Предикција сентимента песме користећи технике за обраду природног језика

Жарко Благојевић
Факултет техничких наука
Универзитет у Новом Саду
Трг Доситеја Обрадовића 6
21400 Нови Сад

Ана Граховац
Факултет техничких наука
Универзитет у Новом Саду
Трг Доситеја Обрадовића 6
21400 Нови Сад

Тара Поганчев
Факултет техничких наука
Универзитет у Новом Саду
Трг Доситеја Обрадовића 6
21400 Нови Сад

***Апстракт* — Категоризација музичких нумера чест је проблем са којим се сусрећу разнолики системи за препоруку музике. Сентимент песме, као један од кључних параметара није увек лак за предвидети, иако он представља примарни параметар када препоруку песме желимо да извршимо на основу корисниковог расположења.**

У раду су примењене и тестиране вишеструке технике за обраду природног језика. Текст је векторизован помоћу рекурентних неуронских мрежа, Word2Vec, GloVe и BERT модела са циљем тестирања која представа текста (енгл. *embedding*) ради најбоље за класификацију песама по сентименту. Поред векторизованог текста песмама, коришћене су и аудио карактеристике песама (табеларни подаци) са циљем пружања додатног контекста датим песмама и побољшања резултата класификације.

Кључне речи — музика, сентимент, NLP, дубоко учење, класификација, BERT, RNN

I. Увод

Системи за препоруку музике све су шире распрострањени на дигиталном тржишту. Сентимент песме уско се повезује са тренутним расположењем корисника, те представља један од кључних параметара на основу којег се препорука може вршити.

Познавање сентимента песме није корисно само у препорукама музике, већ и у природном спајању звучно-сличних песама, које ће се без обзира на извођача или жанр на основу свог „расположења“ добро уклопити заједно. Овакво груписање већ видимо у водећим апликацијама за стриминг музике, као што су Spotify, Deezer, YouTube Music, и сл.

Задатак који је у оквиру рада истражен јесте предикција сентимента песама на основу аудио карактеристика извучених помоћу Spotify API, као и њиховог текста преузетог са Genius сајта за музичке текстове. Како постоји више група сентимената (укупно 5) задатак представља мултикласну класификацију.

Како бисмо дошли до што бољих резултата, у склопу рада искористили смо више приступа за обраду текста песме користећи NLP технике за претварање текста у вектор. Неке од технологија које ће детаљније бити описане у наставку рада јесу LSTM, GRU, BERT, ELMO, Word2Vec... Овако кодирани текстови песама прослеђени су, заједно са аудио карактеристикама, неуронској мрежи чији је задатак био да се обучи над подацима и врши предикцију над новим песмама.

Пред крај самог рада упоредићемо различите приступе, те приказати који су и због чега дали најбоље резултате на крају.

II. Сродна истраживања

A. *"Text-based sentiment analysis and music emotion recognition"*

У раду "*Text-based sentiment analysis and music emotion recognition*"[1], аутор је темељно и систематично описао више аспеката музичких препорука базираних на емоцији коју песме носе. Зашао је дубље у анализу емоција и сентимента, као и моделе њиховог груписања у сродне кластере. У раду су описани различити приступи креирања система препорука који прати расположење слушаоца (*Mood-Aware Music Recommenders*). Такође су истражени начини репрезентације речи, од *Bag-Of-Words* приступа, па до различитих word vector-а, укључујући *CBOW, Skip-Gram* и *Glove*. Показано је да тачност у анализи сентимента коришћењем различитих претренираних модела зависи од природе текста који класификујемо и да се за класификацију текстова песама најбоље показао *GloVe* модел трениран на *twitter* корпусом.

