Predlog projekta iz SIAP-a

Ovaj dokument sadrži kratak opis onoga što je tema projekta i definicija, motivacija za odabranu temu. Nakon motivacije sledi pregled vladajućih stavova i shvatanja u literaturi, zatim skup podataka koji je ukratko opisan. Takođe je naveden i softver koji će biti korišćen, kao i metod evaluacije. Na samom kraju dokumenta nalazi se plan rada na projektu.

Tema projekta je semantička analiza tweet-ova koje su objavili korisnici nakon putovanja avionom. Izvor podataka sadrži 14485 postova objavljenih u periodu Februar, 2015. godine. Tweet-ovi se odnose na vodeće aviokompanije u SAD-u: American, Delta, Southwest, United, US Airways i Virgin America.

Definicija projekta

Analizom tweet-ova koji se odnose na letove utvrditi da li je iskustvo korisnika na tom letu bilo pozitivno, negatovno ili neutralno. Ideja ovog projekta je pružanje povratnih informacija aviokompanijama o zadovoljstvu svojih klijenata, kako bi oni mogli da odreaguju na odgovarajući način.

Motivacija

Cilj svake kompanije je zadovoljan klijent, a ukoliko je osoba izrazila nezadovoljstvo, zadatak kompanije je da utvrdi šta je uzrok tome i kako to sprečiti. Ako je utvrđen obrazac ponašanja da su klijenti nezadovoljni na tačno određenom letu jedne avio-kompnanije, potrebno je analizati šta je to šta im je smetalo (da li je loša usluga, kašnjenje leta ili neljubazno osoblje). Osnovni izvor prihoda i indikator dobrog poslovanja aviokompanije je broj putnika koji na dnevnom/mesečnom/godišnjem nivou koristi njihove usluge. Živimo u svetu društvenih mreža gde je najvredniji resurs informacija i metapodaci informacija, pa ukoliko osoba koja se još nije odlučila za aviokompaniju kojom želi da putuje dobije infomacije u negativnim iskustvima sa određenog leta, najverovatnije da se neće odlučiti za tu kompaniju.

Pegled vladajućih stavova i shvatanja u literaturi

[1] Fotis Misopoulos, Miljana Mitic, Alexandros Kapoulas, Christos Karapiperis, (2014) "Uncovering customer service experiences with Twitter: the case of airline industry", Management Decision, Vol. 52 Iss: 4, pp.705 - 723 http://www.emeraldinsight.com/doi/abs/10.1108/MD-03-2012-0235

Apstrakt:
Cilj rada je bio otkrivanje zadovoljstva potrošača servisima koje pružaju avio kompanije. Da li su potrošači zadovoljni ili ne su pokušavali da oktriju preko tweet-ova korisnika koji su spomenuli korisničke naloge avio kompanija preko kojih su leteli. Jedan od ciljeva je bio da se ustanovi kako su korisnici zadovoljni i nezadovoljni pojedinačnim servisima, i da se na osnovu tih informacija zaključi kako određene servise treba poboljšati.
Podaci:

Autori su imali na raspolaganju 67 953 javnih tweet-ova koji su bili upućeni ka 4 avio kompanije (u tweer-u se nalazi pomen - “@” i naziv korisničko ime avio kompanije).
Korišćeni algoritmi:
U istraživanju je korišćen lexicon-based pristup (procenjivanje orijentacije tweet-a na osnovu semantičke orijentacije reči ili frazi unutar njega) i vector space model obrade teksta.
Ostvareni rezultati:
Pozitivni sentimenti su se uglavnom odnosili na onlajn i check-in servise, cene letova i ocenu iskustva tokom leta. Negativni sentimenti su se odnosili na korišćenjem onlajn sajtova avio kompanija, pomerene letove i izgubljen prtljag.

[2] Adeborna, Esi and Siau, Keng, "An approach to sentiment analysis – the case of Airline Quality Rating" (2014). PACIS 2014 Proceedings. Paper 363. https://www.researchgate.net/profile/Keng_Siau/publication/265381265_An_Approach_to_Sentiment_Analysis_-_The_Case_of_Airline_Quality_Rating/links/554d27880cf29f836c9cd7f0.pdf

Apstrakt:
Ovaj rad otkriva orijentaciju i temu sentimetna iz teksta putem *sentiment mining* pristupa. Tekst koji se analizira je prikupljen iz tweet-ova korisnika avio kompanija. U istraživanju su takođe formirali ocenu avio kompanija (Airline Quality Rating (AQR)) koja je bazirana na rezultatima pomenutog pristupa.
Podaci:
Podaci su skinuti sa Twitter-a, leksikon od oko 6800 reči je korišćen sa 2006 pozitivnih reči i 4783 negativnih reči
Korišćeni algoritmi:
Za klasifikaciju sentimenata je korišćen Naïve Bayes algoritam. Da bi se odredila tema sentimenta korišćen je Sentiment Topic Recognition Model.
Ostvareni rezultati:
Koristeći Naïve Bayes algoritam dostigli su tačnost od 86.4%. Analizirali su tweet-ove usmerene ka tri kompanije:

