ივანე ჯავახიშვილის სახელობის თბილისის სახელმწიფო უნივერსიტეტი
ზუსტ და საბუნებისმეტყველო მეცნიერებათა ფაკულტეტი

სტუდენტი: ჯოლოხავა მადლენა

განწყობის ანალიზი

სამაგისტრო პროგრამა - ინფორმაციული ტექნოლოგიები

ნაშრომი შესრულებულია ინფორმაციული ტექნოლოგიების
მაგისტრის აკადემიური ხარისხის მოსაპოვებლად

ხელმძღვანელი: გელა ბესიაშვილი

თბილისი
2019

ანოტაცია

განწყობის ანალიზი არის განსაკუთრებული კლასიფიკაციის ალგორითმი, რომელიც მიზანმიმართულია შეხედულების პოზიციის პოვნაზე და გამოკვეთავს ამ პროცესში ინფორმაციას ინტერესის მიხედვით. განწყობის ანალიზი, ასევე ცნობილია როგორც შეხედულების ანალიზი, წარმოადგენს ნატურალური ენის დამუშავების (NLP) ნაწილს,რომელიც ქმნის სისტემას ,რათა ამოიცნოს და ამოიღოს შეხედულება ტექსტიდან. როგორც წესი, შეხედულების განსაზღვრის გარდა, ეს სისტემები გამოყოფენ გამოხატვის ისეთ ატრიბუტებს, როგორიცაა:

- პოლარობა: გამოხატავს თუ არა მოსაუბრე პოზიტიურ ან ნეგატიურ აზრს.
- თემა: რის შესახებ საუბრობდა.
- აზრის მფლობელი: პირი, რომელიც გამოხატავს აზრს.

შეხედულების ანალიზი სწრაფად მზარდი სფეროა. არსებობს მრავალი ელექტრონული სავაჭრო საიტები რომლებიც საშუალებას აძლევს მომხმარებლებს,გამოხატონ აზრი სხვადასხვა პროდუქტის შესახებ. ეს შეფასებები სასარგებლოა იმ მომხმარებელთათვისს, რომლებიც აპირებენ პროდუქტის შეძენას. განწყობის პროგნოზირების ზუსტი მეთოდი საშუალებას იძლევა ინტერნეტიდან მივიღოთ კლიენტის შეხედულება პროდუქტზე და ამის მიხედვით განვსაზღვროთ კლიენტის ინტერესები. არსებობს სხვადასხვა ალგორითმები განწყობის ანალიზისთვის. პოლარობის გამოვლენის ნებისმიერი ალგორითმის გამოყენებამდე ხორციელდება უკუკავშირის წინასწარი დამუშავება. ამ წინასწარ დამუშავებული მიმოხილვებიდან გამომდინარეობს სიტყვები და ობიექტები, რომლებზეც გენერირდება შეხედულება და გამოიყენება ნებისმიერი ანალიზის მეთოდი, რათა ნაპოვნი იყოს მიმოხილვის პოლარობა. შეხედულების ანალიზს აქვს დეტალიზაციის სამი დონე: დოკუმენტის ეტაპი, წინადადების დონე და ასპექტის დონე. ამ ნაშრომში წარმოდგენილია სხვადასხვა ალგორითმები განწყობის ანალიზისთვის და განხილულია გამოწვევები და აპლიკაციები ამ სფეროში.

საკვანძო სიტყვები: განწყობის ანალიზი, მანქანური სწავლება, დამხმარე ვექტორების მანქანა, გადაწყვეტილების ხეები, რეკურენტული ნეირონული ქსელები, ბაიესი

Sentiment analysis

Abstract

Sentiment analysis is a predominantly classification algorithm aimed at finding an opinionated point of view and its disposition and highlighting the information of particular interest in the process. Sentiment Analysis also known as Opinion Mining is a field within Natural Language Processing (NLP) that builds systems that try to identify and extract opinions within text. Usually, besides identifying the opinion, these systems extract attributes of the expression e.g.:

- Polarity: if the speaker express a positive or negative opinion,
- Subject: the thing that is being talked about,
- Opinion holder: the person, or entity that expresses the opinion.

Opinion mining and sentiment analysis is rapidly growing area. There are numerous e-commerce sites available on internet which provides options to users to give feedback about specific product. These feedbacks are very much helpful to both the individuals, who are willing to buy that product and the organizations. An accurate method for predicting sentiments could enable us, to extract opinions from the internet and predict customer‟s preferences. There are various algorithms available for opinion mining. Before applying any algorithm for polarity detection, pre-processing on feedback is carried out. From these pre-processed reviews opinion words and object on which opinion is generated are extracted and any opinion mining technique is applied to find the polarity of the review. Opinion mining has three levels of granularities: Document level, Sentence level and Aspect level. In this paper various algorithms for sentiment analysis are studied and challenges and applications appear in this field are discussed.

Keywords: Sentiment Analysis, Machine Learning, Support Vector Machines, Decision Trees, Recurrent Neural Networks, Naive Bayes