B. *"Lyric document embeddings for music tagging"*

Рад "*Lyric document embeddings for music tagging*"[2] се такође бави проблематиком одређивања сентимента песама на основу њиховог текста. Овде су употребљени различити приступи добијања вектора речи - прво *baseline* модели: *bag-of-words* и *TF-IDF*, а затим следећи приступи *word2vec* модела: трениран на корпусу који се састојао од текстова музичких нумера, претренирани *google-300* модел и на крају *warm-start*, односно *google-300* модел дотрениран текстовима из корпуса. Циљ је био да се изабере најбољи од набројаних модела и искористи за добијање document вектора узимањем средње вредности вектора речи, и као *embedding* слој за *LSTM* неуронску мрежу. Трећи приступ који је имплементиран је тренирањем *doc2vec* модела над корпусом. Као најбољи се показао *word2vec* модел трениран над корпусом и димензионалности 512. Изненађујуће је да је он давао боље резултате чак и од *LSTM* модела, и претпоставка аутора је да је то из разлога што је *word2vec* модел трениран над веома великим корпусом, те је успео да направи веома добре репрезентације речи.

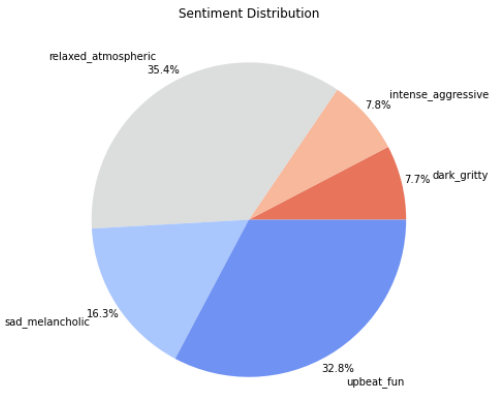
III. Опис скупа података

У овом поглављу биће описани скупови података који су коришћени, и укратко представљене трансформације које су над њима извршене како бисмо дошли до коначног скупа података.

Са сајта „*Kaggle*“ преузета су два скупа која су поседовала податке о музичким нумерама. Први, „*MuSe: The Musical Sentiment Dataset*’’ је био кључан за нас јер је у себи садржао колону *seeds*, која представља информације о сентименту песме - за нас циљно обележје. Поред тога, овај скуп података садржи и име песме и извођача, жанр, три нумеричка обележја која описују осећања која одређена песма изазива (*valence*, *arousal* и *dominance* - редом: пријатност изазваних емоција, њихов интензитет и ниво контроле над њима) и три колоне које садрже идентификациона обележја, међу којима је за нас значајан *spotify_id*.

Анализом скупа података уочили смо да он садржи велики број поновљених песама које се разликују по скупу тагова у колони *seeds*. На пример, песма „*Everyone's A V.I.P To Someone*’’ се појавила четири пута са следећим вредностима: [*'uplifting'*], [*'driving'*], [*'soothing', 'soft', 'uplifting', 'nostalgic'*], [*'driving', nostalgic'*]. Да бисмо добили скуп података без дуплираних редова, а да не бисмо изгубили податке садржане у различитим скуповима тагова, спојили смо вредности из *seeds* колоне и избацили поновљене тагове у оквиру исте песме. Након тога смо могли слободно да одбацимо ,,дупликате’’ песама без бојазни од губитка информација и тиме свели скуп са почетних ~90 хиљада редова на троструко мањи скуп од ~30 хиљада редова.

Други коришћени скуп података увели смо с намером да увећамо број информација о музичким нумерама. ,,*Spotify Tracks DB*’’ представио је музичке нумере помоћу још 18 колона које су највећим делом описивале аудио карактеристике песама, али и друге податке, попут популарности песме. Слично као у претходном скупу података, и овде смо имали проблем поновљених песама које су се овога пута разликовале по жанру. Како број жанрова није био нарочито велик, овај проблем је решен трансформацијом колоне жанр у посебну колону за сваки појединачни жанр из скупа података, попуњену вредностима 0 и 1. Последично, број редова је смањен са ~233 хиљаде на ~176 хиљада. Категорички подаци попут поменутог жанра, мода (дур, мол), и такта трансформисани су помоћу *OneHotEncoding* методе, док су аудио карактеристике песама (нумерички подаци) адекватно скалирани.



