ტექსტის ანალიზი და ყალბი ინფორმაციის აღმოჩენა

გურამი ასანიშვილი

თ.ს.უ ზუსტ და საბუნებისმეტყველო მეცნიერებათა

ფაკულტეტის დოქტორანტი

ხელმძღვანელი: მანანა ხაჩიძე

თბილისი 2019

სარჩევი

აბსტრაქტი

ინტერნეტის და სოცილური ქსელის განვითარებამ გამოიწვია ადამიანების ყოველდღიური დამოკიდებულება ინფომარციაზე და ინფორმაციის წყაროებზე, მისი საშუალებით ის ეცნობა ახალ ამბებს, ყიდულობს პროდუქტებს, გეგმავს მოგზაურობას და ა.შ . სამწუხაროდ ამ ყველაფერმა დადებითი მხარეების გარდა რომელიც ადამიანს ეხმარება ცხოვრებაში, იჩინა თავი უარყოფითმა მხარემაც, კერძოთ ბოლო პეროდში რთული გახდა ერთმანეთისგან განვასხვავოთ სანდო და ყალბი ინფორმაცია. ყალბი ინფორმაციის სიმრავლემ გამოიწვია დიდ მასებზე ურთიერთობა და შემოიჭრა ჩვენს ყოვედღიურ ცხოვრებაში, ცდილობენ ჩვენს გადაწყვეტილებებზე ზეგავლენას და არასაიკეთოდ გამოყენებას.

ამ ნაშრომის მიზანია, ჩამოვაყალიბოთ პრობლემა, რომელიც გამოწვეულია ყალბი ინფორმასიის დიდი დოზით არსებობით ინტერნეტში, დავსბათ ამოცანა თუ როგორ შეიძლება ამ პრობლემასთან გამკლავება და შევიმუშავეთ მეთოდი ან მეთოდები რომელებიც დაგვეხმარება დროულად აღმოვაჩინოთ და შესაბამისად შემდგომში თავი დავიცვათ ყალბი ინფორმციისგან.

Abstract

Internet and social networks development caused a person's dependence on information and on sources of information, with their help, we learn the news, buys groceries, plans a trip and so on. Unfortunately it has a lot of goodness that helps a person in life and a denial side. The last time it has become difficult to distinguish between pravdivie and false information from each other. A lot of false information causes an impact on society and try to use them for non-sour purposes. The purpose of this article is to understand the problem that causes false information, our task is to find a way to deal with this problem, so that we could detect and prevent fake information.

აქტუალურობა

21-ე საუკუნეში თანამედროვე მსოფლიო ცხოვრობს ინფორმაციულ ერაში. არ არის ადვილი განისაზღვროს ელექტრონული მონაცემების მოცულობა,რადგან ელექტრონული ინფორმაციის მოცულობა საკმაოდ სწრაფად იზრდება და მათი მოცულობის წინასწარმეტყველება დროის გარკვეული პერიოდისთვის შეუძლებელია IDC (International Data Corporation)-ის გათვლებით "ციფრული სამყაროს მოცულობა" 2006 წელს შეადგენდა 0,18 ზეტაბაიტს[1], 2011 წელს კი 1,8 ზეტაბაიტს და მონაცემების მოცულობაც დღითიდღე უფრო იზრდება. არსებობს რამდენიმე წყარო რომ ზოგადი წარმოდგენა შეგვექმნას მონაცემების რაოდენობასთან დაკავშირებით, მაგალითად:

- ნიუ-იორკის საფონდო ბირჟა ყოველღდე აგენერირებს რამდენიმე ტერაბაიტ მონაცემს;
- Facebook-ის მონაცემთა საცავი ყოველდღე იზრდება 500 ტერაბაიტით;
- პროექტი Internet archive უკვე ინახავს 2 პეტაბაიტ[2] მონაცემებს და ყოველთვიურად 20 ტერაბაიტი ემატება;
- დიდ ადრონულ კოლაიდერში ჩატარებულ ექსპერიმეტებს წამში შეუძლია დააგენერიროს დაახლოებით 1 პეტაბაიტი.

ინფორმაციის შენახვა და მისი შემდგომი დაცვა ინფორმაციული ტექნოლოგიების ერთ-ერთი უმთავრესი შემადგენელი ნაწილია. ყოველ წამს კომპანიები თუ კერძო პირები ქმნიან დიდი მოცულობის და რაოდენობის ინფორმაციას. ეს ინფორმაცია აუცილებელია შევინახოთ, დავიცვათ, გავუკეთოთ ოპტიმიზაცია და ვმართოთ. ბოლო ორი ათეული წელია ინფორმაციის დიდი ნაკადების და მოცულობების გამო უფრო და უფრო რთულდება მონაცემების შენახვის, უსაფრთხოების, ოპტიმიზაციის პრობლემების გადაწყვეტა. ამას ისაც ემატება რომ

[1] 1 მილიადი ტერაბაიტი
[2] 1000 ტერაბაიტი

დღითიდღე იზრდება ქსელში ჩართული მომხმარებლების რიცხვიც, შესაბამისად მონაცემების რაოდენაობაც პროპორციულად იზრდება.