შინაარსი

შესავალი

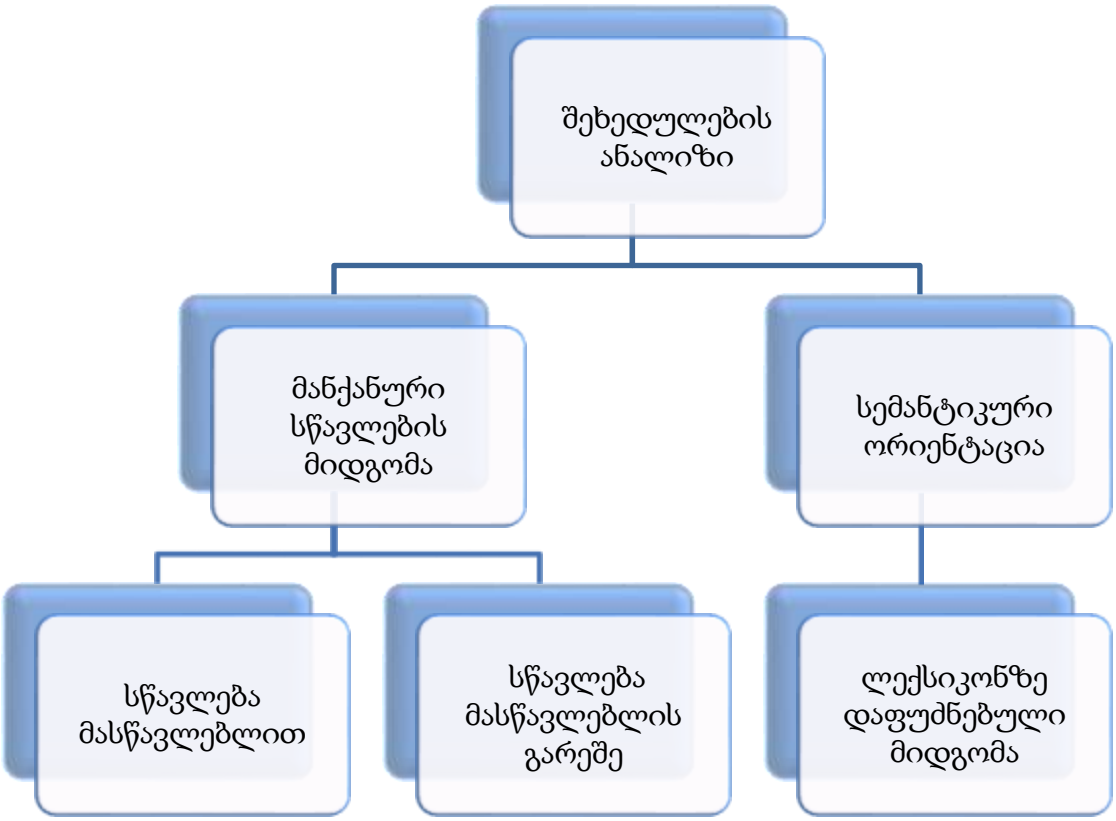
განწყობის ანალიზი(SA) , იგივე შეხედულების ანალიზი(OM) - ეს არის ხალხთა აზრის კომპიუტერული შესწავლა, მათ შეხედულებების და ემოციებისა ამოცნობა კონკრეტულ ობიექტთან მიმართებაში. საერთო ჯამში, ანალიზით შესაძლებელი ხდება შეგროვდეს ინფორმაცია დადებით და უარყოფით ასპექტებში განსაზღვრულ თემებზე. შედეგად, დადებითი და მაღალი შეფასების მქონე პროდუქტები რეკომენდირებული ხდება მომხმარებლებისთვის. იმისათვის რომ ხელი შეუწყონ მარკეტინგს, მსხვილი კომპანიები და ბიზნესმენები ახდენენ შეხედულების ანალიზის გამოყენებას.

ამჟამად, განწყობის ანალიზი სარგებლობს გაზრდილი ინტერესითა და განვითარებით, ვინაიდან გააჩნია ბევრი პრაქტიკული გამოყენება. დიდი რაოდენობით ტექსტები, გამოხატული აზრები ხელმისაწვდომი ხდება საიტებზე, ფორუმებზე, ბლოგებსა და სოციალურ ქსელებში. განწყობის ანალიზის სისტემის გამოყენებით არასტრუქტურირებული ინფორმაცია შესაძლოა ავტომატურად ტრანსფორმირდეს საზოგადოებრივ სტრუქტურირებულ შეხედულებებად, პროდუქტების შესახებ, მომსახურებების შესახებ, ბრენდების შესახებ, პოლიტიკური თუ ნებისმიერი თემის შესახებ, რომელზეც ხალხი აზრს გამოხატავს. ეს მონაცემები შესაძლოა იყოს სასარგებლო კომერციული აპლიკაციებისთვის, ასევე მარკეტინგული ანალიზისთვის, საზოგადოებრივი ურთიერთობებისთვის, ქსელური პრომოუთერების შეფასებისთვის, პროდუქტის განხილვისთვის და კლიენტების მომსახურებისთვის.

არსებობს მრავალი ნაშრომი განწყობის ანალიზის შესახებ, მომხმარებლების შეხედულებების მიხედვით, რომელიც ძირითადად განიხილავს მომხმარებლის შეხედულების პოლარობას. ამ კვლევებში განწყობის ანალიზი ხშირად წარმოდგენილია როგორც ერთერთი სამი დონიდან: დოკუმენტური დონე, წინადადების დონე ან ატრიბუტის დონე. რაც შეეხება განწყობის ანალიზს, ჩატარებული კვლევების მიხედვით არსებული ლიტერატურა მიუთითებს მეთოდის ორ ტიპზე, მანქანური სწავლება და სემანტიკური ორიენტაცია.

