Rīgas 64. Vidusskola

**Kāda Konkrēta YouTube Video Komentāru Sadaļas Sentimenta Nolasīšana: VADER un RoBERTa Modeļu Salīdzinājums**

Zinātniski pētnieciskais darbs: Datorzinātņu un datorlingvistikas sadaļā

**Darba autors:**

Rīgas 64. vidusskolas 12. klases skolnieks

Kristers Laganovskis

**Darba vadītājs:**

Rīgas 64. vidusskolas programmēšanas skolotājs

Edvards Bukovskis

Rīga 2023

**Anotācija**

Zinātniski pētniecisko darbu: “Sentimenta novērtēšana YouTube komentāru sadaļai” Izstrādājis autors Rīgas 64. vidusskolas 12.DIT klases skolēns – Kristers Laganovskis. Darba Vadītājs Rīgas 64. vidusskolas programmēšanas skolotājs - Edvards Bukovskis.

Zinātniski pētnieciskais darbs tika veikts ar mērķi, izpētīt, kā ar dabiskās valodas apstrādi (NLP) iegūst sentimenta analīzi. Kādas ir būtiskākās atšķirības starp VADER un RoBERTTa modeļiem. Kā, tiek veikta efektīva datu ievākšana. Kā, tiek apstrādāti dati. Kā arī, cik precīzi ir iespējams novērtēt komentāru sadaļu balsoties uz mākslīgo intelektu un programmēšanas palīdzību. Zinātniski pētnieciskajā darbā balstoties uz analizēt un aplūkoto, tika izveidots pārlūka paplašinājums, ar kura palīdzību iespējams analizēt sentimentu kādam konkrētam YouTube video komentāru sadaļai.

Teorētiskajā daļa tika detalizēti izpētīti dabiskās valodas apstrādes (NLP) pamati, teksta sentimenta analizēšanai. Tās pielietošana sociālajā un ētiskajā kontekstā, kā arī tās nozīme mūsdienu un vēsturiskajā kontekstā. Tika apskatīts, kas ir VADER modelis un kas ir RoBRTTa modelis un to atšķirības. Kā, tiek veikta efektīva datu ievākšana un tā principi. Kā arī datu apstrādes principi, lai nodrošinātu pēc iespējas precīzāku sentimenta novērtējumu. Kā arī tika apsvērti sentimenta analīzes ierobežojumi.

Praktiskā daļā tika izstrādāts pārlūka paplašinājums, kas veidots uz VEDER un RoBERTa modeļu bāzes, šis paplašinājums sniedz iespēju vizualizēt sentimentu statistiski par kāda konkrēta YouTube video komentāru sadaļu, sniedzot vērtīgu perspektīvu par to, kāda ir publikas reakcija uz konkrēto saturu video. Paplašinājums nodrošina iespēju efektīvi filtrēt komentārus pēc to sentimenta, padarot analīzes procesu vēl efektīvāku. Taču šāda veida analīzei kļūdas ir tik lielas, ka visticamāk tam ir neliela praktiska pielietojuma vērtība. Iespējamie metodes uzlabojumi ietver nevēlamo reklāmu komentāru filtrēšanu un “emoji” sentimenta iekļaušanu.

Atslēgas Vārdi: Datu ievākšana, datu apstrāde, Komentāri, sentimenta analīze, pārlūka paplašinājums, dabiskās valodas apstrāde (NLP), mašīnmācīšanās, VADER, RoBERTa.

**Abstract**

Scientific Research Work: "Sentiment Analysis of YouTube Comment Section" Developed by the author, a student of the 12th-grade DIT class at Riga Secondary School No. 64 – Kristers Laganovskis. Supervisor: Edvards Bukovskis, programming teacher at Riga Secondary School No. 64.

The scientific research work was conducted with the aim of exploring how sentiment analysis is obtained through Natural Language Processing (NLP). It delves into the significant differences between the VADER and RoBERTa models, as well as how to efficiently gather data and process it. Additionally, it evaluates the precision of assessing the comment section using artificial intelligence and programming assistance. Based on the analysis and examination of the scientific research work, a browser extension was developed to analyze the sentiment of specific YouTube video comment sections.

In the theoretical part, the fundamentals of Natural Language Processing (NLP) for text sentiment analysis were thoroughly studied, including its application in social and ethical contexts, as well as its significance in modern and historical contexts. The VADER model and the RoBERTa model, along with their differences, were examined. The effective principles of data collection and processing were also discussed to ensure the most accurate sentiment assessment possible. The limitations of sentiment analysis were also considered.

In the practical part, a browser extension was developed based on the VADER and RoBERTa models. This extension provides the opportunity to statistically visualize the sentiment of a specific YouTube video comment section, offering valuable insight into the audience's reaction to the specific content of the video. The extension allows for the efficient filtering of comments by their sentiment, making the analysis process even more effective. However, the errors in this type of analysis are likely so significant that it has little practical application value. Possible method improvements include filtering out unwanted advertising comments and incorporating “emoji” sentiment.