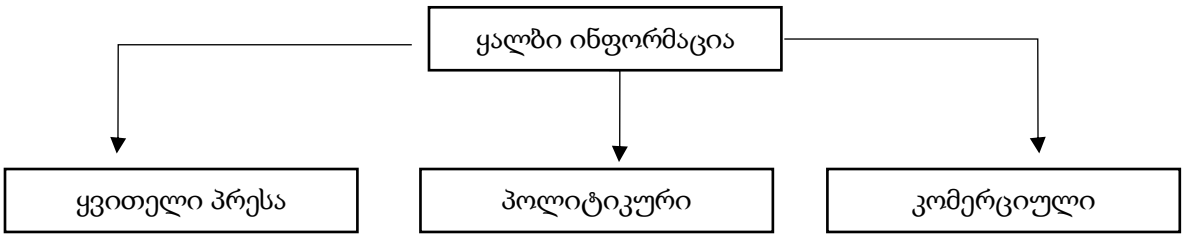
ქსელური ტექნოლოგიების სწრაფმა განვითარებამ გამოიწვია სოციალური მედიის და სოციალური ქსელების განვითარება. ბოლო პერიოდში კომერციული თუ პოლიტიკური მიზნებისთვის ფართოდ გამოიყენება ყალბი ინფორმაციის გავრცელება რაიმე კონკრეტული ადამინთა ჯგუფზე, რომლებზეც შესაძლებელია ამა თუ იმ გავლენის მოხდენა. სოციალური ქსელის მეშვეობით ადამიანების ჯგუფებზე შესაძლებელია გარკვეული ზემოქმედების განხორციელება კომერციული ან პოლიტიკური სარგებლის მიღების მიზნით ყალბი (ცრუ) ინფორმაციის გავრცელების გზით. ყალბი ინფორმაციის გავრცელებით თავისუფლადაა შესაძლებელი ადამიანთა მასის შეცდომაში შეყვანა. დღევანდელობის ერთ-ერთი მთავარი გამოწვევაა სოციალურ ქსელებში, საინფორმაციო საიტებზე და სრულიად ინტერნეტ სივრცეში ინფორმაციის სანდოობის სწორი განსაზღვრა. დღითიდღე უფრო და უფრო მნიშვნელოვანი ხდება ახალი მეთოდოლოგიების და ალგორითმების დანერგვა ყალბი ინფორმაციის აღმოსაჩენად.

მონაცემთა ანალიზის კერძოდ კი დიდი მონაცემების ანალიზზე დაყრდნობით შესაძლებელია ახალი ავტომატიზირებული მოდელის შექმნა რომელიც დაგვეხმარება ყალბი ინფორმაციის ამოჩენაში და პრევენციაში.

იმის გამო რომ ყალბმა და დაუზუსტებელმა ინფორმაციამ შესაძლოა დიდი ნეგატიური გავლენა მოახდინოს საზოგადოებაზე ძალზედ მნიშვნელოვანია მსგავსი სისტემების შექმნა, ტესტირება და გამოყენება.

ამოცანა

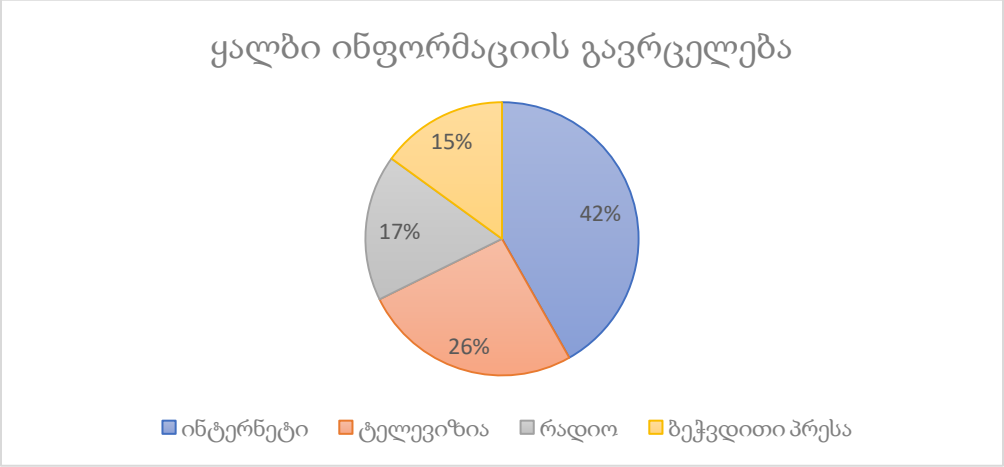
ადრეულ პერიოდში როდესაც ინფორმაციის მიღების წყაროს წარმოადგენდა ბეჭდითი მედია, ყალბი ინფორმაცია განეკუთვნებოდა ყვითელი პრესის ტიპს, რომლის ძირითადი მიზანი იყო მკითხველის დეზინფორმაცია, მოტყუება და რაც შეძლება მეტი გაზეთის გაყიდვა. მას შემდეგ რაც გამოჩნდა ინტერნეტი, ონლაინ მასმედია და სოცალური ქსელი, ყალბმა ინფორმაციამ არ დაკარგა ფუნქცია და მისი მოქმედების არეალი გაიზარდა და მოიცვა მსოფლიო. მისი გამოყენების ფუქციები გაზარდა და გარდა დეზინფორმაციისა, ყვითელი პრესისა შეიძინა სარეკლამო ხასიათიას მიზნებიც, რომელიც მიმართულია რაც შეიძლება მეტი პროდუქტის გაყიდვაზე ასევე სხვადასხვა სახის პოლიტიკურ თუ კომერციულ მიზნები