ეს მეთოდები ნაჩვენებია

სურათზე:



არსებობს რამდენიმე პრობლემა განწყობის ანალიზში. ერთი ეს არის განწყობის გამომხატველი სიტყვა, რომელიც ერთ სიტუაციაში უნდა ჩაითვალოს დადებითად, თუმცა ნეგატიურად ითვლება სხვა სიტუაციაში. მეორე პრობლემა მდგომარეობს იმაში რომ ხალხი ყოველთვის ერთნაერად არ აფიქსირებს შეხედულებას. ტრადიციულად დამუშავებული ტექსტების დიდი ნაწილი დაფუძნებულია იმ ფაქტზე რომ ტექსტის ორ ნაწილს შორის მცირე სხვაობა ძალიან არ ცვლის შეხედულებას საერთო ჯამში. განწყობის ანალიზში „სურათი იყო მშვენიერი“ ძალიან განსხვავდება „სურათი არ იყო მშვენიერი“. ხალხს შეუძლია შეაბრუნოს თავისი გამოთქმული წინადადება. ბევრი შეხედულება შესაძლოა იყოს როგორც პოზიტიური,ასევე ნეგატიურიც, რომელიც გარკვეულწილად მართვადია წინადადების ანალიზით მიმდინარე დროში.

მომხმარებლები აფიქსირებენ თავიანთ შეხედულებებს პროდუქტების ან მომსახურებების შესახებ, რომლითაც სარგებლობენ ისინი საიტებიდან ან ბლოგებიდან, ეს

სასარგებლოა როგორც სხვა მომხმარებლებისთვის, ასევე მწარმოებლებისთვის, რათა იცოდნენ რას ფიქრობს საზოგადოება კონკრეტულ პროდუქტსა თუ მომსახურებაზე.

არაფორმალურ მედიაში, როგორიცაა ეს Twitter ან ბლოგები, ხალხი აერთიანებს განსხვავებულ აზრებს ერთ წინადადებაში, რომელიც მარტივად გასაგებია ადამიანისთვის, მაგრამ კომპიუტერისთვის რთული დასამუშავებელი. ხანდახან შესაძლოა ხალხისთვისაც გაუგებარი დარჩეს დაფიქსირებული შეხედულება , რომელიც გადმოცემულია კონტექსტს მოკლებული მოკლე ტექსტით.

არსებობს შეხედულების ანალიზის <u>სამი დონე:</u>

1. *დოკუმენტური დონე:* დოკუმენტი ამ მიდგომისას განიხილება როგორც ერთი მთლიანობა და ანალიზი კეთდება მთლიან დოკუმენტზე. შედეგი რომელიც დოკუმენტზე გენერირდება ,ყოველთვის შესაბამისი არ არის.
2. *წინადადების დონე:* წინანდადება განიხილება როგორც ერთი მთლიანობა და ანალიზი ტარდება ინდივიდუალურად ყოველ წინადადებაზე. შემდეგ შედეგები ერთიანდება რომ მთლიანი დოკუმენტის შედეგი დაიდოს.
3. *ასპექტის დონე:* ხდება ფრაზების დონეზე ანალიზი, რომელიც ახდენს განწყობის გამოვლენას ასპექტების მიხედვით ობიექტებზე.

(ასპექტები, რომლებიც მკაფიოდ არის ნახსენები როგორც არსებითი სახელები , ეწოდებათ პირდაპირი ასპექტები. არაპირდაპირი ასპექტები-რომლებიც პირდაპირ არ არიან მოხსენიებული წინადადებაში,მაგრამ არიან ნაგულისხმები.)

ნაშრომში განხილულია გასხვავებული ალგორითმები განწყობის ანალიზისთვის, პრობლემები და განწყობის ანალიზის მეთოდები ასევე განიხილება.

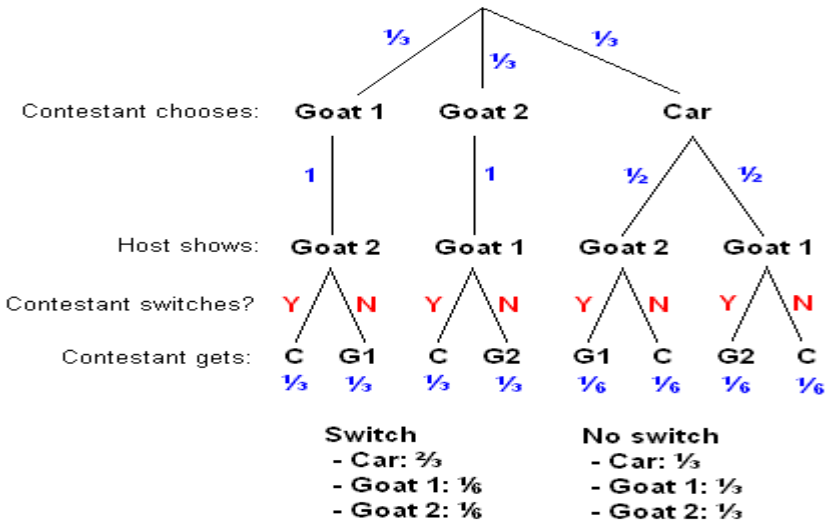
განწყობის ანალიზის მეთოდები

სწავლება მასწავლებლით

ეს მეთოდი შეიცავს დოკუმენტის ორ ნაკრებს, სასწავლო და სატესტო ნაკრები. დოკუმენტის შესასწავლად გამოიყენება სასწავლო ნაკრები კლასიფიკატორის მიერ. ვალიდაციისთვის კი გამოიყენება სატესტო ნაკრები. მიმოხილვის კლასიფიკაციისთვის კი შესაძლოა გამოყენებული იქნას მრავალი მეთოდი.

<u>სწავლება მასწავლებლით-ტიპები :</u>

კლასიფიკატორი-გადაწყვეტილებათა ხე



გადაწყვეტილებათა ხე უზრუნველყოფს იერარქიულ დეკომპოზიციას სასწავლო მონაცემების სივრცისათვის, რომელშიც ატრიბუტის მნიშვნელობის პირობა გამოიყენება მონაცემების გასაყოფად. პირობა ან პროგნოზირება არის ერთი ან მეტი სიტყვის არსებობა ან არარსებობა. სივრცეში მონაცემების გაყოფა ხორციელდება რეკურსიულად, სანამ საბოლოო კვანძები შეიცავს ჩანაწერების განსაზღვრულ მინიმალურ რაოდენობას,რომელიც გამოიყენება კლასიფიკაციისთვის.

