Предлог пројекта из СИАП-а

Овај документ садржи кратак опис онога што је тема пројекта и дефиниција, мотивација за одабрану тему. Након мотивације следи преглед владајућих ставова и схватања у литератури, затим скуп података који је укратко описан. Такође је наведен и софтвер који ће бити коришћен, као и метод евалуације. На самом крају документа налази се план рада на пројекту.

Тема пројекта је *Предвиђање популарности огласа за посао на Linkedin друштвеној мрежи*.

Дефиниција пројекта

Циљ пројекта је анализа и предикција популарности огласа за посао на Linkedin друштвеној мрежи. Методологија обухвата анализу карактеристика огласа и профила корисника који их постављају. Параметри попут броја прегледа и броја пријављених на оглас биће анализирани. Идеја је утврдити у којој мери ови фактори утичу на популарност огласа и користећи моделе машинског учења предвидети потенцијалну трајекторију огласа који имају потенцијал да буду популарни.

Мотивација

1. Ефикасније оглашавање послова: Разумевањем фактора који доприносе популарности огласа, послодавци могу прилагодити своје огласе како би привукли већи број квалификованих кандидата.
2. Оптимизација Linkedin профила: Кандидати могу користити информације о популарности огласа како би побољшали своје Linkedin профиле и прилагодили их оним вештинама и искуствима које послодавци траже.
3. Стратегије рекрутовања: Рекрутерске службе могу користити анализу популарности огласа како би оптимизовале стратегије за привлачење кандидата.

Преглед владајућих ставова и схватања у литератури

[1] „*Popularity Prediction of Instagram Posts*“

Линк: https://www.mdpi.com/2078-2489/11/9/453

У овом раду анализирана је популарност будуће објаве на Инстаграм друштвеној мрежи. Аутори истражују ову тему као кључан изазов у друштвеним медијима, са фокусом на постизање тачног класификовања популарности објава. Рад користи Gradient boosting као основни модел за предвиђање популарности. Проблему је приступљено као задатку класификације, користећи релевантне карактеристике објава за тренирање модела. Рад користи скуп података са инстаграм мреже за тренирање и евалуацију модела. Скуп података садржи информације о објавама као што су слика, текст, време овјављивања, број лајкова, број коментара и сл. Евалуација се врши кроз одређене метрике које омогућавају процену тачности предложеног модела. Метрике, као што су тачност, прецизност, одзив или F1 мера, користе се за оцену перформанси модела. Рад треба да пружи резултате експеримената који показују перформансе предложеног модела. Ово укључује тачност предикција, анализу релевантних метрика и упоређивање са другим приступима или моделима. Релевантност овог рада за овај пројекат се огледа у томе што пружа увиде у сличан проблем и приступ решавању истог.

[2] "*Predicting the Popularity of Online Content*"
https://www.researchgate.net/publication/23417017_Predicting_the_Popularity_of_Online_Content

У овом раду анализира се популарност онлајн садржаја на платформама YouTube и Digg. Показано је како је могуће предвидети популарност YouTube видео снимака на основу њихове почетне популарности. Сакупљање података са YouTube платформе. Одабрано је 7146 видео снимака из листе "most recently uploaded" како би се избегла пристрасност ка одређеним категоријама видео садржаја. Затим су, током наредних 30 дана, свакодневно ажурирани подаци о броју прегледа тих видео снимака. За предвиђање популарности видео садржаја коришћена су три модела. Први модел је линеарна регресија на логаритамској скали; least-squares absolute error. Други модел је модел константног скалирања: релативна квадратна грешка. Трећи коришћени модел је модел раста профила. Подаци о видео снимцима подељени су у две групе, тренинг скуп и тестни скуп, са расподелом од 50% за сваки скуп, односно по 3573 видео снимака у сваки скуп. Иако су се сва три модела показала одлична у предвиђању популарности видео снимака на YouTube-у, најбоље предвиђање даје други модел, односно модел константног скалирања. Релевантност овог рада се поново огледа у томе што се решава проблем који је сличан проблему који ће се решавати кроз овај пројекат.

[3] *Shan Li, Baoxu Shi, Jaewong Yang „Deep Job Understanding at Linkedin“*

Линк: https://arxiv.org/pdf/2006.12425.pdf

Циљ истраживања је унапређење разумевања послова на Linkedin платформи кроз примену дубоког преноса знања. Коришћена су дубока трансферна учења за развој модела специфичних за домен послова. Интеграција ових модела у Linkedin омогућава континуирано побољшање разумевања послова. Коришћени су необрађени огласи за посао са широким спектром информација попут броја пријављених, броја аплицираних као и детаљи о пословима који се налазе у понуди попут локације, информације о платама, начин рад (од куће, хибридно, из канцеларије) и сл. Евалуација се спроводи кроз Feedback loop са корисницима, укључујући поступак прикупљања повратних информација од постављача огласа.

Резултати: Побољшање задовољства корисника, повећање броја пријава за посао и ефикасност у препоручивању послова.

Закључак: Рад је релевантан за пројекат предикције популарности огласа за посао јер истражује ефикасне методе разумевања необрађених огласа.

Скуп података

Биће коришћен јавни скуп података који се налази на „Kaggle“ сајту, представљен у CSV формату. Сваког дана, хиљаде компанија и појединаца се окрећу Linkedin друштвеној мрежи у потрази за талентом. Овај скуп података садржи скоро свеобухватан запис о 33000 евидентирана посла, укључујући следеће атрибуте: job_id, company_id, title, description, location, applies, views, remote_allowed, currency, pay_period. Скуп података такође садржи и информације о платама на

конкретним пословима. Анализа ће примарно бити вршена на основу броја пријављених на оглас и броја прегледа огласа.

Линк ка скупу података: https://www.kaggle.com/datasets/arshkon/linkedin-job-postings/

Методологија

1. Претпроцесирање података: Анализа података уз помоћ Pandas библиотеке.
2. Анализа огласа: Проучавање веза између броја прегледа, броја пријављених и других атрибута огласа.
3. Подела података: Раздвајање података на обучавајучи и тест скуп. Добијени обучавајући скуп заједно са атрибутима се користи у одабраним моделима за учење.
4. Избор модела: Коришћење модела машинског учења попут SVM, Decision Trees, XGBoost, NN.
5. Евалуација модела: Поређење предвиђених вредности са стварним подацима кроз коришћење macro-F1 мере, RMSE (Root Mean Squared Error) и Спирмановог коефицијента корелације.

Метод евалуације

Класификација популарноси огласа се врши на основу броја прегледа, броја пријављених. Тачност модела ће се евалуирати кроз F1 meru, RMSE и Спирманов коефицијент корелације.

Софтвер

Пројекат ће се имплементирати користећи Python програмски језик и радно окружење PyCharm. Користиће се Pandas библиотека за претпроцесирање података, а за моделе ће се користити Keras и Scikit-Learn библиотеке.

План

План рада на пројекту обухвата следеће тачке:

- Експлоративна анализа података
- Обрада података
- Обучавање модела
- Евалуација модела

Тим

Дејан Курдулија (E2 39/2023)

Немања Малиновић (E2 45/2023)

Здравко Милинковић (E2 49/2023)