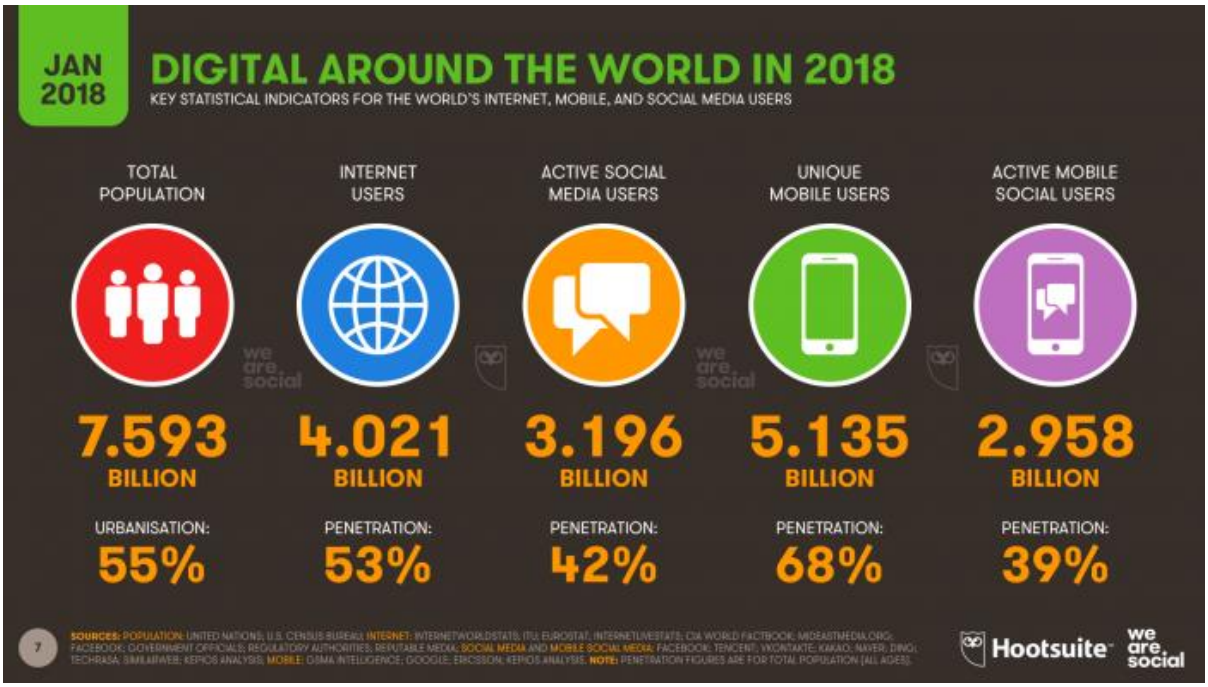
ყალბი საინფორმაციო სტატიების ან ახალი ამბების გამოყენებით და მარტივი სატყუარა სიტყვების გამოყენებით შესაძლებელია ონლაინ მომხმარებლებზე სრული ზემოქმედების მოხდენა, მათ აზრების არასწორედ ჩამოყალიბება, მაგალითად 2016 წლის ამერიკის შეერთებული შტატების საპრეზიდენტო არჩევნების დროს მრავლად ვრცელდებოდა მსაგავსი ინფორმაციები სოციალურ ქსელებში კერძოდ Facebook-ში და Twitter-ში, რომლებსაც შეეძლოთ მნიშვნელოვანი გავლენა მოეხდინათ არჩევნების შედეგებზე.

ბოლო წლების სტატისტიკის მიხედვით ინფორმაციის ნაკადის 41,8% მოდის ყალბ ინფორმაციაზე რომელიც თავის პიკს აღწევს სწორედ არჩევნების პერიოდში.

სტატისტიკური განაწილების მიხედვით კი ეს უფრო მეტია ვიდრე სატელევიზიო, რადიო, ბეჭდით პრესაში გავრცელებული ყალბი ინფორმაციები.