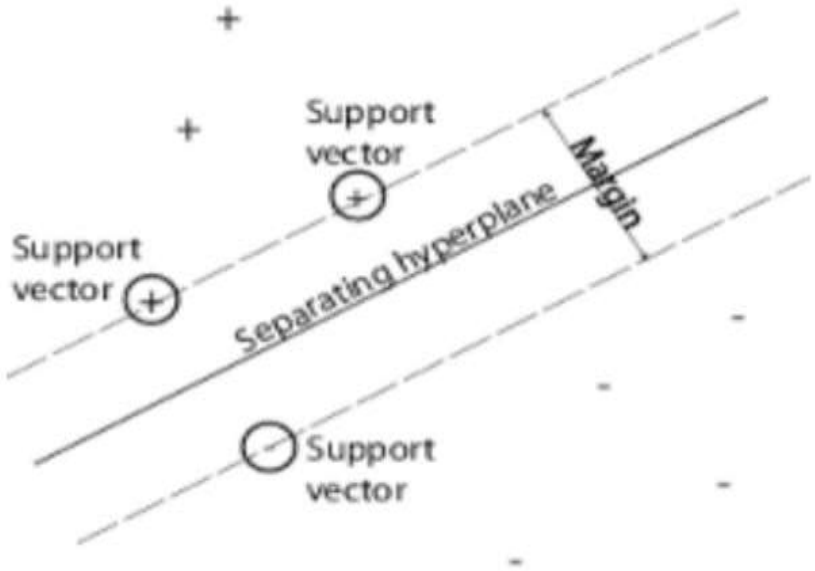
ემოციების კლასიფიცირებისა და გადაწყვეტილებათა ხის გამოყენებით ანალიზირდება სხვადასხვა ემოციური ვარიაციები. თუ-მაშინ(if-then) წესი ასევე გენერირდება გადაწყვეტილებათა ხისგან.

წრფივი რეგრესიის კლასიფიკატორი

ა. დამხმარე ვექტორების მანქანა(Support vector machine SWM):

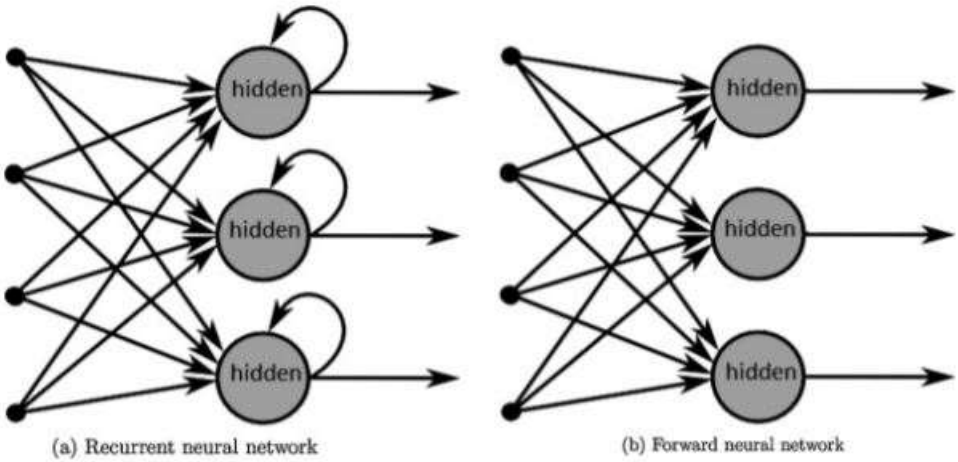
ტექსტური მონაცემები იდეალურად ერგებიან SWM კლასიფიკაციას,რადგან ტექსტის მეჩხერი ბუნება განაპირობებს რომ ზოგიერთი ნაწილები არ ფლობენ მნიშვნელობას,მაგრამ ისინი დაკავშირებულია ერთმანეთთან და როგორც წესი ორგანიზებულია ხაზოვნად გაყოფად კატეგორიებად. SVM გამოიყენება პოლარობის კლასიფიკაციისთვის ყოველი ასპექტისთვის. ექსპერიმენტული შედეგები მიუთითებენ რომ მეთოდი აღწევს 78%იან სიზუსტეს ემოციის დადგენაში.

SWM წარმოადგენს მეთოდს,რომელიც მიიჩნევს რომ ყოველი ფუნქციის ნაკრები წარმოადგენს შიდა ჰიპერსივრცეს, SWM კი ცდილობს დაყოს იგი ჰიპერსიბრტყისა და ფუნქციის საშუალებით მინიმიზირებული ვექტორების დახმარებით. ასეთი სივრცობრივი გაყოფა რთული შესასრულებელია, ხანაც თითქმის შეუძლებელი,რისთვისაც SWM იყენებს დასაშვები გადახვევას, რომელიც საშუალებას აძლევს ზოგიერთი მაგალითი არასწორად დაყოს,თუმცა ზრდის საერთო შედეგიანობას.



