**Opinion Mining U SOCIJALNIM MREŽAMA
Opinion Mining and Social Networks**

Aleksandar Ponjavić

**Elektrotehnički fakultet**

**Univerzitet u Beogradu, Srbija**

aleksandar.ponjavić@gmail.com

**Sadržaj -** U ovom radu je opisana uloga i značaj socijalnih mreža kao glavni izvor informacija za opinion mining, kao i rad samog algoritma.

**Abstract -** In this paper we discuss the role and importance of social networks as preferred environments for opinion mining and the algorithm itself.

1. uvod

Zadnjih decenija ogroman tehnološki razvoj je omogućio da internet bude dostupan skoro svim ljudima širom planete. Kao posledica toga, omogućena je komunikacija ljudi koji su veoma udaljeni, što je uzrokovalo pojavu i nagli razvoj socijalnih mreža.

Socijalne mreže omogućavaju korisnicima iz bilo kog dela planete da komuniciraju na vrlo jednostavan i unifikovan način, čime se razmenjuje ogromna količina podataka. Ti podaci su idealan izvor informacija za jednu od grana *Data Mining-a* koja se naziva *Opinion Mining.*

Postoji jako puno *Data Mining* algoritama i njihov broj je sve veći i veći. Takodje i broj aplikacija, koje nam pružaju podršku u Data Mining-u, povećava se svakodnevno. *Data Mining* algoritmi iz velikog broja jednostavnih podataka pronalaze skriveno znanje, odnosno veze koje na prvi pogled i ne moraju biti direktno povezane.

*Opinion mining* je tehnika koja omogućuje praćenje razmišljanja ljudi ili njihove stavove, kako bi donela opšte mišljenje javnosti o nekom proizvodu..

U ovom radu je opisana *Opinion Mining* tehnika, kao i socijalne mreže koje same po sebi predstavljaju najbolji izvor podataka. Glavni cilj je prikazati ključne faze tenika za *Opinion Mining,* opisati algoritam koji se primenjuje,kao i diskusije vezane za eventualna poboljšanja opisanog algoritma.

1. Socijalne mreže

Socijalne mreže su široko rasprostranjene na internetu. Omogućavaju svojim korisnicima brzu i laku komunikaciju. Podaci koji interaguju izmedju korisnika su pisani, jako često, na maternjem jeziku korisnika te mreže.

Socijalne mreže se najbolje predstavljaju pomoću grafova, gde svaki korisnik predstavlja jedan čvor. Povezanost čvora, odnosno broj veza sa ostalim čvorovima, nam omogućaja da pratimo uticajnost tog korisnika na ostale u mreži.

Jako bitan faktor u socijalnim mrežama, a koji je takodje veoma značajan i za *Opinion Mining,* jeste centralnost čvora (*eng. centrality*).

Centralnost se definiše pomoću:

1. Broj veza izmedju čvorova

*(veze predtsavljaju konverzacije izmedju korisnika)*

1. Broj najkraćih putanja koje prolaze kroz taj čvor

*(može da se definiše kao protok novih informacija od jednog korisnika do drugog)*

1. Srednje najkraće rastojanje do ostalih čvorova u mreži

*(može da se definiše koliko novih informacija taj korisnik/čvor prosledjuje drugima)*

Pojam centralnosti se poistovećuje sa pojmom *Opinion Leader*. Korisnik koje je proglašen za *Opinion Leader-*a ima veći uticaj na kompletno mišljenje korisnika i njegov stav kao takav se više i vrednuje. Ovaj pojam ćemo često pominjati u nastavku dokumenta i jako je bitan za razmevanje ekstrakcija finalnog stava.

1. OPinion mining

*Opinion Mining* je vrsta procesiranja nativnog jezika, koji predstavlja osnovni izvor informacija, kako bi pratili javno mišljenje grupe ljudi.

*Opinion Mining* podrazumeva izgradju kompletnog sistema koji treba da vrši sakupljanje i kategorizaciju mišljenja, njihovu obradu, tj. primenu odgovarajućih tehnika, kao i ekstrakciju finalnog stava. Taj sistem je jako veliki i često uključuje više različitih algoritama koji vrše pretprocesiranje podataka, a ne tiču se direktno *Opinion Mininga.* Mi ćemo se ovde fokusirati na deo koji se tiče isključivo *Opinion Mininga,* dok ćemo ostale delove samo pomenuti.

Postoje nekoliko *Opinion Mining* zadataka, a to su:

1. Analiza sentimenata
2. Eliminisanje subjektivnosti
3. Ekstrakcija mišljenja

**3.1 Analiza sentimenata**

Ovaj faza predstavlja prvo od faza pri primeni *Opinion Mininga.* Ulazni podaci su zasnovani na sirovom tekstu dobijenim direktno od socijalnih mreža, na nativnom jeziku mreže. Oni su obično sakupljaju pri diskusiji neke od tema, kao što su globalne teme (rast ekonomije, povećanje ugljen-dioksida u atmosferi, itd...) ili diskusije o nekom proizvodu. Cilj nam je da odgovarajuće komentare klasifikujemo, kako bismo mogli da izršimo predikciju o reakciji javnosti na neki od sledećih proizvoda ili da pokušamo da izvedemo zaključak na neku od diskutovanih tema.