თუ გავითვალისწინებთ რომ 2018 წლის მონაცემებით სოციალურ მედიას 3,196 მილიარდი ადამიანი იყენებს და მომხმარებლების რიცხვი ყოველწელს 12%-ით იზრდება(https://wearesocial.com/blog/2018/01/global-digital-report-2018), მათზე მოსული 41,8% ყალბი ინფომაცია საკმაოდ დიდი რიცხვია.



მომხმარებელთა ზოგირეთი ნაწილი მიიჩნევს რომ ყალბი ინფორმაცია მსგავსია ე.წ. სპამების. მათ საერთო არაფერი არ აქვთ, რადგან სპამი როგორც წესი არსებობს პირად მეილებში ან რაიმე კონკრეტულ ვებსაიტზე და აქვთ მხოლოდ ლოკალური გავლენა მომხმარებლების მცირე ნაწილზე, ხოლო ყალბი ინფორმაციის ზეგავლენის არეალი შეიძლება იყო საკმაოდ დიდი ვინაიდან სოციალური ქსელები წარმოადგენენ ინფორმაციის გავრცელების მთავარ საშუალებას და ყალბი ინფორმაციის გავრცელების შემთხვევაში ადამიანების დიდ ჯგუფს მისდის შესაბამისი ინფორმაცია, რაც უფრო მეტად ამძაფრებს მის მოქმედებას. გარდა ამისა, სპამის პასიური მოქმედებისგან გასხვავებით, რომელიც მხოლოდ შეიძლება მეილზე მივიღოთ ან რეკლამის სახით საიტზე გამოჩნდეს, ყალბი ინფორმაცია უფრო აქტიურია, რადგან მას მომხმარებელი ეძებს, იღებს, აზიარებს, განიხილავს და რაც უფრო მეტი ადამიანი აქტიურად განიხილავს ამათუ იმ ინფორმაციას მით უფრო ძნელია ყალბი ინფორმაციის განასხვავება ნამდვილი ინფორმაციისგან, რამდენად შეესაბამება ის სინამდვილეს და დადგინდეს რელევანტურობის ხარისხი.

ზემოთ ხსენებული მახასიათებლების მიხედვით, ყალბი ინფორმაციის სიჭარბე ართულებს ინფორმაციის ნამდვილობის დადგენის საკითხს და ქმნის ახალ გამოწვევებს. გარდა ამისა, დიდი მოცულობის ინფორმაციაზე შემცირდეს ყალბი ინფორმაციის ზეგავლენა.ასევე მნიშვნელოვანია სტატიის ან ინფორმაციის წარმომავლობის საკითხი, რადგან სანდო პირველწყაროდან მოსული ინფორმაციას უფრო მაღალი საიმედოობის ხარისხი აქვს, ვიდრე იმ პიროვნებისგან რომელის სტატიების უმრავლესობა არ არის გადამოწმებული. მსგავსი კორელაცია შეიძლება დაფიქსირდეს ახალ ამბების სტატიებსა თუ თემებში.

ყალბი ინფორმაციის აღმოჩენა არ არის ადვილი, ამის მიზეზი შემდეგია:

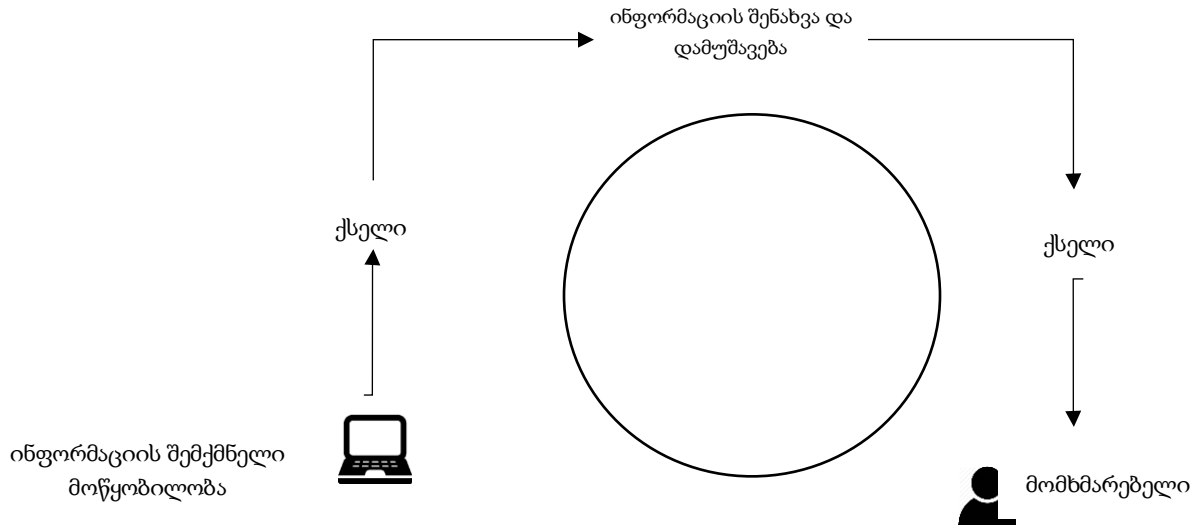
- **პრობლემის ფორმულირება:** ყალბი ინფორმაციის აღმოჩენის საკითხი, პარამეტრების დადგენა, რითაც განისაზღვრება ან შეიძლება განისაზღვროს ინფორმაციის საიმედოობა;