ბ. ნეირონული ქსელები

ნეირონული ქსელი-ეს არის მეთოდი, რომელიც ცდილობს ოპტიმიზირება გაუკეთოს ზოგ წონას, ნეირონის სხეულს, რომელიც მრავლდება მახასიათებელი ვექტორებით, დენდრიტებით. ამ გამრავლების შედეგი არის პროგნოზირება მოხდეს ნეირონების საშუალებით, აქსონის ტერმინალით. იგი შესაძლოა გამოყენებულ იქნეს როგორც შედეგი ან ფუნქცია შემდეგი ნეირონების ნაკრებისთვის, რომელსაც ეწოდება მრავალდონიანი ნეირონული ქსელი (MLP) . მიზანი კი არის შიდა წონების მომზადება გრადიენტული დაშვების მეთოდის და უკუგავრცელების ალგორითმის საშუალებით, სადაც შეფასების ფუნქცია გამოითვლება და შედეგი გადაეცემა უკან ნეირონების წონებს, რომლებიც განახლდებიან, რათა მოახდინონ საბოლოო ფუნქციის მინიმიზაცია. არსებობს რეკურენტული ნეირონული ქსელები, როდესაც გამოიყენება შიდა მეხსიერება ყოველ ნეირონზე, რომელიც წარმოადგენს შუალედურ გაგებას ფუნქციებს შორის, სადაც შესაძლოა იყოს დაგროვილი ან დავიწყებული ნეირონები.



(a) Recurrent neural network (b) Forward neural network

ნეირონული ქსელები შეიცავს მრავალ ნეირონს,სადაც ნეირონი წარმოადგენს საბაზო ერთეულს. ნეირონების შეყვანა აღინიშნება ვექტორით Xi,რომელიც წარმოადგენს i-ურ სიხშირის სიტყვას დოკუმენტში. A არის წონების ნაკრები,რომელიც დაკავშირებულია ყოველ ნეირონთან და იყენებს გამოსათვლელ ფუნქციას მისი შემსვლელისთვის. შემავალ მონაცემებზე დაყრდნობით და წონის კოეფიციენტებით გენერირდება შედეგი.

წესებზე დამყარებული კლასიფიკატორი

წესებზე დამყარებული კლასიფიკატორში მონაცემების სივრცე მოდელირდება წესების ნაკრებით. მარცხენა მხარე წარმოადგენს ფუნქციების ნაკრების პირობებს, დიზუნქციურ ნორმალურ ფორმაში წარმოდგენილს,როცა მარჯვენა მხარე წარმოადგენს კლასების მისამართს. ეს კლასიფიკატორი წარმოადგენს ემოციურ მოდელს, სადაც ემოციური ლექსიკა შესაძლოა აიწყოს ხელით ან ავტომატურად ,ამასთანავე გამომწვევი კომპონენტების პროპორციები შესაძლოა გამოთვლილ იქნეს მრავალენოვან ემოციებშიც. ექსპერიმენტის შედეგები ამტკიცებენ მიდგომის მოქნილობას.

სავარაუდო(ალბათური) კლასიფიკატორი

ა. Naïve bayes

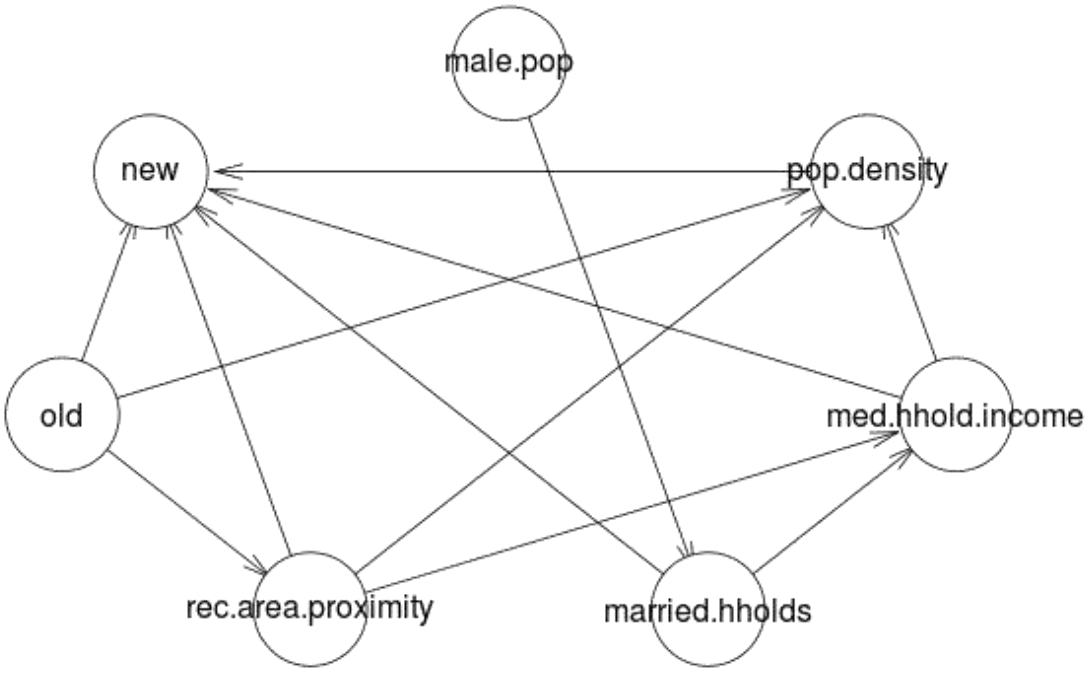
ბაიესის კლასიფიკატორი არის ყველაზე მარტივი და ყველაზე გავრცელებული კლასიფიკატორი. იგი გამოითვლის კლასის შუალედურ ალბათობას, დოკუმენტში სიტყვების განაწილების საფუძველზე. მოდელი მუშაობს BOW-თან, რომელიც უგულებელყოფს სიტყვის პოზიციას დოკუმენტში. იგი იყენებს ბაიესის თეორემას ალბათობის პროგნოზირებისთვის რომ მოცემული ფუნქციის ნაკრები ეკუთვნის კონკრეტულ ნიშანს.

