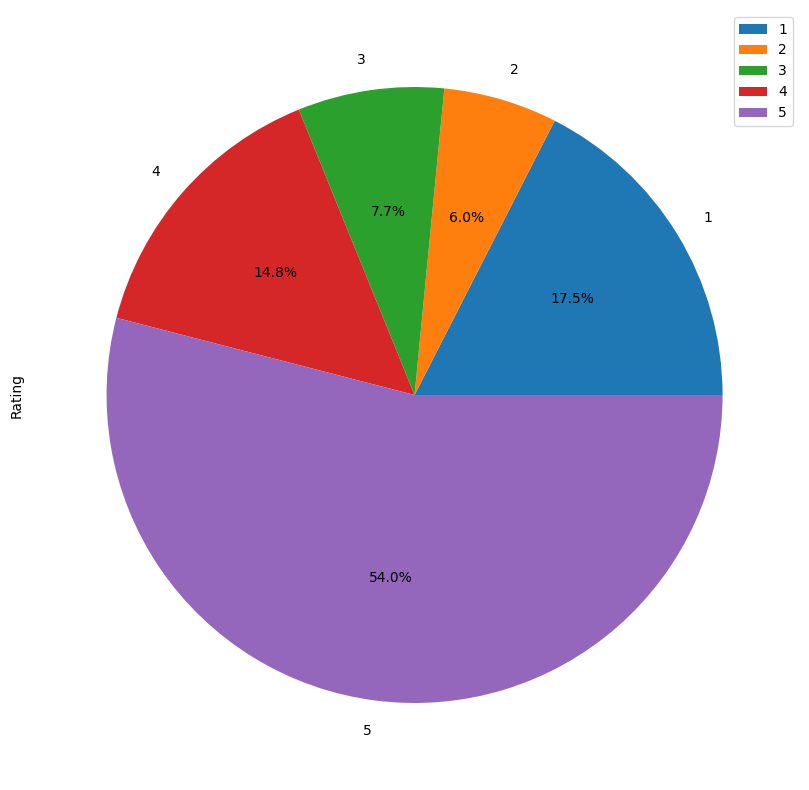
Експлоративна анализа података служи како би се путем различитим графичких алата омогућила боља визуализација података, расподела вредности по посматраним атрибутима, учесталост одређених токена у скупу података, дистрибуција дужина реченица у скупу података и сл.

Визуализација расподеле сентимената

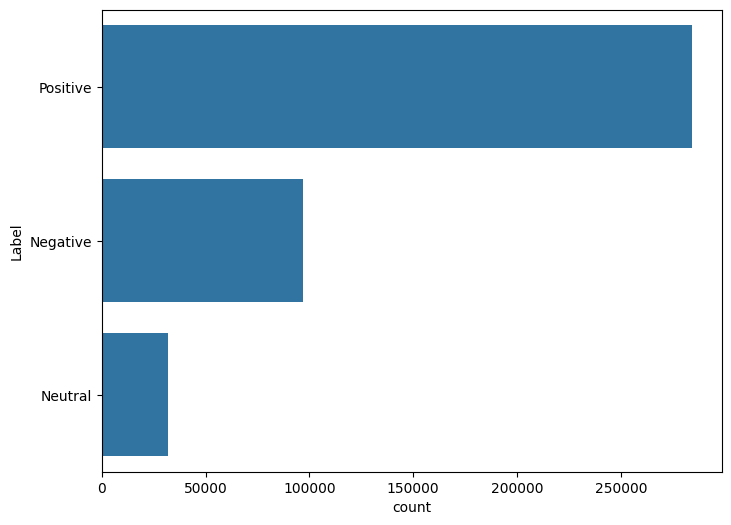
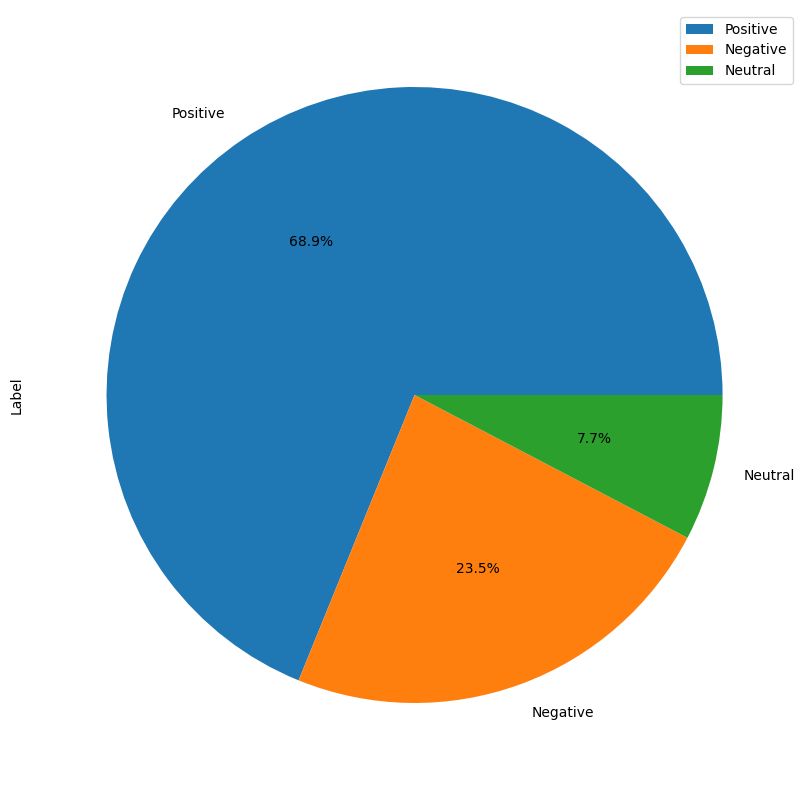
Визуализација расподеле сентимената, тј. категоричког атрибута који описује којој категорији сентимента посматрани текст припада, приказује учешће одређених категоричких атрибута у целокупном скупу података и број понављања.