Слика 1 – Количнички однос песама из сваког сентимента

Како ниједан од поменутих скупова података није садржао текстове песама, њихово прикупљање извршено је веб скрејповањем светски популарног сајта за текстове песама *genius.com*. Скрипта за прикупљање текстова песама је на основу имена извођача и назива песме, ослањајући се на Гугл претраживач, пронашла линкове до страница текстова датих песама на циљаном сајту. Затим су текстови песама извучени уз помоћ библиотеке *BeautifulSoup*. Прикупљање текста било је успешно за око 5500 песама (од 6500 из претходно спојеног скупа података), па су остале песме одбачене због немогућности да над њима примењујемо моделе процесирања природног језика.

Последњи корак у припреми података је да на основу већ описаних тагова у *seeds* колони одредимо сентимент коме песма припада. Иако тагови сами по себи представљају сентимент, њихов број је превелики да би се класификација могла успешно извршити над релативно малим скупом података који поседујемо. Стога смо одлучили да ове сентименте групишемо у 5 већих група: *Upbeat/Fun, Intense/Aggressive, Dark/Gritty, Relaxed/Atmospheric* и *Sad/Melancholic*, где је свакој песми додељена група из које су тагови који преовлађују. Заступљеност песама које носе одређени сентимент приказана је на слици 1. Ова слика указује да је финални скуп података небалансиран, што значајно утиче на сам процес тренинга модела, али и избор адекватних метрика за евалуацију модела.

IV. Методологија

Након што су подаци о песама прикупљени и обрађени, доводе се на алгоритме дубоког учења са циљем класификације сентимента. Како постоји више од два сентимента који нису подједнако заступљене у скупу података, овај проблем представља класификацију на више класа над небалансираним подацима.

Модел приказан на слици *2* има две врсте улаза, табеларне податке о карактеристикама песама и текстове песама. Превођење текста различите дужине из људски читљивог облика у вектор фиксне величине, тиме омогућујући његово довођење на неуронску мрежу, извршено је на неколико приступа:

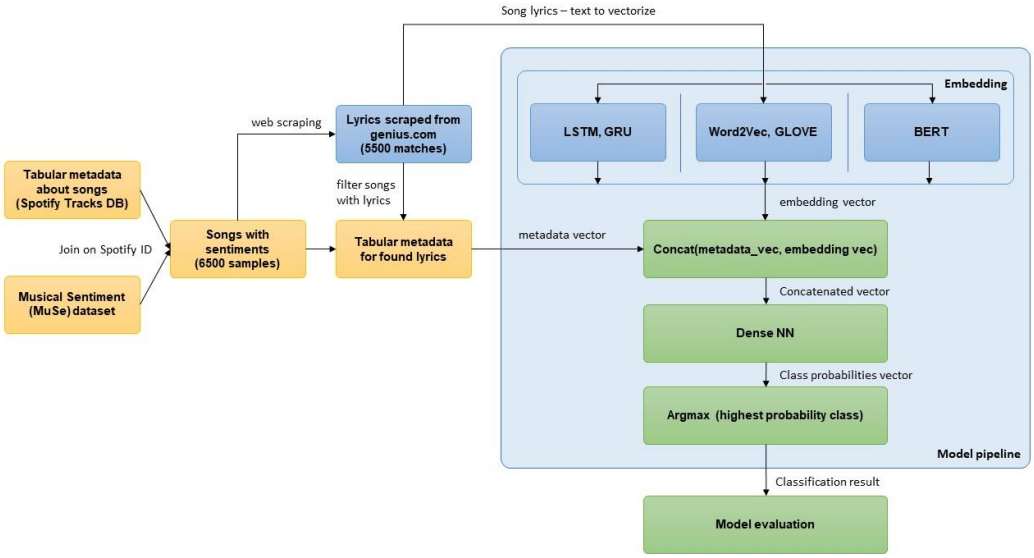
1) Рекурентном неуронском мрежом (*LSTM/GRU* ћелија) са вокабуларом креираним на основу скупа података