- **ტექსტური ინფორმაციის გამოყენება**: საინფორმაციო სტატიების შემქმნელები და შექმნის ობიექტები, შემავალი ინფორმაციის კრებული შეძლება მიღებული იქნას სოციალური ქსელებიდან. იმითვის რომ მოხდეს დეტექცია საჭიროა ეფექტური მოდელი რომელიც აღმოაჩენს და შეისწავლის ამ პროცესსს.
- **ჰეტეროგენური ინფორმაციის სინთეზი**: გარდა იმისა რაც ნახსენები იყო სანდოობის ხარისხების შესახებ სტატიების, ავტორების და თემებზე რომლებსაც აქვს ძლიერი კორელაცია რომელზაც მიუთეთებს ავტორისა და სტატიის ურთიერთობაზე შესაძლია ეფექტურად ჩაირთოს საბაზისო კორელაციებიც რომლებიც დაგვეხმარება ყალბი ინფორმაციის ზუსტად დეტექციის შესწავლაში.

იმისთვის რომ გადაიჭრას ზემოხსენებული პრობლემა უნდა შეიქმნას გარკვეული ალგორითმი რომელიც გამოავლენს ინფორმაციის საიმედოების სანიშნეებს, რომლის მიხედვითაც შესაძლებელი იქნება გამოვლინდეს სანდოობა.

კონცეფციების და ფორმულირების გასაზღვრა, რომელიც საჭიროა ყალბი ინფორმაციის აღმოსაჩენად

როგორც წესი, ინფორმაცია შექმნის მომენტში განთავსებულია მომხმარებლის მოწყობილობებზე როგორიცაა: მობილური ტელეფონები, კამერები, პლანშეტები, ნოუთბუკები და ა.შ იმისთვის რომ მოხდეს ინფორმაციის გაცვლა საჭიროა ეს მოწყობილობები დაკავშირებული იყვნენ ქსელთან, რომლის მეშვეობითაც ახდენენ მონაცემთა საცავების გავლით ინფორმაციის გაზიარებას.



ზემოთ ნაჩვენებ ციკლის მიხედვით, ინფორმაციის შემოწმება მის ვალიდურობაზე და სანდოობაზე უნდა ხდებოდეს „ინფორმაციის შენახვა და დამუშავების“ ეტაპზე იმისთვის რომ ზუსტად განვსაზღვრთ კონცეფცია და ჩამოვაყალიბოთ ფორმულირება შემოვიტანოთ აღნიშვნები:

- ახალი სტატიების ნაკრები, რომლებიც ქვეყნდება ინტერნეტში ან სოციალურ ქსელში აღვნიშნოთ N=$\{n_1, n_2, \dots n_m\}$, სადაც n_i€ N რაიმე ახალი სტატიაა;
- ნამდვილი ინფორმაცია აღვნიშნოთ Y-ით რომელიც n_i^c€ Y n_i€ N-თვის;

- ახალი თემა- ჩვენ შეგვიძლია ქსელში არსებული საინფორმაციო თემების ნაკრები წარმოვადგინოთ როგორც S={$s_1, s_2, \dots s_m$};
- ავტორები რომლებიც ქმნიან ახალ სტატიებს და ახალ ინფორმაციას აღვნიშვნოთ A={$a_1, a, \dots a_m$}-თი;

პრობლემა რომელიც ეხება ინფორმაციის სანდოობას შეიძლება ჩამოყალიბდეს შემდეგნაირად : F:AUNUS→Y სადაც F არის სწავლების ფუქნცია რომელიც ადგენს ინფორმაციის ნამდვილობას.

მონაცემთა ანალიზი

ინფორმაციის და მოწყობილობების რომლებიც ამ ინფორმაციას აგენერირებენ რაოდენობის ზრდის ფონზე მატულობს ინფორმაციის და მონაცემების მართვის პრობლემები. პრაქტიკამ აჩვენა რომ პრობლემები რომლებიც ეხება დიდ მონაცემებს შეიძლება დაიყოს რამდენიმე ნაწილად:

- **ციფრული სამყაროს სწრაფი ზრდა:** ინფორმაციის იზრდება გეომეტრიული სისწრაფით. მონაცემები კოპირდება სხვადასხვა სერვერებზე იმისთვის რომ მოხდეს ინფომაციაზე სწრავი წვდომა.
- **გაზრდილი ინფორმაციული დამოკიდებულება:** ინფორმაციის სტრატეგიული გამოყენა, თამაშობს დიდ როლს ბიზნესის წარმატების მიღწევაში და ბაზარზე ზრდის კონკურენტუნარანობას.
- **ინფორმაციის ღირებულების ცვლილება:** ინფორმაცია ღირებულია დღეს, ის ხვალე შეიძლება უფრო ნაკლებად ფასეული გახდეს. ინფორმაციის ფასი ხშირად იცვლება დროის განმავლობაში.