Визуализација оваквих података је обављена уз помоћ *pd.Dataframe.plot.pie* и *sns.countplot* при чему је у зависности од приступа потребно прилагодити улазне податке на одређени начин.

На визуализацији је приказана расподела не енкодираних вредности *Rating* атрибута, као и енкодиранoг *Label* атрибута.



Сл. 7 – Графички приказ удела појединачних вредности *Rating* атрибута пре фазе енкодирања



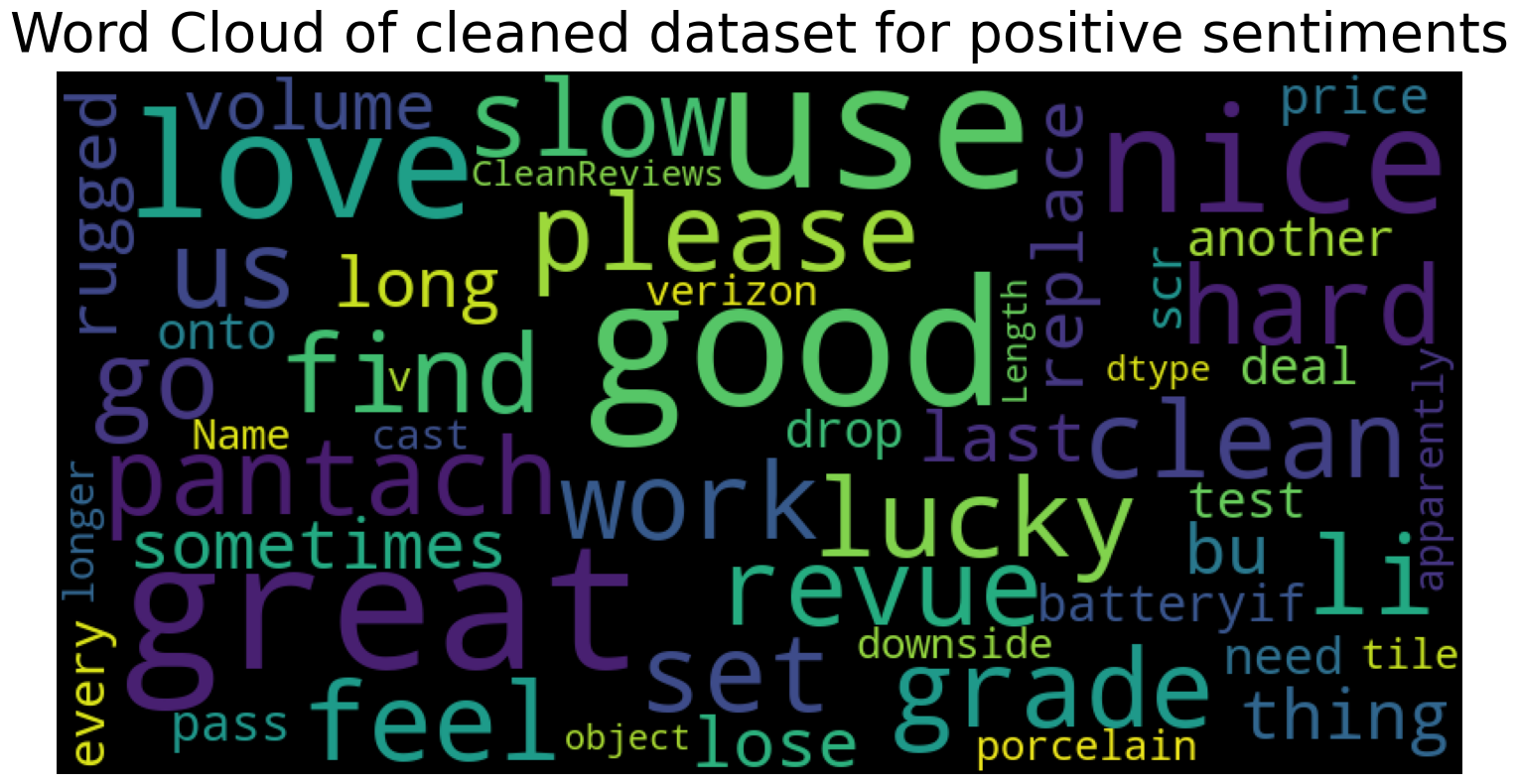
Сл. 8 – Графички приказ удела појединачних вредности *Label* атрибута након фазе енкодирања

