Экстремистік контентке қол жеткізу және оларды жинақтау әдістері

Әлеуметтік медиа идеялармен және қол жетімді және ғаламдық онлайн-коммуникациялармен алмасу алаңын ұсынады. Көптеген адамдар әлеуметтік медианы әртүрлі себептерге қолдау көрсету немесе қарсы пікір білдіру үшін пайдаланады, бұл пайдаланушы мазмұнының көп бөлігі мәтіндік ақпарат болып табылады. Күнделікті өмірдің әртүрлі тақырыптары бойынша нақты уақыт режимінде өз пікірлерін жариялайтын адамдар туралы көптеген мәліметтер болғандықтан, үкіметке дұрыс шешім қабылдау немесе қоғамдық пікірді бақылау үшін пайдалы болуы мүмкін деректерді жинау және талдау үшін зерттеу жүргізген жөн. Әлеуметтік желілерде қол жетімді деректер ықтимал террористік немесе радикалды топты анықтауға тырысқанда қызығушылық тудыруы мүмкін ақпараттың бір түрі ғана болып табылады. Мысалы, радикалдар әлеуметтік медианы зорлық-зомбылыққа деген радикалды көзқарастары бар фанаттарға ықпал ету және ынталандыру үшін қолданған бірнеше жағдайлар бар. Сондықтан, бұл дәрісте біз пікірлерді талдауға және интернеттегі әлеуметтік медиа деректеріндегі экстремистік мазмұнды анықтауды қарастырамыз.

Social networks

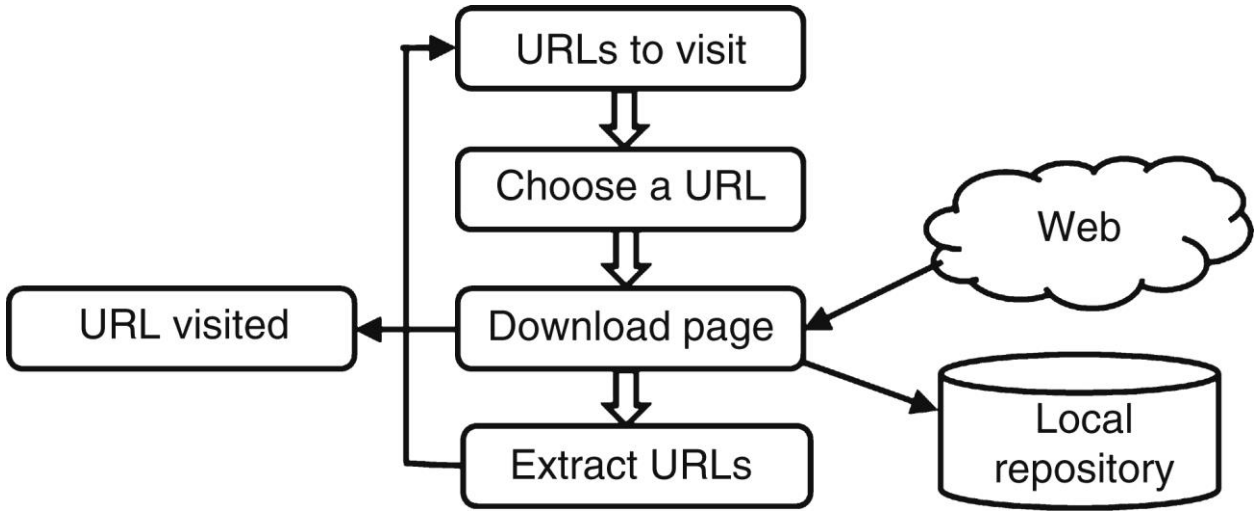
Әлеуметтік медиа оларды қолданудың қарапайымдылығына байланысты көпшілік арасында танымал болды. Миллиондаған хабарламалардың күнделікті пайда болуымен, әлеуметтік медиа, қосымшалар немесе қолданушыларға интернетте өзара әрекеттесуге мүмкіндік беретін қосымшалар соңғы жылдары маңызды бола бастады. Қол жетімді сайттар пайдаланушылар нақты уақыт режимінде түсініктеме бере алатын сайттардан бастап әлеуметтік медиа қызметтеріне дейін. Әлеуметтік медианы талдаудың артықшылығы - деректер көпшілікке қол жетімді болады, өйткені қатысушылар осы ортаға белсенді және пассивті пайдаланушылардың толық көрінісінде қатысады. Бұл электрондық пошта немесе Интернет және телефония сияқты басқа сандық байланыс құралдарымен пайдаланушылармен өзара әрекеттесуден ерекшеленеді, мұнда талдаушылар пайдаланушылардың жеке өмірін қорғау үшін тиісті заңдарды ұстануы керек.