სისტემა მიმოიხილავს მომხმარებლების შეხედულების ასპექტებს. ყოველი წინდადებიდან ხდება არსებითი სახელების და არსებითი ფრაზების ამოღება. ბაიესის ალგორითმი მასწავლებლით დასწავლასთან ერთად გამოიყენება იმის გასაგებად, წინადადება არის ნეგატიური თუ პოზიტიური და ამასთანავე განსაზღვრავს მის რაოდენობას.

for each class $c \in C$ # Calculate $P(c)$ terms
N_{doc} = number of documents in D
N_c = number of documents from D in class c
$logprior[c] \leftarrow \log \frac{N_c}{N_{doc}}$
$V \leftarrow$ vocabulary of D
$bigdoc[c] \leftarrow$ **append**(d) **for** d $\in$ D **with** class c
for each word w in V # Calculate $P(w|c)$ terms
$count(w,c) \leftarrow$ # of occurrences of w in $bigdoc[c]$
$loglikelihood[w,c] \leftarrow \log \frac{count(w,c) + 1}{\sum_{w' \text{ in } V} (count(w',c) + 1)}$
return *logprior*, *loglikelihood*, *V*

<u>გ. ბაიესის ქსელი</u>

წარმოდგება ორიენტირებული აციკლური გრაფით, რომლის კვანძები წარმოადგენს შემთხვევით ცვლადებს და წიბოები წარმოდგენილია პირობითი დამოკიდებულებებით. NB ითვლება ცვლადების და მათი დამოკიდებულებების სრულ მოდელად. ტექტის დამუშავებაში NB-ს კომპლექსურობის გამოთვლა ძალიან ძვირია,ამიტომ არცისე ხშირად გამოიყენება.



გ. მაქსიმალური ენტროპია

მაქსიმალური ენტროპიის კლასიფიკატორი (ცნობილია, როგორც პირობითი ექსპონენციალური კლასიფიკატორი) გარდაქმნის მნიშვნელობების სიას ვექტორებად,კოდირების გამოყენებით. კოდირებული ვექტორი გამოიყენება წონების გამოსათვლელად, რომელიც შეიძლება გამოყენებულ იქნეს ყველაზე ხშირი მნიშვნელობის განსაზღვრისთვის ფუნქციების სიაში.

ლექსიკონზე დაფუძნებული მიდგომა

ასეთი მიდგომის დროს იკრიბება მცირე ნაკრები სენტიმენტალური სიტყვებისა,რომელთაც ეწოდებათ საწყისი სიტყვები მათი დადებითი და უარყოფითი ორიენტაციების მიხედვით.შემდეგ ეს ნაკრები იზრდება სინონიმების და ანტონიმების მოძებნით WordNet-ში ან სხვა ონლაინ ლექსიკონებში. ახალი სიტყვები ემატება არსებული სიტყვების სიას, შემდეგ იწყება მომდევნო იტერაცია. ინტერაცია შესაძლოა შეჩერდეს,თუ არიქნა მოძიებული არცერთი ახალი სიტყვა. ბოლოს გამოიყენება ინსპექტირების ნაკრები სიის გასაწმენდად.პოლარობა გამოითვლება აზრთა უმრავლესობის საფუძველზე.

სწავლება მასწავლებლის გარეშე

ეს მეთოდი გამოიყენება ისეთი კლასიფიკაციისას როცა მიმოხილვა რეკომენდირებულია თუ არა. გამოიყენება დამალული ან შეუმჩნეველი მონაცემების აღმოსაჩენად. ალგორითმი ღებულობს შემსვლელ მონაცემებად მიმოხილვას და გვაძლევს შედეგად კლასიფიკაციას. კლასიფიკაციაზე მუშაობა ხორციელდება განწყობის ანალიზის სხვადასხვა დონეებზე.

კლასიფიკაციის პრობლემები

ყოველი ადამიანის მიერ წარმოთქმული წინადადება ფლობს რამდენიმე მნიშვნელობას. ხალხი თავის აზრს გამოხატავს რთული საშუალებებით, რიტორიკული მიდგომებით,მაგალითად სარკაზმი,ირონია, გადატანითი მნიშვნელობა, რამაც შესაძლოა გამოიწვიოს განწყობის ანალიზის შეცდომაში შეყვანა. ერთადერთი საშუალება სწორად წარიმართოს პროცესი, არის კონტექსტის გაანალიზება: იმის ცოდნა ,თუ როგორ იწყება აბზაცი, ახდენს გაწყობაზე გავლენას მთლიან წინადადებაში.

განწყობის ანალიზში არსებული აზროვნების უმრავლესობა ხდება კატეგორიულ ჩარჩოში: განწყობა გაანალიზდება კონკრეტულ ჯგუფთან გაერთიანებით, მაგალითად მოცემული წინადადება შესაძლოა იყოს 45% ბედნიერი, 23%სევდიანი ,50% იმედიანი, ეს ციფრები არის წარმოდგენილი 100ის ფარგლებში და ინდივიდუალურ ინდიკატორს წარმოადგენენ იმისას, თუ რა განწყობის მატარებელია X წინადადება.

რატომ არის განწყობის ანალიზი მნიშვნელოვანი?