Keywords: Data collection, data processing, comments, sentiment analysis, browser extension, Natural Language Processing (NLP), machine learning, VADER, RoBERTa.

**Saturs**

Anotācija…......................................................................................................................................2

Abstract...........................................................................................................................................3

Ievads…………………………………………………………………………………………………………………………………..…..5

Secinājumi………………………………………………………………………………………………………………….……………..0

Izmantotie informācijas avoti…………………………………………………………………………………………………….0

Pielikumi……………………………………………………………………………………………………………………………………0

**Ievads**

Mūsdienās informācijas plūsma ir pārāk bagātīga, un dažkārt tas var rasties kā problēma gan saturu patērētājam, gan to veidotājiem. Būtisks šķērslis ir tas, ka cilvēki bieži vien vērš pārāk lielu uzticību satura veidotājiem, kas, kā rāda pieredze, var nebūt vienmēr objektīvi. Šāda pieeja rada risku, ka satura veidotājs varētu apslēpt vai pat manipulēt ar informāciju, ietekmējot patērētājus. Tādēļ ir ārkārtīgi svarīgi uzzināt, kā citi cilvēki novērtē konkrētā video saturu. Īpaši šo var novērot interneta platformā "YouTube", kur satura veidotāji bieži veicina savu subjektīvo viedokli, ko patērētāji bieži uztver kā objektīvu faktisko informāciju. Tomēr tas var radīt lielas problēmas, jo ne vienmēr šim veidotājam būs pareiza nostāja. Tāpēc ir svarīgi uzzināt, ko citi cilvēki domā par konkrēto viedokli, vai saturu, un veikt pašam savus secinājumus. Veikt secinājumus pārskatot, pāris komentārus ir ātri un ērti, bet, kad to ir simtiem vai tūkstošiem, šāda analīze prasa ievērojami vairāk laika. Tāpēc sentimenta analīze ir lietderīga, jo tā ļauj strukturēt komentāru sadaļu nostājās - pozitīvas, neitrālas vai negatīvas.

Šajā pētījumā tiks apskatīts, kāda YouTube video komentāru sadaļa, izmantojot valodas apstrādes (NLP) modeļus, spēj nolasīt komentāru noskaņojumu, izmantojot VADER un RoBERTa modeļus. Mēs analizēsim metodes, kas tiek izmantotas, lai iegūtu nepieciešamos datus analīzei, kā arī kā šie dati ir jāapstrādā, lai sasniegtu veiksmīgu sentimenta novērtējumu. Turklāt tiks izvērtēta mākslīgā intelekta spēja precīzi noteikt komentāru noskaņojumu. Papildus tam, tiks izveidots paplašinājums, kas ļaus ikvienam analizēt komentārus un tos filtrēt pēc noteiktām kategorijām. Tas būs noderīgs ne tikai satura patērētājiem, bet arī veidotājiem, ļaujot tiem ietaupīt laiku, filtrējot komentārus.

**Darba mērķis:** Izpētīt dabiskās valodas (NLP) principus kā arī salīdzināt VADER un RoBERTa modeļus un izveidot paplašinājumu, kas ļaus analizēt komentārus.

**Hipotēze:** Autoram izdosies veiksmīgi salīdzināt VADER un RoBERTa modeļus kā arī analizēt un interpretēt YouTube video komentārus, izmantojot (NLP) modeļus un programmēšanu.

**Darba uzdevumi:**

1. Izpētīt (NLP) modeli un kā to pielietot sentimenta novērtēšanai.
2. Izpētīt atšķirības starp VADER un RoBERTa modeļiem.
3. Izpētīt, kā iegūt nepieciešamos datus no YouTube video komentāru sadaļas.
4. Izpētīt, kā apstrādāt datus efektīvi, lai tie sniegtu precīzāku novērtējumu.
5. Izstrādāt paplašinājumu, kas ļauj ikvienam analizēt un filtrēt komentārus pēc noteiktām kategorijām.
6. Veikt secinājumus

**Izmantotās darba metodes:** Literatūras apskats, lai varētu izpētīt, Dabiskās valodas apstrādes (NLP). Salīdzināšanas metode, lai varētu salīdzināt VADER un RoBERTa (NLP) modeļus. Sentimenta nolasīšanas metode, lai interpretētu iegūtos datus. Statistikas un analīzes metodes. Atvērtā pirmkoda pielietošana, “Python” programmēšanas valodas pielietošana radošās daļas veikšanai.

**Darba struktūra:** Darbs sastāv no x nodaļām, x apakšnodaļām, secinājumiem, izmantoto literatūras avotu saraksta un x pielikumiem. Darbā ir x attēli.