2) Усредњавањем *Word2Vec/GloVe* вектора речи из текста како би се добио јединствени вектор за цео текст

3) Коришћењем *BERT* модела (*Bidirectional Encoder Representations from Transformers*) – претренирани трансформер који кодира текст песме у вектор

Вектор добијен од текста песме се затим спаја са вектором карактеристике песама (табеларни подаци), са циљем давања додатног контекста тексту и побољшавањем резултата класификације. Затим се новонастали вектор доводи на потпуно повезану неуронску мрежу која врши класификацију. Излазни слој овог модела садржи 5 неурона (јер је присутно 5 класа) са *softmax* активационом функцијом и враћа вектор вероватноћа припадности одређеној класи. Као резултат класификације узима се класа са највећом вероватноћом.

Скуп података подељен је на три дела: тренинг (80%), валидациони (10%) и тест (10%) скуп. Валидациони и тест скуп су стратификовани (задржана је заступљеност



Слика 2 - Шема модела

класа из тренинг скупа), а тренинг скуп узоркован са повећањем броја узорака мање заступљених класа (енгл. oversampling). Са циљем да се спречи претренирање (енгл. *overfitting*) имплементирано је рано заустављање (енгл. *early stopping*) које прати вредност функције грешке на валидационом скупу и зауставља тренирање уколико се грешка не смањи након три везане епохе тренирања.

Након прегледа модела на вишем нивоу апстракције позабавићемо се приступима за векторизацију текста. Текстови су векторизовани помоћу следећих модела:

1) Рекурентне неуронске мреже (LSTM/GRU), Рекурентне неуронске мреже (РНН, енг. *Recurrent neural networks*) представљају посебну врсту неуронских мрежа дизајнирану за процесирање секвенцијалних података попут текста или временских серија. Текст се у РНН убацује реч по реч, тако да на кодирање сваке нове речи утичу речи које јој претходе. Ово им омогућава да открију зависности између различитих делова секвенце и ухвате контекст у којем се извесне речи јављају. Самим тим репрезентација тих секвенци постаје знатно богатија. Мана им је што не умеју лепо да репрезентују веома дугачке секвенце, те се дужи текстови пре прослеђивања на неуронску мрежу често сумаризују или деле на параграфе.

Long Short-Term Memory (*LSTM*) је тип РНН-а који може помоћи у кодирању текстова због његове способности да пронађе дугорочне зависности у ово постиже тако што ажурирање скривених слојева замењује комплекснијим меморијским ћелијама.

Други тип РНН-а који се често користи је *Gated Recurrent Unit (GRU*). *GRU* је сличан *LSTM*-у по томе што такође има могућност складиштења и приступа информацијама кроз време, али има једноставнију меморијску ћелију са мање параметара. Ово га чини лакшим за коришћење и бржим за обуку, али често и не толико ефикасним у хватању сложених зависности у секвенцама.

Један од имплементираних приступа за класификацију сентимента песама користи рекурентне неуронске мреже са циљем богате репрезентације текста. За векторизацију текста неопходно је прво пречистити текст од скраћеница, знакова интерпункције и честих речи (енгл. *stop words*). Ово омогућава изградњу мањег и квалитетнијег вокабулара који *LSTM* и *GRU* користе при тренирању. Слој са 128 меморијских ћелија и *Dropout* слојем од 0.2 дали дао је најбоље резултате при класификацији и за *LSTM* и *GRU*, с тим да се *LSTM* показао мало боље.

2) BERT – Bidirectional Encoder Representations from Transformers, један од најактуелнијих модела природног језика данашњице јесте BERT. Заснива се на технологији трансформера, и његова главна улога јесте да рачунарима помогне у разумевању текста уз импликацију контекста у коме се он налази.

Овакав приступ био је идеалан за проблем представљања текста песме у неуронској мрежи, како се сигурније можемо ослонити валидност на контекста артистичне природе текстова.