განწყობის ანალიზი ახდენს ბიზნესის რეალური პრობლემების გადაწყვეტას:

- იგი ეხმარება კლიენტის ქცევის პროგნოზირებაში კონკრეტულ პროდუქტთან მიმართებაში
- ამოწმებს პროდუქტის ადაპტირებულობას
- ავტომატიზირებას უკეთებს მომხმარებლის უპირატესობებს
- მარტივად ახდენს პროცესის ავტომატიზირებას, თუ რამდენად კარგი იყო ფილმი, მისი მიმოხილვების განწყობის ანალიზის საშუალებით

განწყობის ანალიზის სისტემური მოდელი

გაწყობის ანალიზის კლასიფიკაციის ამოცანის პრობლემის გაცნობამდე, საჭიროა ზოგადი ტექსტის კლასიფიკაციის პრობლემის ნათელი წარმოდგენა. ფორმალურად ზოგადი ტექსტის კლასიფიკაციის ამოცანა შეიძლება წარმოდგენილი იქნეს შემდეგნაერად:

- **შემსვლელი:**
  - A - დოკუმენტი d
  - A- ფიქსირებული კომპლექტი კლასებისა C = {c_1,c_2,..,c_n}
  - A fixed set of classes C = {c_1,c_2,..,c_n}
- **გამომსვლელი**: პროგროზირებული კლასი c ∈ C

(დოკუმენტის როლში მოიაზრება აზრები, ფრაზები, ახალი სტატიები,ისტორიები და ა.შ .)

დასასწავლი ნაკრები n დოკუმენტიდან გამოიყურება შემდეგნაერად: *(d_1,c_1), (d_2,c_2),...,(d_n,c_n)* და საბოლოო რეზულტატი იქნება შემსწავლელი კლასიფიკატორი.

განწყობის ანალიზთან მუშაობისას აუცილებელია გვახსოვდეს, რომ ფრაზის ყოველი სიტყვა არ გადმოსცემს ამ ფრაზის აზრს. ასეთი სიტყვებია „მე“, „ვარ“, „ხარ“ და ა.შ ისინი არ არიან არანაერი განწყობის მატარებლები და შესაბამისად არ გააჩნიათ განწყობის ანალიზთან არანაერი კავშირი. ფუნქციის შერჩევისას საჭიროა განისაზღვროს მეტნაკლებად მნიშვნელოვანი ფუნქციები,რომლებიც კლასთან იქნება ახლო კავშირი. მხოლოდ რამდენიმე სიტყვა მონაწილეობს ფრაზიდან კლასიფიკაციის პროცესში და მათი ფრაზიდან ამოკრეფა წარმოადგენს რთულ ამოცანას.

მაგალითი

წარმოდგენილია ფილმის განხილვა:

"I love this movie! It's sweet, but with satirical humor. The dialogs are great and the adventure scenes are fun. It manages to be romantic and whimsical while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I have seen it several times and I'm always happy to see it again....."

ფილმის ეს მიმოხილვა ატარებს პოზიტიურ განწყობას კონკრეტული ფილმიდან გამომდინარე,მაგრამ რომელია ის სიტყვები,რომელიც ამ პოზიტიურობას განსაზღვრავს?

გადავხედოთ ხელახლა:

*"I **love** this movie! It's **sweet**, but with **satirical** humor. The dialogs are **great** and the adventure scenes are **fun**. It manages to be **romantic** and **whimsical** while laughing at the conventions of the fairy tale genre. I would **recommend** it to just about anyone. I have seen it several times and I'm always **happy** to see it **again**....."*

გამუქებული სიტყვები წარმოადგენენ ყველაზე მნიშვნელოვან სიტყვებს რომლებიც განწყობის პოზიტიურ ბუნებას აყალიბებენ.

შედეგი ნაბიჯი არის, წარმოვადგინოთ ეს სიტყვები შემდეგნაირად:

great	2
love	2
recommend	1
laugh	1
happy	1
.	.
.	.
.	.

თითოეული სტრიქონი შეიცავს სიტყვას და მისი შეხვედრის სიხშირეს არსებულ დოკუმენტში.

აქ გამოყენებულია **bag-of-words** წარმოდგენა, რომელიც არის NLP-თვის მნიშვნელოვანი კონცეფცია და პირველი ნაბიჯია ნებისმიერი კლასიფიკაციის პროცესში. **bag-of-words** არ შეიცავს მხოლოდ სპეციფიურ სიტყვებს, იგი შეიცავს სპეციფიურ და უნიკალურ სიტყვებს და მათ სიხშირეებს, დოკუმენტში შეხვედრის მიხედვით.

დოკუმენტში სიტყვების მიმდევრობა შეიძლება არ იყოს მოწესრიგებული და თანმიმდევრობითი, მაგრამ განწყობის ანალიზის დროს ამ მიმდევრობას არ აქვს დიდი მნიშვნელობა.

თუ გვექნება ფილმის რამდენიმე მიმოხილვა,იგივე დოკუმენტი, რომლებიც წარმოდგენილია bag-of-words პრინციპით და დავალაგებთ მათი განწყობის მიხედვით, სასწავლო ნაკრები იქნება შემდეგნაერი:

document	w1	w2	w3	w4	...	wn	sentiment
d1	2	1	3	1		1	positive
d2	1	5	5	5		1	negative
d3	3	8	6	8		2	positive
d4	2	5	1	5		3	positive
d5	3	0	3	0		0	negative
d6	0	0	0	0		0	negative
d7	2	0	0	0		0	positive
d8	9	2	9	2		2	negative
...	...	...	...	...	...	...	...

ამ წარმოდგენას ეწოდება კორპუსი.

ცხრილის ყველა რიგი არის დამოუკიდებელი ვექტორი, რომელიც შეიცავს ინფორმაციას განსაზღვრული დოკუმენტისას, მიმდინარე სიტყვებით და მისი განწყობით. განწყობა ხშირად გამოისახება როგორც (+,-) ან (+ve, -ve). ასევე w1, w2, w3, 34, ..., wn გენერირებულია bag of words-სგან და არ არის აუცილებელი ყველა დოკუმენტი შეიცავდეს მათ.