Kada neki korisnik ostavlja komentar, može da u istoj rečenici daje i pozitivne i negativne komentare. Generalno ljudi imaju različite stavove prema nekim temama i ne moraju uvek da budu pravolinijske. Npr.pojedinac može prema nekom proizvodu da ima i pozitivno i negativno mišljenje, jer svaki proizvod ima svoje prednosti i mane. To značajno komplikuje *Opinion Mining,* pa je potrebno izvršiti kategorizaciju mišljenja na ograničen broj klasa.

Tri osnovne klase su:

* Pozitivno mišljenje
* Negativno mišljenje
* Neutralno mišljenje

Nakon što smo izvršili generalizaciju mišljenja, nephodno je da svaki komentar pojedinca svrstamo u odredjenu klasu. To je kraj ove faze.

**3.2 Eliminisanje subjektivnosti**

Nakon analize sentimenata mi smo izvršili klasifikaciju mišljenja, medjutim mi ne znamo verodostojnost tih iskaza. Tačnije pre nego što ta mišljenja tretiramo kao validma, moramo da izvršimo neku vrstu procesiranja nad njima. Sada na ulogu stupaju neki drugi algoritmi *Data Mininga* koje smo pominjali malopre.

Glavni zadatak ove faze jeste da eleminiše nivo subjektivnosti pojedinaca. Tačnije, cilj je identifikovati samo objektivni deo kod mišljenja pojedinaca. Taj postupak nije lak i zasniva se na *Link Mining* algoritmima pomoću koji se identifikuju veze izmedju pojmova. Na taj način je moguće eliminisati subjektivnost pojedinca (*npr. davanje uvek poitivnih komentara najboljim prijateljima na mreži*). Sam *Opinion Mining* se ne bavi direkto ovom problematikom, ali je usko povezan i njegovom primenom se značajno poboljšava kvalitet predikcije mišljenja.

**3.3 Ekstrakcija mišljenja**

Ova faza podrazumeva predikciju stava ili mišljenja o nekom proizvodu ili temi na osnovu prethodnih informacija. Obično su to neki statistički algoritmi, koji sa odgovarajućom verovatnoćom prognoziraju mogućnost pojave nekog ishoda, pomoću gore spomenutih podataka.

U ovoj fazi je jako bitno osvrnuti se na same socijalne mreže. Tačnije, treba da se osvrnemo na pojam centralnosti, odnosno *Team Leadera,* jer ćemo mišljenja od takvih pojedinaca mnogo više vrednovati nego ostala.

Ova faza je ujedno i finalna pa bi kao krajnji rezultat trebalo da izbaci prognozirano mišljenje.

1. Algoritam

Glavni deo ove tehnike jeste identifikacija klase sentimenata, odnosno glavni deo se realizuje u prvoj fazi. Da bismo odredili semantičku orijentaciju nekog dokumenta ili komentara neophodno je da izvršimo analizu celokupnog teksta. Vrši se analiza svakog dela dokumenta reč, po reč.



(3)

(2)

(1)

gde su, *Ti – i-ti* pojam u dokumentu, *|d| -* broj pojmova u dokumentu, *Cp i Cn* – pozitivne i negativne klase i *score() –* funkcija koja dodeljuje pozitivne ili negativne vrednosti u zavisnosti od prirode pojma.

Semantička orijentacija se pronalazi pomoću *eng. dictionary method.* Pomoću ove metode se u vreme učitavanja dokumenta vrši klasifikacija pojmova u neku od klasa u zavisnosti od njegove bazične orijentacije ili u zavisnosti od orijentacije pojmova oko njega. Na početku rada ove metode kreira se početni skup pojmova i definiše im se bazična orijentacija, nakon toga se svaki sledeći pojam klasifikuje do kraja dokumenta. Kada se završi procesiranje svih pojmova, onda se odredjuje klasifikacija čitavog dokumenta u jednu od tri gore pomenute klase: pozitivnu, negativnu, neutralnu.

Za dobar rad algoritma je najbitnija funkcija *score()* u jednačinama (2) i (3) i ona je jednaka:



gde su, *p(t|Cp) i p(t|Cn)* – uslovne promenljive koje predstavljaju verovatnoću pojavljivanja pojma *t* u pozitivnim, odnosno negativnim klasama.

1. Poboljšanja

Gore navedeni postupak je uzimao u obzir samo 3 klase: pozitivno mišljenje, negativno mišljenje i neutralno. Ta podela je suviše restriktivna, jer ljudi nikada ne razmišljaju na takav način. Poboljšanje bi se moglo napraviti ukoliko bi smo definisali više različitih nivoa, tj. klasa. Na taj način bi dobili finije rezultate uz nesto komplikovaniji algoritam selekcije.