Пре самог генерисања вектора, било је неопходно обрадити текстове на начин који ће претрениране моделу BERT-а бити једноставније за рад. Ограниченост на 512 токена/речи захтевала је да покушамо да скратимо текст који му прослеђујемо, али да и даље задржимо битне информације.

Примењене промене су подразумевале:

а) Отклањање помоћних тагова за означавање строфе, рефрена, инструментала и слично,

b) Брисање понављајућих стихова, тако да се уникатни стихови и даље налазе у иницијалном редоследу

c) Додавање текста налик „Ово је текст песме ... од извођача ...“ пре почетка самог текста, како би модел имао контекст материјала са којим ради.

Овако обрађени текстови спојени су са извученим и адаптираним табеларним подацима, и прослеђени неуронској мрежи.

3) ELMo - Embeddings from Language Models – као и BERT, овај модел даје нам могућност да текст представимо у облику вектора чувајући његов контекст.

Иста реч може имати различита значења у зависности од реченице у којој се налази. Овај бидирекциони модел заснива се вишеслојној LSTM мрежи која у обзир узима и леви и десни контекст речи коју ембедује.

Иницијална идеја приликом овог истраживања јесте да исте текстове ембедујемо користећи ELMo и BERT, те да на крају упоредимо резултате који су дали. Међутим, током имплементације дошли смо до проблема временског извршавања ELMo модела, као и презахтевности његове потребе за радним ресурсима.

Природа овог модела чинила је рад са дугачким текстовима песама (обрађених на исти начин описан у секцији IV. 1) веома захтевним и спорим, те нисмо били у прилици да скуп од оквирно 5000 песама преточимо у векторе.

Још један начин на који би ово негативно утицало на рад јесте временска нефлексибилност по питању мењања процеса. Мале промене у методологији обраде текстова захтевале би поновно ембедовање истих које би трајало више дана.

Због непогодности рада са ELMo моделом, и недовољних ресурса, одлучили смо да га изоставимо из крајњих инкремента пројекта. У његову замену, BERT ће остати као представник комплексних претренираних модела за ембединг текста одржавајући контекст истог.

4) Word2Vec i GloVe - Један од приступа који је коришћен да би се добио *document level embedding* јесте помоћу *Word2Vec* и *GloVe* модела.

Word2Vec покушава да разуме речи на основу контекста у којем се речи јављају, гледајући речи које се јављају заједно у реченицама (локални контекст).

GloVe покушава да разуме речи на основу тога колико често се речи јављају заједно у великом корпусу, креирајући глобалну матрицу заједничког појављивања (енгл. *co-ocurrences matrix*)

Како је наш скуп података релативно мали, а садржај текстова песама није доменски специфичан текст, одлучили смо се за коришћење претренираних модела из gensim библиотеке. Изабрали смо Word2Vec модел који је трениран над корпусом од око 100 милијарди речи из чланака Google News-a и садржи 300-дименензионалне векторе за око 3 милиона речи на енглеском језику (word2vec-google-news-300). Од *GloVe* модела коришћени су 2 модела. Први модел је тениран над твитовима и димензионалности 200 (glove-twitter-200). Добра ствар овог модела јесте што је трениран над корпусом који садржи речи из сленга које се релативно често јављају и у текстовима песама. Међутим, како су твитови, уопштено говорећи, доста краћи него текстови песама, могуће је и да овај модел неће бити најбољи избор за наш случај коришћења. Из тог разлога, ради поређења, узет и модел трениран над текстовима са Википедије и Gigaword-a и димензионалности 300 (glove-wiki-gigaword-300).

Текст је очишћен на следећи начин: пребачен је у мала слова, уклоњени су знакови интерпункције, честе скраћенице у енглеском језику где су коришћени апострофи замењене су својом пуном верзијом (i'm ->i am), и уклоњени су метаподаци из текста који су се налазили унутар угластих заграда *([Verse 1: Eminem])*. Даље је овај текст токенизован, односно претворен у низ појединачних речи. *Document vector* за сваки текст песме добили смо рачунањем средње вредности вектора речи који нам је вратио неки од наших претренираних модела. Дата је предност средњој вредности над сумом вектора због велике разлике у дужини текстова различитих песама. Последично, добили смо по колону за сваки од изабраних претренираних модела која садржи 300-димензионални, односно 200-димензионални вектор за сваку песму.