ამოცანის პრაქტიკული რეალიზაცია

მარტივი განწყობის ანალიზი Python-ში

გამოყენებული იქნება ოფლაინ ფილმების მიმოხილვის კორპუსი, რომელიც წარმოდგენილია წიგნში NLTK , nltk გვაწვდის მონაცემების ნაკრებს. იგი ახდენს ყოველი მიმოხილვის კლასიფიცირებას პოზიტიურად ან ნეგატიურად. თავდაპირველად საჭიროა ჩამოიტვირთოს იგი CMD-ს საშუალებით:

```
python -m nltk.downloader all
```

როცა მონაცემები ჩამოიტვირთება, საჭიროა მოვახდინოთ მათგან ფილმების მიმოხილვების იმპორტირება ,ამის შემდეგ გვექნება დოკუმენტების სია, მონიშნული შესაბამის კატეგორიებად.

```
import nltk
from nltk.corpus import movie_reviews
import random

documents = [(list(movie_reviews.words(fileid)), category)
             for category in movie_reviews.categories()
             for fileid in movie_reviews.fileids(category)]

random.shuffle(documents)
```

შემდეგ უნდა განისაზღვროს დოკუმენტში სიტყვების ექსტრაქტორი, რათა კლასიფიკაციის ალგორითმმა იცოდეს მონაცემების რა ნაწილზე იმუშაოს. ამ შემთხვევაში ყოველი სიტყვისთვის განისაზღვრება ფუნქცია, მიმთითებელი, შეიცავს თუ არა მას დოკუმენტი. რათა გამოითვალოს ფუნქციის რიცხვი,კლასიფიკატორის მიერ დასამუშავებელი, ხდება 2000-ზე მეტი სიხშირის შემხვედრი სიტყვის სიის შედგენა კორპუსისთვის,შემდეგში განისაზღვრება ნიშნების ექსტრაქტორი,რომელიც შეამოწმებს გვხვდება თუ არა ეს სიტყვები მიმდინარე დოკუმენტში.

```
all_words = nltk.FreqDist(w.lower() for w in movie_reviews.words())
word_features = list(all_words)[:2000]

def document_features(document):
    document_words = set(document)
    features = {}
    for word in word_features:
        features['contains({})'.format(word)] = (word in document_words)
    return features
```

ამის შემდეგ საჭიროა მოვამზადოთ ბაიესის კლასიფიკატორი რათა მოხდეს ფილმის მიმოხილვის განწყობის გამოცნობა. ქვემოთ ნაჩვენებია კლასიფიკატორის სიზუსტის გამოთვლა:

```
featuresets = [(document_features(d), c) for (d,c) in documents]
train_set, test_set = featuresets[100:], featuresets[:100]
classifier = nltk.NaiveBayesClassifier.train(train_set)
```

```
print(nltk.classify.accuracy(classifier, test_set))
```

```
0.71
```

კლასიფიკატორი აღწევს 71%იან სიზუსტეს, ნებისმიერი პარამეტრის კორექტირების გარეშეც კი.

როგორც აღმოჩნდა, ამ ნაკრებში, ისეთი მიმოხილვები, სადაც მოხსენიებულია "Illogical" ,არის 8-ჯერ უფრო მეტად მოსალოდნელი, რომ ის იქნება ნეგატიური ვიდრე პოზიტიური, ვიდრე ის რომ მიმოხილვები, სადაც გვხვდება "Captures", იქნება 6-ჯერ უფრო პოზიტიური.

```
    contains(winslet) = True              pos : neg    =      8.4 : 1.0
  contains(illogical) = True              neg : pos    =      7.6 : 1.0
   contains(captures) = True              pos : neg    =      7.0 : 1.0
     contains(turkey) = True              neg : pos    =      6.5 : 1.0
     contains(doubts) = True              pos : neg    =      5.8 : 1.0
```

დასკვნა

ნაშრომში განხილულია ერთერთი ყველაზე აქტუალური და მნიშვნელოვანი კლასიფიკაციის მეთოდი განწყობის ანალიზი, რომელიც დღეისთვის სარგებლობს გაზრდილი ინტერესით და საკმაოდ ხშირად გამოიყენება პრაქტიკაში.

ნაშრომში აღწერილია მეთოდები მანქანური სწავლების მიდგომიდან და განხილულია თითოეული მეთოდის: სწავლება მასწავლებლით და მასწავლებლის გარეშე მუშაობის პრინციპი. ამ მეთოდებიდან ნაშრომში ნაჩვენებია მხარდამჭერი ვექტორების მანქანების, ნეირონული ქსელების,გადაწყვეტილებათა ხეების, თუ ბაისესის ალგორითმის მუშაობის პრინციპები.

მომხმარებლების მხრიდან აზრის გამოხატვა სპეციალიზირებულ საიტებზე მომსახურებისა და პროდუქტის შესაფასებლად, ასევე სოციალური ქსელების პლათფორმების საშუალებით გახდა ერთერთი ძირითადი საშუალება კომუნიკაციისა, ინტერნეტის განვითარების წყალობით ბოლო პერიოდში, სწორედ ამ ინფორმაციის დამუშავებასა და ანალიზს ემსახურება წარმოდგენილი თემის შინაარსი და ნაშრომში აღწერილია თუ როგორ მიმდინარეობს კლასიფიკაციის პროცესი სხვადასხვა მეთოდის გამოყენებისას.

ლიტერატურა

[1] A Survey on Sentiment Analysis Algorithms for Opinion Mining, International Journal of Computer Applications (0975 – 8887)Volume 133 – No.9, January 2016

[2] blog.algorithmia.com/introduction-sentiment-analysis

[3] lct-master.org/files/MullenSentimentCourseSlides.pdf

[4] theappsolutions.com/blog/development/sentiment-analysis/

[5] datacamp.com/community/tutorials/simplifying-sentiment-analysis-python

[6] Real Time Sentiment Classification Using Unsupervised Reviews E.Divya