Из овог случаја се може приметити да удео позитивних коментара има доминантну улогу у скупу података и да на овакву расподелу узорака треба обратити пажњу приликом тренирања класификатора.

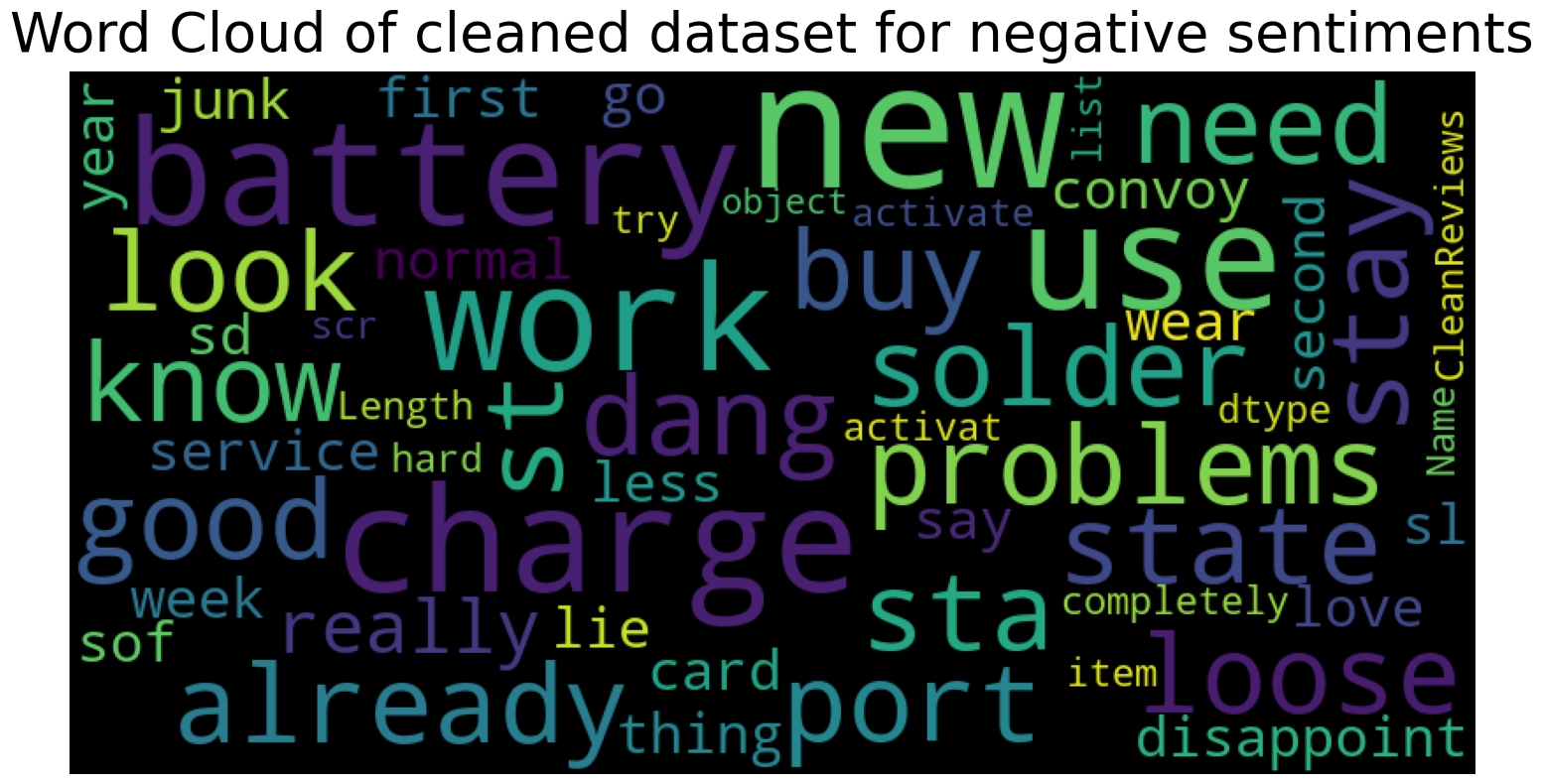
Визуализација најучесталијих речи помоћу *WordCloud*

*WordCloud* представља концепт који олакшава израчунавање и приказ најучесталијих речи из улазног скупа података. У практичном делу је најпре улазни скуп података подељен на скуп података са позитивних и негативних сентиментима.