Drugo poboljšanje bi se ogledalo u tome da se odredjenim korisnicima dodeljuju neke vrste poena, odnosno bonusa u zavisnosti od prethodnih predvidjanja. Ukoliko su u nekim prethodnim temama ti korisnici bili konstanto u pravu, onda bi njihovo mišljenje trebalo vrednovati bolje, u suprotnom slučaju, analogno. Ovaj pojam podseća na *Team Leadera*, medjutim razlikuje se, jer ovi pojedinci ne moraju da imaju dobro definisanu centralnost (pominjano u odeljku socijalnih mreža).

Još jedna vrsta poboljšanja se ogleda u tome da se pored izvora sirovih podataka aktivno koriste i još neke informacije od socijalnih mreža. Jedna od mogućih primena jeste praćenje takozvanih spam korisnika (*eng. troller*), odnosno ljudi koji konstantno ostavljaju komentare koji nemaju smisla ili upotrebnu vrednost, ili na neke druge načine ne doprinose samoj predikciji mišljenja.

1. Prednosti i praktične primene

*Opinion Mining* ima široku primenu, sreće se industriji, ekonomiji, u edukacione svrhe itd. Korišćenje ove tehnike je praktično neograničeno, jer se u skoro svakoj društvenoj sferi vrši obrada takvih informacija.

Najčešće se primenjuje kod marketinške industrije, gde je neophodno da se izvrši evaluacija reakcije ljudi na neki od novih proizvoda. Takodje se primenjuje i u većim firmama koje treba da se odluče pri selekciji nekih od proizvoda, na osnovu reakcija ljudi na izradjene prototipove.

Još jedna široka primena ove tehnike jeste ta, što nije više potrebno vršiti anketiranje. Ankete su često jako dosadne, ma koliko jednostavne bile i ljudi ih masovno izbegavaju, jer konkretno njima ne donose ništa korisno. Pošto *Opiniom Mining* kupi sirove informacije sa raznih izvora, uz odgovarajuću filtraciju je moguće izdvojiti baš one koju su bile deo komentarisanja nekog proizvoda, na taj način se koriste postojeći podaci kako bi se popunila neka anketa, a informacije se nabavljaju legalno bez dodatnog opterećivanja korisnika.

1. moguće prepreke i mane

Prilikom primene ove tehnike može se naići na mnoge poteškoće, koje se pre svega baziraju na fazu Analize sentimenata.

Neke reči u izvornim jezicima imaju ogromne brojeve sinonima, imaju različito značenje u zavisnosti od konteksta ili se pišu različito u zavisnosti od primene. Takve reči su jako teške za analizu i klasifikaciju. Izbor jezika je jako bitan faktor u ovoj fazi, jer neki jezici su posebno teški za obradu.

Kako bi se rešio taj problem potrebno je implementirati jako *machine learning* algoritme, koje će imati mogućnost „učenja“ klasifikacije. Primena ovih algoritama dodatno komplikuje već veliki sistem, tako da se ovo može identifikovati kao jedna od većih mana.

Jedna od dobrih strana upotebe socijalnih mreža u *Opinion Minig* je bila ta, što imamo skoro neograničen izvor informacija. Medjutim upotreba socijalnih mreža ima i negativnu stranu. Jedna od njih se ogleda u tome da bilo ko može da ostavi svoje mišljenjne, čak i da njegova kredibilnost nije velika. Ovaj problem ume da bude veliki u početku, medjutim primenom odgovarajućeg poboljšanja na duge staze se prevazilazi, jer će se postepeno takvi korisnici tretirati kao *eng. spammer/troller.*

1. Zaključak

U ovom radu je opisano kako pomoću *Opinion Mining-a* može da se odrediti predikcija mišljenja. Ova tehnika je usko povezana i sa ostalim vrstama *Data Mining-a* i jako se oslanja na njihov ispravan rad.

Rezultati algoritma su zadovoljavajući i temelje se na osnovu rezultata istraživanja u drugim radovima, koji su navedeni u odeljku Literatura.

Na osnovu generalnog istraživanja i poznavanja društvenih mreža, ova tehnika bi trebalo da daje jako dobre rezultate i u bilo kojim drugim sferama gde ima dosta tekstualnih podataka, a ne samo u socijalnim mrežama.

literatura

1. Bo P, Lillian Lee, „Opinion Mining and Sentiment Analysis“ Now Publishers Inc, 2008.
2. Xindong W. „Top 10 Algorithms in Data Mining“, (<http://home.etf.rs/~vm/os/dmsw/Top10DMAlgorithms.pdf>), University of Vermont, USA, 17.11.2014. @16h
3. Milutinović V. „Data Mining Versus Semantic Web“,(http://home.etf.rs/~vm/os/dmsw/TutorialDMvsSW.ppt), Beograd, Srbija, 17.11.2014. @17h
4. Bing L, „Sentiment Analysis: Mining Opinions, Sentiments and Emotions“, Cambridge University Press, 2009.
5. Jędrzejewski K. „Opinion Mining and Social Networks: Apromising Match“, Institute of Computing Science, Poznan University of Technology, Poznan, Poland, 2010.