ინფორმაციის დამუშავება და მონაცემთა ანალიზი მოიცავს 4 ძირითად ნაწილს, ესენია: მონაცემთა სანდოობის ანალიზი ტექსტის შინაარსიდან გამომდინარე, თემის ნამდვილობის ანალიზი, ავტორის სტატიები და მისი წინა ჩანაწერები. ეს პარამეტრები მოქმედებს მხოლოდ სტატიების დამუშავების დროს, იმ შემთხვევაში თუ სანდოობაზე

ვამოწმებთ ვებსაიტებს შესამოწმებელი პარამეტრების რაოდენობა იზრდება და შემოწმებაც უფრო კომპლექსური ხდება.

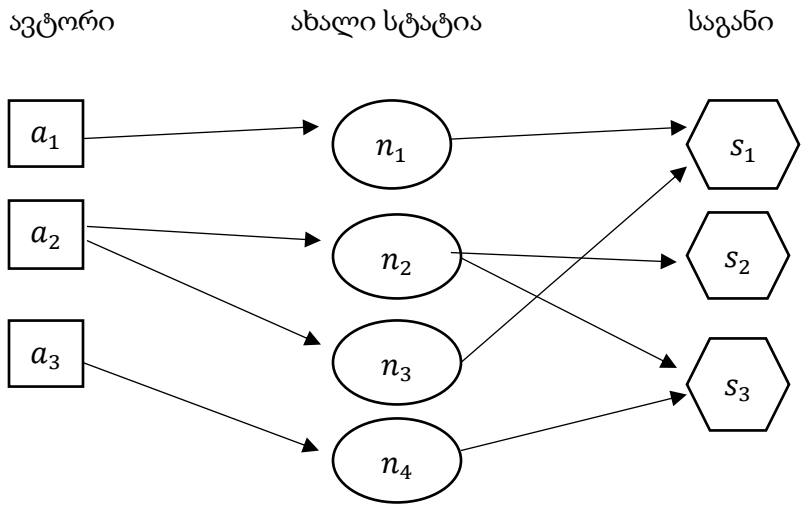
- **მონაცემთა სანდოობის ანალიზი**- მონაცემთა სანდოობის ანალიზი ტექსტის შინაარსიდან გამომდინარე ძირითადად მოიცავს ტექსტის სემანტიკურ ანალიზს იმ სიტყვებზე რომლებიც შეიძლება ნიშნავენ სტატიას როგორც სანდო ან ყალბი. ამისთვის საჭიროა შემოვიღოთ ტერმინები და აღნიშვნა „სანდო“, „უმეტესად სანდო“ ან „ნახევრად სანდო“ სტატიის დამუშავების შემდეგ რეიტინგულად ეს აღნიშვნები შესაძლოა შეიცვალოს „ყალბი“ ან „უმეტესად ყალბად“. ყველაზე მთავარაია განისაზღვროს იმ სიყვების ნაკრები რომელიც საშუალებას მოგვცემს აღმოვაჩინოთ ან დავადგინოთ სანდოა თუ არა ინფორმაცია.
- **თემის ნამდვილობის ანალიზი**- თემის ნამდვილობის ანალიზისთვის საჭიროა განისაზღვროს ტოპ თემების სია, რომელიც დროის გარკვეულ პერიოდში არის აქტუალური ან დამყარდეს კავშირი არსებულ თემასა და აქტუალურ თემას შორის.
- **ავტორის სტატიები და მისი წინა ჩანაწერები**- შედარებით მარტივია ანალიზითვის ავტორის წინა პუბლიკაციები, რათა განისაზღვროს მის მიერ გამოქვეყნებული სტატიების რა პროცენტული წილი იყო სანდო ან ყალბი, ამის მიღეხედვით ადვილია დადგინდეს ავტორის სტატიის სანდოობის ხარისხი ან საჭიროებს თუ არა ის დამატებით შემოწმებას.