На следећим визуализацијама су приказане најучесталије речи у оба споменута скупа података.



Сл. 9 – Речи са највећим број појављивања у позитивном подскупу корпуса

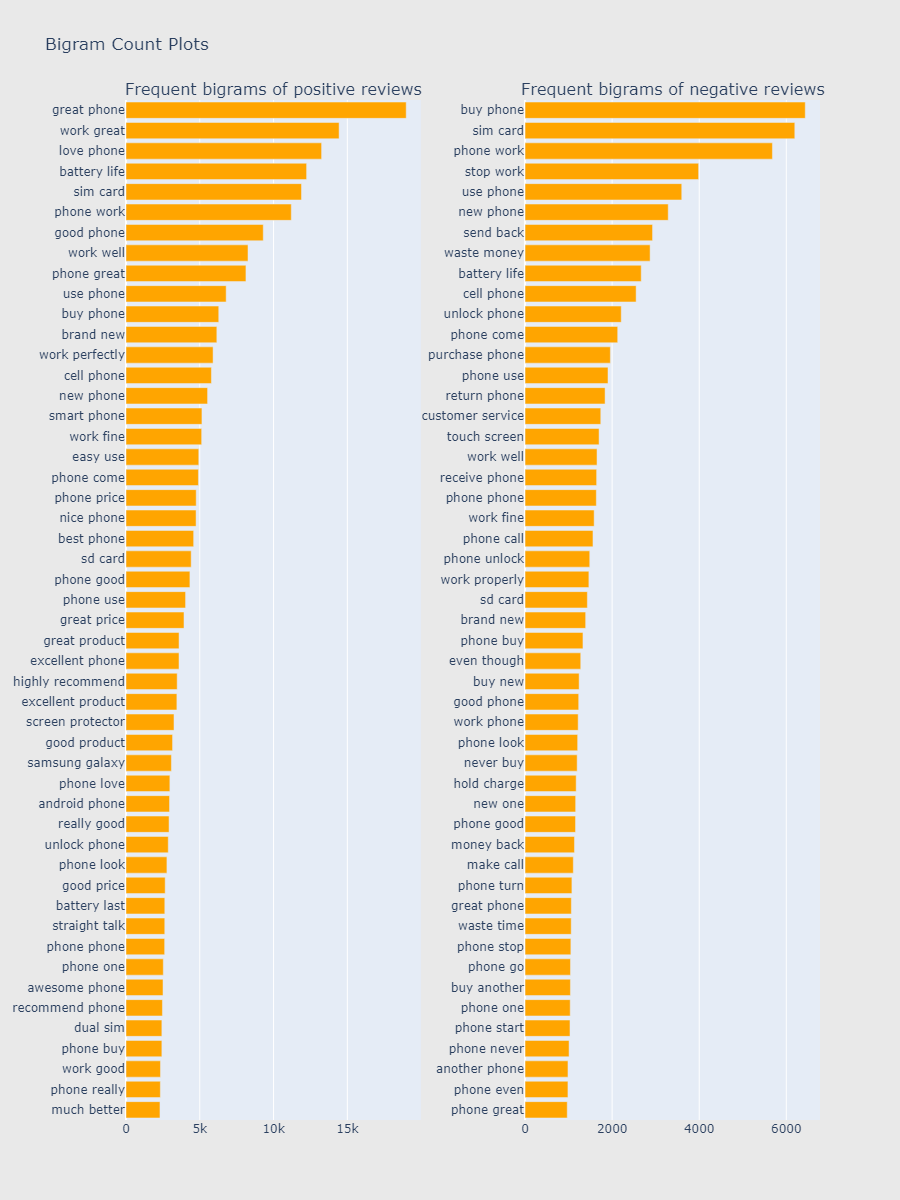


Сл. 10 – Речи са највећим број појављивања у негативном подскупу корпуса

*N-gram* визуалиција

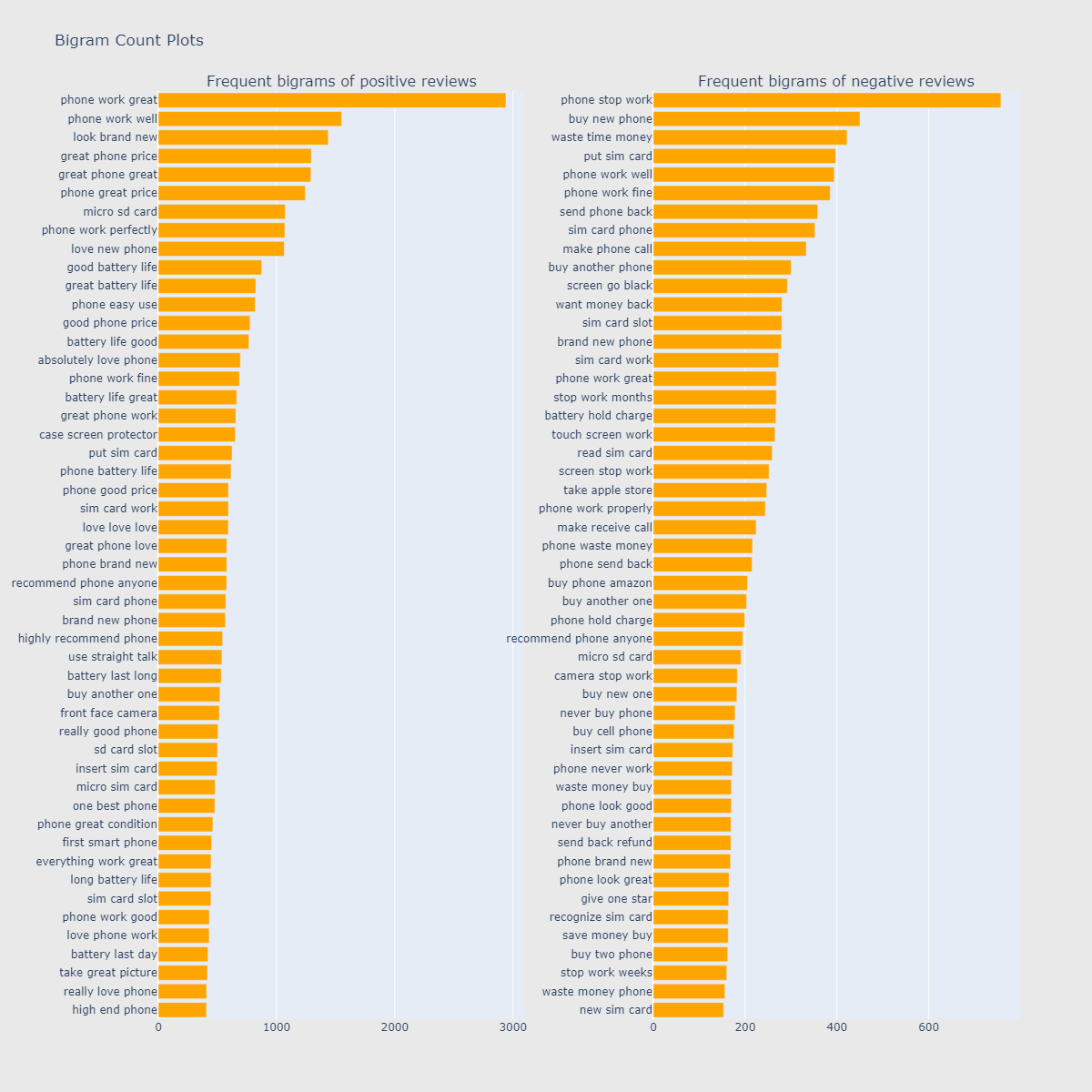
Овај тип визуалиције представља сличан начин представљања фреквенције појединачних *gram-*ова у целокупном корпусу текстова. Уз помоћ ове методе могуће је добити фреквенције вишеструких *gram*-ова (2-*gram*, 3-*gram*…). Предуслов за примену *N-gram* визуализације је претходно креиран *N-gram* скуп података.

На следећим графицима су приказани резултати овакве визуализације.



Сл. 11 – 2-*gram* узорци са највећим бројем појављивања у позитивном подскупу корпуса

На датим сликама је могуће приметити како се позитивни или негативни контекст са различитим фреквенцијама налази у датим подскупима података. Може се уочити позитиван контекст 2-*gram* и 3-*gram* узорака у позитивном подскупу (лево), као и идентична зависност код негативних случајева у негативном подскупу (десно). Из наведених резултата може се даље анализирати утицај на позитивно или негативно задовољство корисника који су остављали своја запажања и критике.



Сл. 12 – 3-*gram* узорци са највећим бројем појављивања у позитивном подскупу корпуса

## Припрема података за тренирање модела

У фази припреме података у практичном делу рада извршена је векторизација података као последњи корак у припреми података за тренирање модела и класификационе алгоритме. У самом процесу векторизације текстова коришћене су обе методе векторизације података претходно описане у теоријском делу овом извештаја.