[a] AirTran (57.5% pozitivnih, 27.6% negativnih i ostatak neutralnih tweet-ova),
[b] Frontier (64.1% pozitivnih, 18.0% negativnih i ostatak neutralnih tweet-ova),
[c] SkyWest (82% pozitivnih, 19.4% negativnih i ostatak neutralnih tweet-ova)

Skup podataka

Za ovu temu je predviđeno da se skup podataka formira na osnovu podataka dostupnih na internetu. Kao glavni izvor korišten je *Tweets* izvor podataka koji je dustupan na Kaggle repozitorijumu (https://www.kaggle.com/crowdflower/twitter-airline-sentiment). Podaci sadrže 14485 tweet-ova koji predstavljaju utiske putnika nakon letova. Pored same objave (tweet-a), podaci sadrže i sledeće atribute:

- **tweet_id** (jedistveni identifikator tweet-a),
- **airline_sentiment** (klasa kojoj pripada: neutralna, negativna ili pozitivna),
- **airline_sentiment_confidence** (indikator sigurnosti klasifikacije),
- **negative_reason** (razlog nezadovoljstva),

- **airline** (naziv aviokompanije),
- **name** (naziv korisničkog naloga),
- **retweet_count** (broj retweet-ova, tj. koliko puta je podeljena objava),
- **text** (sam tweet),
- **tweet_coord** (koordinata tweet-a),
- **tweet_created** (datum i vreme nastanka),
- **tweet_location** (lokacija nastalog tweet-a),
- **user_timezone** (vremenska zona u kojoj se korisnik nalazi).

Metodologija

S obzirom da su podaci dati u obliku tweet-ova potrebno ih je prethodno pripremiti za klasifikacioni algoritam. Koraci metodologije:

- Priprema dataseta, koja podrazumeva pretprocesiranje i tokenizaciju svakog tweet-a. Svaki tweet će biti podeljen u zasebne reči, biće izvršena zamena emotikona odgovarajućim rečima koje predstavljaju datu emociju, zatim zamena znakova interpunkcije (npr. uzvičnici predstavljaju naglašavanje i nešto bitno) kao i lematizacija reči tj. svođenje na zajednički oblik. Reči bismo tagovali u zavisnosti od toga šta je ta reč (imenica, zamenica…), zatim bi prepoznali da li je ta reč vlastita imenica, npr. naziv grada, aviona, aerodroma. Takođe bi bilo uzeto u obzir da li je i cela reč u tweet-u napisana velikim slovima, pa da li je shodno tome izraz osobe bio vikanje. Nakon pretprocesiranja teksta svakoj reči biće dodeljena ocena tj. sentiment koji ćemo odrediti koristeći SentiWordNet leksikon reči. Ukoliko bude potrebe, biće odrađen i TF-IDF weighting reči u tweet-u.
- Obučavanje modela i optimizacija parametara nad validacionim skupom podataka. Od algoritama za klasifikaciju koristićemo SVM, Naive Bayes i Random Forest.
- Testiranje modela nad test podacima
- Poređenje rezultata različitih algoritama

Primarni cilj ovog projekta jeste upoređivanje kako određeni klasifikacioni algoritam raspoznaje same tweet-ove i na koji način ih klasifikuje, kao i u kojoj meri su ti rezultati zadovoljavajući. Kao rezultat ćemo detaljno opisati naša zapažanja tokom samog procesa obrade.

Softver

Za izradu projekta će biti korišteni SQLite kao baza podataka za skladištenje, analizu i obradu podataka. Data mining analize će biti izvedene uz pomoć softvera za obradu i analizu podataka *Rapid Miner*, a slučaju da se javi potreba za dodatnim analiziranjem, koristiće se biblioteke za analizu podata u programskom jeziku Python. Aplikacija će takođe biti izrađena u programskom jeziku Python.

Metod evaluacije

Prilikom analize podataka utvrđeno je da je procenat negativnih tweet-ova (62,7%) znatno veći od pozitivnih (16,1%) i neutralih (21,2%) i iz tog razloga za evaluaciju tačnosti

modela biće korišćena F-mera. Evaluacija će se vršiti nad test podacima koji će biti izdvojeni kao 20% od ukupnog skupa podataka, dok će preostalih 80% opet biti podeljeno u odnosu 80:20 na validacioni skup podataka i trening skup podataka, respektivno. Nakon evaluacije modela biće odrađena analiza grešaka tako što ćemo izvojiti određeni podskup primera na kojima model greši. Nad tim primerima biće izvšena ručna analiza da bi se utvrdili uzroci nastanka grešaka.

Plan

Plan rada na projektu obuhvata sledeće tačke:

- eksplorativna analiza podataka,
- obrada podataka,
- obučavanje modela,
- evaluacija modela.

Članovi tima

- Nikola Đuza E2 111/2016
- Marina Nenić E2 11/2016
- Jana Vojnović E2 28/2016