სწავლის მოდელი

სწავლების დანიშნულებაა რაიმე სისტემის შემავალი და გამომავალი სიდედეების, მდგომარეობების დაკვირვებათა ნაკრებიდან შედგეს ამ სისტემის მოდელი, რომლის საშუალებითაც შესაძლებელი იქნება სხვდასხვა ამოცანების გადაწყვეტა. ინტელექტუალური სისტემები ეფუძვნება მანქანურ სწავლებას აქედან გამომდინარე უდიდესი მნიშვნელობა აქვს შესაბამისი მოდელის აგებას. ასეთი ტიპის სისტემები შეიძლება გათვლილი იყოს მხოლოდ კონკრეტული ტიპის ინფორმაციაზე. არსებობს უამრავი სხვადასხვა ტიპის ალგორითმი, რომლებიც გამოიყენება ასეთ სისტემებში, მაგ: კლასიფიკაციის, კლასტერიზაციის ტიპის ალგორითმები. როგორც წესი მათი ცალ ცალკე გამოყენება ვერ იძლევა ეფექტურ შედეგს, აქედან გამომდინარე საჭირო ხდება ასეთი ტიპის ალგორითმების დაჯგუფება, ერთმანეთში კომბინაცია.ინტელექტუალური სისტემა შეიძლება მრავალნაირი იყოს, აქედან გამომდინარე სრულიად შესაძლებელია შემავალი ინფორმაცია იყოს ნებისმიერი სახით წარმოდგენილი. ინფორმაციის ტიპები მრავალნაირია, შესაბამისად თავდაპირველად მოწოდებული შემავალი ინფორმაცია შეიძლება იყოს სხვადასხვა ტიპის: ტექსტური, რიცხვითი და ა.შ. თუმცა ინტელექტუალური ტიპის სისტემა ძირითადად ემყარება ორი ტიპის ინფორმაციას, რომელთაგან პირველია რაოდენობრივი ხოლო მეორე ხარისხობრივი, ყოველი ინფორმაცია საჭიროებს დამუშავებას და შედეგის მიღებას. სწორედ ამიტომ, მანქანურ სწავლებაში გამოიყენება სხვადასხვა დანიშნულების ალგორითმები, რომელიც საშუალებას გვაძლევენ ეფექტურად დავამუშავოთ ინფორმაცია, რათა მოხდეს მათი სწორად წარმოდგენა და გაანალიზება. რეფერატში განხილულია ერთ-ერთი ასეთი მანქანური სწავლების მონაცემთა დამუშავების ალგორითმი.

იმითვის რომ, განსხვავდეს ნამდვილი და ყალბი ინფორმაცია ერთმანეთისგან საჭიროა დასწავლის ალგორითმის შემუშავება, რომელიც კონკრეტული მარკერების საფუძველზე განსაზღვრავს ინფორმაციის სანდოობას.

მოდელი რომელიც წარმოადგენს ამ ალგორითმს მდგომაარეობას შემდეგში: გამომავალი ინფორმაცია წარმოადგენს ვექტორს რომელიც მიმართულია ავტორიდან ახალი სტატიისკენ და საგნისკენ.



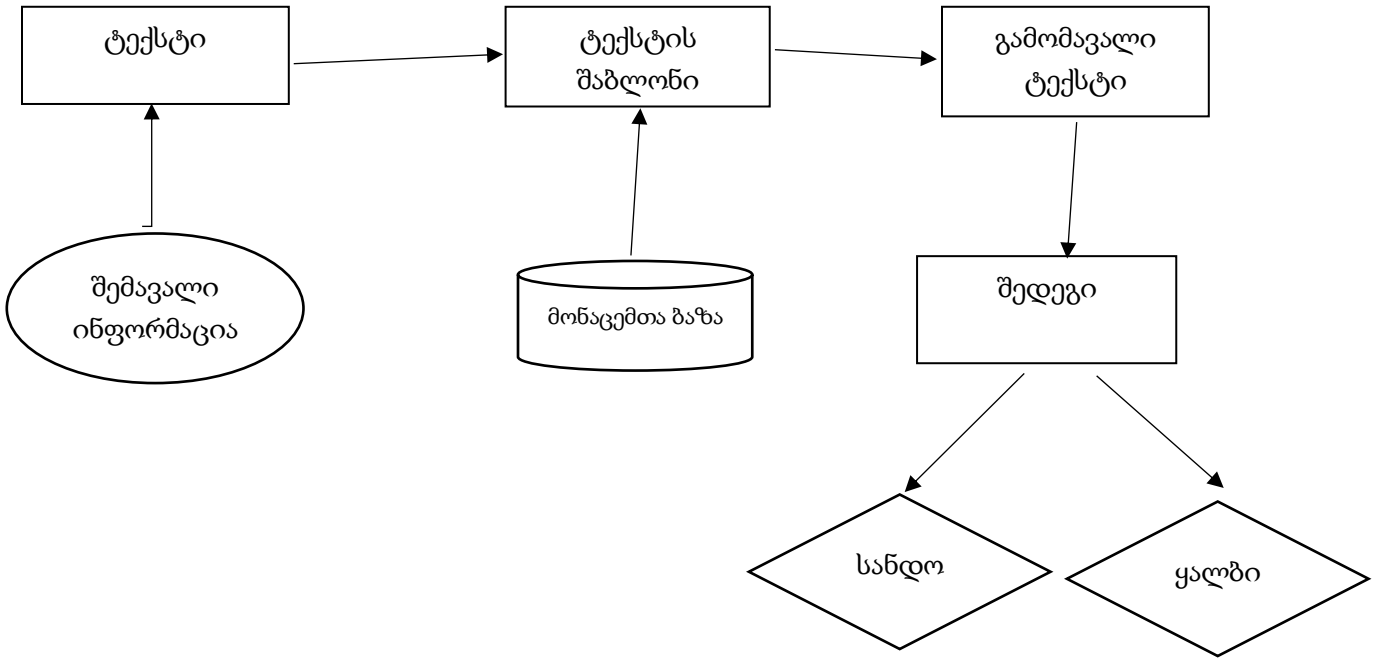
ალგორითმი საძიებო სისტემის დახმარებით ქსელში ეძებს ინფორმაციას ავტორის შესახებ განსაზღვრავს მის წინაისტორია და შემოგვთავაზებს სანდოობის ხარისხს W, ტექსტის სემანტიკური ანალიზის შემდეგ ქსლეში ეძებს მსგავს სტატიებს და საგასაღებო სიტყვების (key word) მეშვეობით განსაზღვრავს ტექსტის სანდოობის ხარისხს H, ასევე მოწმდება ტექსტი საგნის შინაარსის აუთენტურობა და მისი ხარისხი T.