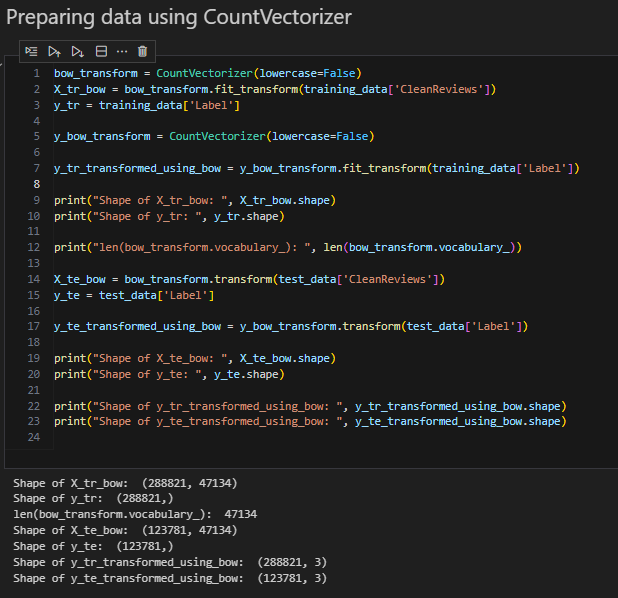
У практичном делу векторизација података је урађена за документе корпуса при чему је коришћена и 1-gram и 3-gram репрезентација текстова.

Коришћени објекти за векторизацију текстова су:

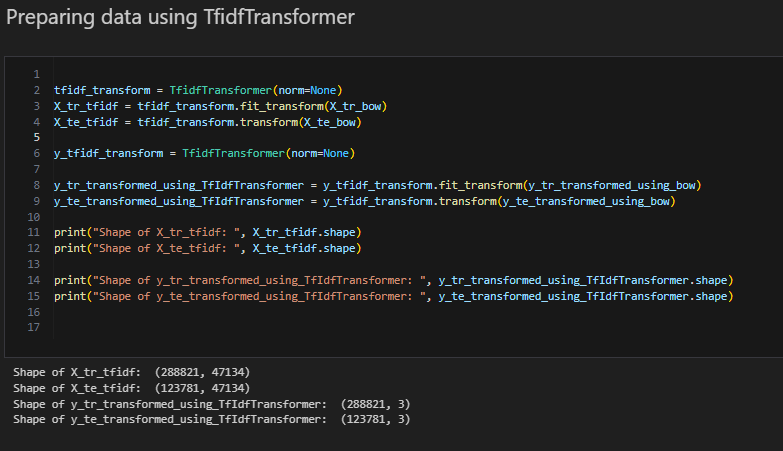
* CountVectorizer
* TfidfTransformer
* TfidfVectorizer

Векторизација текстова коришћењем 1-*gram* репрезентација

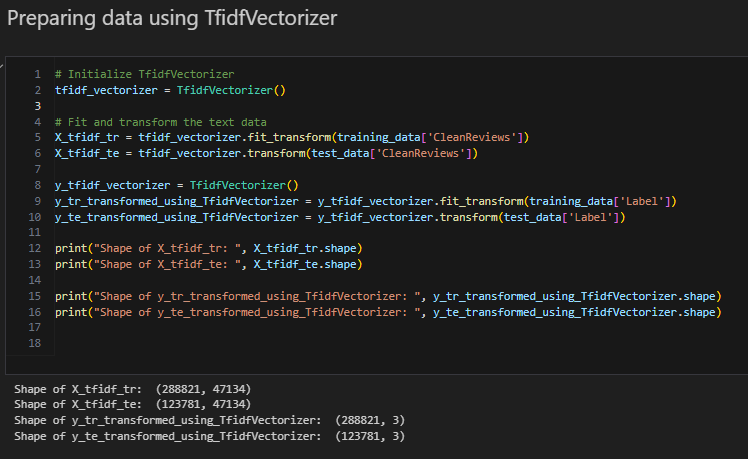
На наредним сликама је приказан начин за припрему података над 1-gram репрезентацијама.



Сл. 13 – Векторизација 1-*gram* репрезентација текстова коришћењем CountVectorizer