V. Резултати

Финална евалуација је извршена над тест скупом где су за поређење искоришћене макро усредњене (све класе подједнако битне) вредности прецизности, одзива и F1 мере. Због небалансираности података тачност није узета у обзир. Резултати су приказани у табели 1.

Model Name	Macro AVG precision	Macro AVG recall	Macro AVG F1 Score
LSTM	0.58	0.52	0.54
GRU	0.55	0.52	0.53
word2vec-google-news	0.68	0.44	0.42
glove-wiki-gigaword-300	0.63	0.46	0.49
glove-twitter-200	0.66	0.4	0.4
BERT	0.65	0.41	0.42

Табела 1 - Резултати

Сваки модел искоришћен за векторизацију има одговарајући потпуно повезани слој за класификацију, који је морао бити прилагођен због различитих димензија произведених вектора, а који није наведен због прегледности табеле.

Рекурентне неуронске мреже имале су лошије резултате од осталих приступа уз претпоставку да нису имале довољно примерака да боље генерализују сентимент из текста који је претежно у колоквијалан уз доста сленга. РНН би потенцијално дале боље резултате да вокабулар није грађен од недовољно великог скупа података текстова песама који је био доступан, већ да је преузет из колоквијалних *word2vec-google-news* или *glove-twitter* модела тренираних над много већим корпусом.

Модел са највећом прецизношћу по класама (0.68) користи *word2vec-google-news* модел за векторизацију тиме показујући да је успео да направи најбогатије репрезентације текстова песама.

Модел који користи BERT за векторизацију се по метрикама налази између рекурентних неуронских мрежа и *Word2Vec/GLOVE*. Резултати са овим моћним енкодером потенцијално би се побољшали када би се дотренирао (енгл. *fine-tuning*) над колоквијалним текстовима песама.

VI. Закључак

У овом раду приказани су различити приступи класификације сентимента песме на основу текста и аудио карактеристика песме. Ови подаци су прикупљени из различитих извора, тачније *kaggle.com* на којем су пронађени велики број тагова сентимената и аудио карактеристике песама и *genius.com* са којег су прикупљени текстове песама. Над тако прикупљеним подацима је извршена експлоративна анализа, а затим су обрађени и пречишћени. Тагови сентимената организовани су у 5 група по сличности, тако да представљају циљно обележје за класификацију.

Скуп података подељен је на три дела: тренинг, валидациони и тест скуп, са очувањем заступљености класа у валидационом и тест скупу. Како је креиран скуп података био небалансиран, за евалуација модела коришћени су макро усредњене метрике прецизности, одзива и F1 метрике.

Резултати добијени употребом *word2vec-google-news* кодираних текстова песама показали су се најбољи по прецизности (0.68) док је најбољи F1 скор имао модел са *LSTM* кодираним текстовима (0.54).

Даљи развоја решења ишао би у правцу прикупљања већег скупа података на основу кога би модели могли боље да генерализују, као и комбинације приступа описаних у овом раду (*LSTM* који користи *glove-twitter* колоквијални вокабулар), или дотренирања BERT модела над текстовима песама како би вектори још прецизније репрезентовали контекст и сентимент текста. Још један потенцијални правац развоја решења био би у тестирању различитих хеуристика за креирање циљног обележја од тагова сентимента, што би потенцијално довело до бољих резултата.

Литература

[1] Çano, Erion. "Text-based sentiment analysis and music emotion recognition." arXiv preprint arXiv:1810.03031 (2018).

[2] McVicar, Matt, et al. "Lyric document embeddings for music tagging." arXiv preprint arXiv:2112.11436 (2021).