საბოლოოდ ალგორითმი ამ პარამეტრების გამოყენებით ადგენს ინფორმაციის საბოლოო სანდოობას L, სადაც $L(n)=\sum W, H, T/100$

ავტორის, სტატიის და საგნის სანდოობის ხარისხი მდებარებობს [0,5] შუალედში, სადაც 0 არის ყალბი ხოლო 5 მაქსიმალურად სანდო ინფორმაცია.

ყალბი ინფორმაციის დამუშავაბა ემყარება ტექსტის დამუშავებას, რომლის მიხედვითაც დგინდება ინფორმაციის საიმედოება. ძირითადი იდეა მდგომარეობს შემდეგში: იმისთვის რომ განისაზღვროს ყალბი ინფორმაცია და ალგორითმა შეძლოს

საიმედო ტექსტის აღთქმა, სწავლის პროცესში მას სპეციალურად ვაანალიზებინებთ ტექსტებს რომლებიც შეიცავს ცრუ ინფორმაციას რათა ჩამოყალიბდეს ტექსტის ანალიზის შაბლონები (თარგები), რომელთანაც შემდგომში მოხდება სხვა ტექსტების შედარება. ყოველი ახალი შედარების დროს იქმნება მონაცემთა ბაზა, გამოყოვა სიტყვები, წინადადებები და მათ შორის დგება სემანტიკური კავშირები, რომლის საშუალებითაც ალგორითმი სწავლობს ტექსტი საიმედოობის დადგენას. ყოველი ახალი ტექტის ანალიზი საშუალუებას მოგვცემს უფრო დავხვეწოთ და გავაუჯობესოთ ინფორმაციის. მოდელი შეიძლება გამოიხატო გრაფიკულად:



ტექსტის ანალისზი დროს ყურადღება ექცევა ისეთ მნიშვნელოვან პარამეტრებს როგორიცაა: ავტორის მონაცემები: მისი წინა სტატიების განსაზღვრისთვის, საგანი ან ტემა თუ რის შესახებ არის დაწერილი ტექსტი, ტექსტში საგასაღებო სიტყვებს შორის კავშირი რომელიც დაგვეხმარება განვსაღვროთ რამდენად შეესაბამება იგივე შინაარსის სხვა სტატიების შაბლონს.

ასევე შესაძლებელია რეგრესიის ალგორითმის გამოყენებით ინფორმაციის კლასიფიკაცია ორგანზომილებიან სივრცეში, რაც ვიზუალურად უფრო წარმოგვიდგენს სანდოობის ხარისხის განაწილებას სიბრტყეზე.

დასკვანა

მიუხედავად იმისა რომ ჯერჯერობით ეს მოდელი თეორიულად დონეზეა, მსგავის სისტემის დამუშავება და შექმნა საშუალებას მოგვცემს გავზარდოთ ინფორმაციის სანდოობა, შევამციროთ ყალბი ინფორმაციის ზეგავლენა საზოგადოებაზე და რაც იწვევს არასწორი აზრის ჩამოყალიბებას, შესაბამისად არამართებული გადაწყვეტილებების მიღებას

გამოყენებული ლიტერატურა

- J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei. Line: Large-scale information network embedding.
- Q. Zhan, J. Zhang, S. Wang, P. Yu, and J. Xie. Influence maximization across partially aligned heterogenous social networks.
- B. Wu and B. Davison. Detecting semantic cloaking on the web.
- Sliva, S. Wang, J. Tang, and H. Liu. Fake news detection on social media: A data mining perspective.
- Machine Learning by Tom M Mitchell
- R. Al-Rfou, and S. Skiena. Deepwalk: Online learning of social representations.