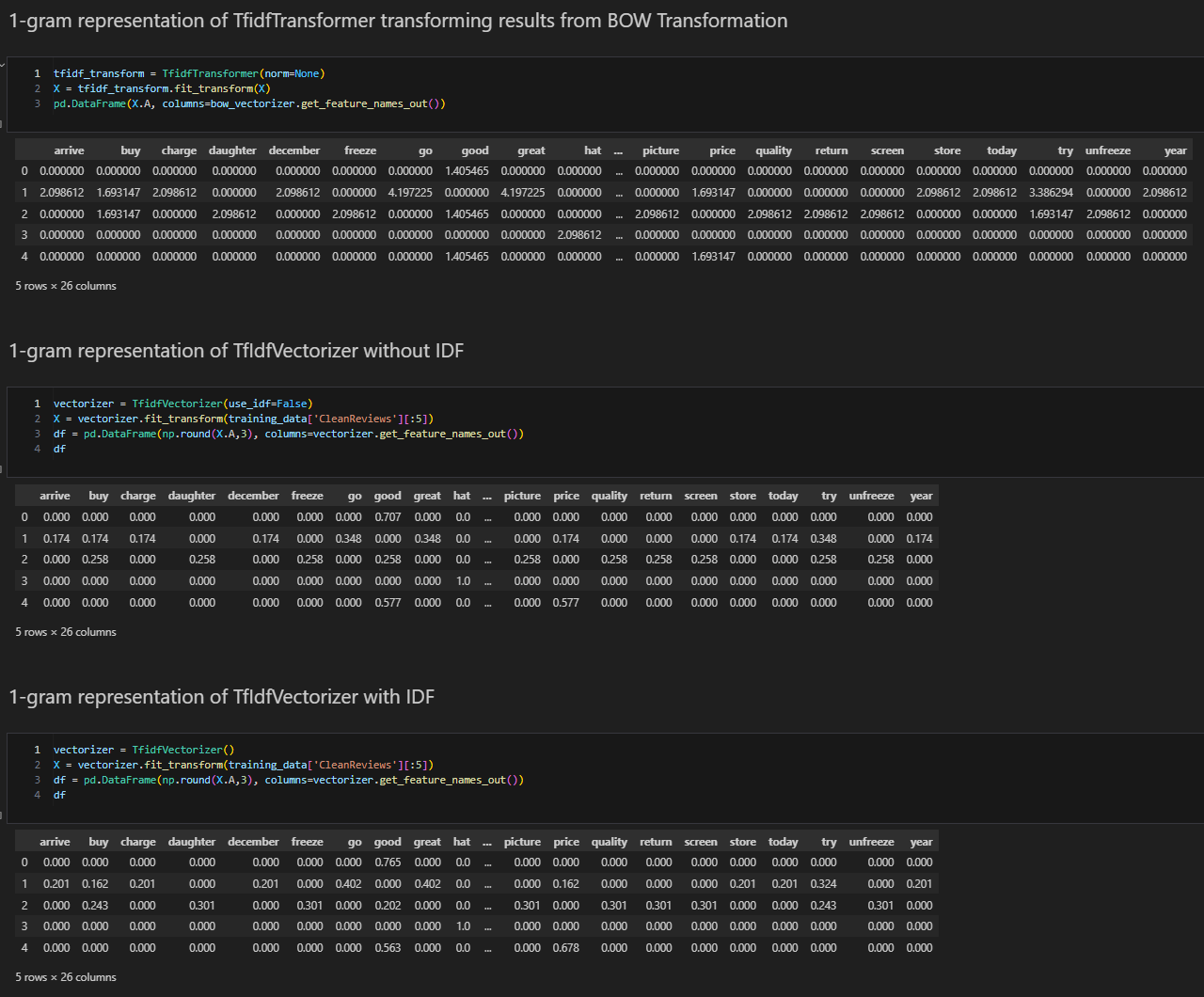


Сл. 14 – Векторизација 1-*gram* репрезентација текстова коришћењем TfidfTransformer



Сл. 15 – Векторизација 1-*gram* репрезентација текстова коришћењем TfidfVectorizer

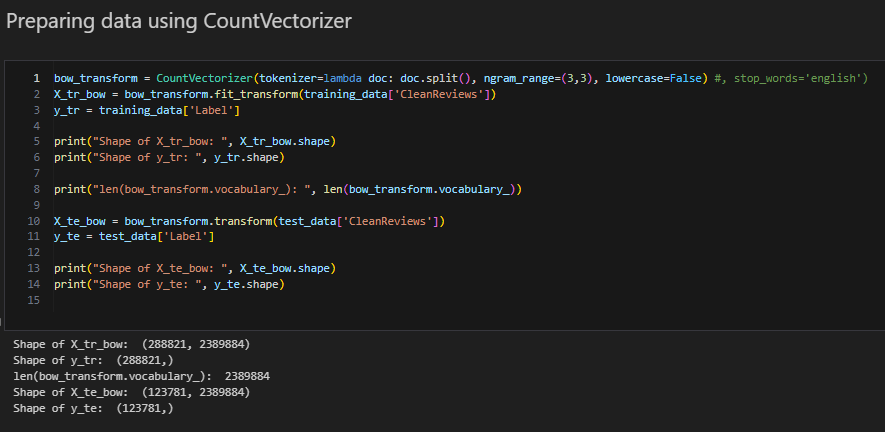
Као што је и у теоријском делу извештаја описано, применом векторизације података се добија ретко поседнута матрица. Примери таквих матрица над улазним скупом од 5 текстова су приказани на следећoј слици.



Сл. 16 – Примери ретко поседнутих матрица креираних на три различита начина

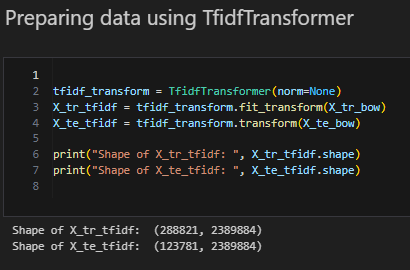
Векторизација текстова коришћењем 3-*gram* репрезентација

На наредним сликама је приказан начин за припрему података над 3-gram репрезентацијама.



Сл. 17 – Векторизација 3-*gram* репрезентација текстова коришћењем CountVectorizer

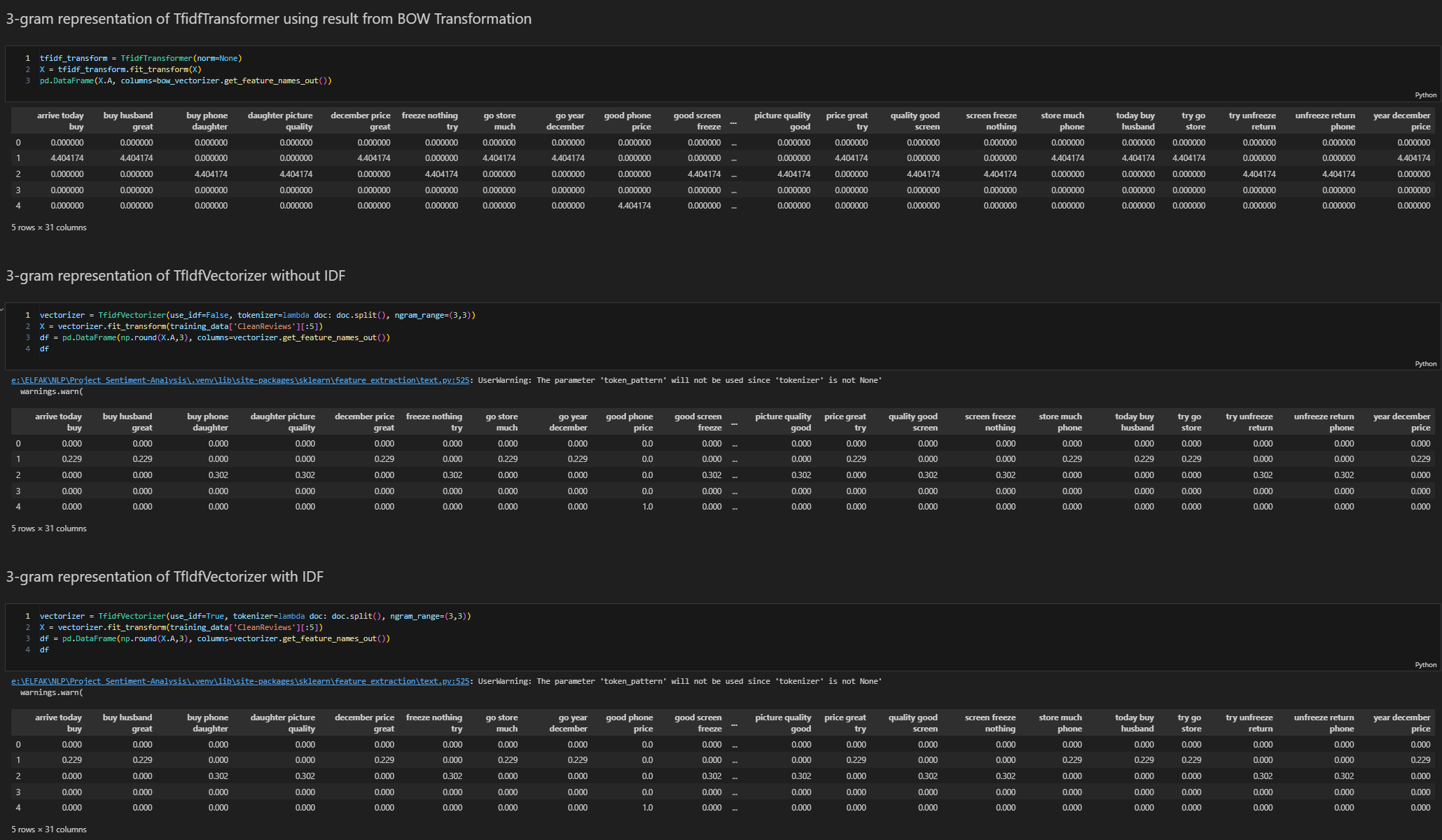
TfifdfTransformer се користи приликом трансформације резултата добијених коришћењем CountVectorizer-а, како би се добили улазни подаци у Tfidf формату.



Сл. 18 – Векторизација 3-*gram* репрезентација текстова коришћењем TfidfTransformer-a



Сл. 19 – Векторизација 3-*gram* репрезентација текстова коришћењем TfidfVectorizer-a



Сл. 20 – Примери ретко поседнутих матрица креираних на три различита начина

# Закључак

# Литература