Әлеуметтік медиа біздің пікірімізді білдірудің танымал құралына айналды, сондықтан оларды іс жүзінде құнды идеяларды, тіпті ең радикалды идеяларды алуға болатын тірі фокус-топ ретінде пайдалануға болады. Саяси сайлау, дін немесе реалити-шоудың жеңімпазы туралы айтатын болсақ та, әлеуметтік медиа қоғамдық пікірді анықтауда маңызды рөл атқара алады. Бұл адамдар нақты уақыт режимінде өз көзқарастарымен бөлісетін, ұйымдар мен барлауға қоғамдық пікірге баға жетпес көзқарас беретін платформа. Әлеуметтік медиа біздің күнделікті өмірімізде маңызды рөл атқаруды жалғастыра отырып, оларды әлеуметтік көңіл-күйді түсіну және оған жауап беру тәсілі ретінде қолдануға болмайды.

Facebook

Facebook-2004 жылдың ақпан айында іске қосылған, Facebook корпорациясының басқаруындағы және жеке меншігіндегі әлемдегі ең танымал әлеуметтік желі. Оның мақсаты - адамдарға әлемді ашық және өзара байланыстыруға мүмкіндік беру. Facebook қолданушылары Жеке немесе қоғамдық профиль жасай алады, басқа пайдаланушыларды достарға қосады және хабарламалармен бөлісе алады (профиль туралы ақпаратты жаңарту кезінде автоматты хабарландыруларды қоса). Сонымен қатар, пайдаланушылар фотосуреттерімен, бейнелерімен, мәртебесімен, жаңалықтарымен, жазбаларымен бөлісе алады және достарына түсініктеме бере алады. Сонымен қатар, пайдаланушылар топтарға қосыла алады немесе жұмыс орнына, мектепке немесе тіпті брендке фанаттар парағын жасай алады немесе іс-шаралар ұйымдастыра алады. Алайда, бұл платформа экстремистік пайдаланушыларды әртүрлі заңсыз әрекеттерді жасауға итермелейтіні анық.

Facebook Graph API-бұл қолданбалы бағдарламалау интерфейсі, ол бағдарлама жасаушыларға Facebook қолданушысының есептік жазбасынан деректерге қол жеткізуге мүмкіндік береді. Деректерге қол жеткізуден басқа, API-ді пайдаланушының есептік жазбасына Деректерді жіберуді автоматтандыру үшін де пайдалануға болады. API өкілдік күйді берудің веб-қызметін (REST) жобалауға негізделген. REST дизайнының архитектурасы кез-келген нақты желілік технологияға тәуелді емес, rest веб-қызметі әзірлеушіге хабарлама алу, жаңарту сияқты HTTP стандартты әдістерін қолдана отырып, қызмет провайдерінен ақпаратты оқуға және жазуға мүмкіндік береді.


URLs to visit
Choose a URL
Web
URL visited
Download page
Local
repository
Extract URLs

Үлкен деректерді талдау

Үлкен деректерді талдау-бұл құрылымданбаған пайдаланушы деректері туралы бай, терең және нақты ақпарат алудың технологиялық стратегиясы. Үлкен деректердің үш негізгі сипаттамасы, олар шешілуі керек негізгі мәселелерді анықтайды - көлем, әртүрлілік және жылдамдық.

- Көлемі: соңғы бірнеше жылда әлемдік деректердің 90% - ы құрылды, құрылымданбаған деректердің жаппай ауқымы мен өсуі дәстүрлі сақтау және талдау шешімдерінен асып түседі.
- Әртүрлілік: үлкен деректер бұрын талдау үшін пайдаланылмаған жаңа көздерден жиналады. Деректерді басқарудың дәстүрлі процестері электрондық пошта, әлеуметтік медиа хабарламалары, бейнелер, суреттер, блогтар, кіру журналдары және веб-іздеу тарихы сияқты әртүрлі форматта келетін үлкен деректердің өзгермелі сипатын жеңе алмайды.
- Жылдамдық: деректер нақты уақыт режимінде, қажет болған жағдайда берілетін ақпарат қажеттіліктерін ескере отырып жасалады.

Нақты уақыт режимінде тұрақты деректер ағынын өңдеу арқылы Үкімет пен құқық қорғау органдары шұғыл шешімдерді бұрынғыдан тезірек қабылдай алады, дағдарыс пен дамып келе жатқан тенденцияларды қадағалап, бағытты тез түзетіп, мүмкіндіктерді қолдана алады.

Үлкен деректерді талдау деректерді талдаушы мен сандық сот тергеушісіне жасырын заңдылықтарды, корреляцияларды, мәтінді талдауды, көңіл-күй мен көңіл-күйді және басқа ақпаратты анықтау үшін әлеуметтік желілерде пайда болған деректердің үлкен көлемін зерттеуге мүмкіндік береді. Заманауи технологиялардың көмегімен кез-келген адам өзінің Facebook-тегі қоғамдық постының мазмұнын талдау негізінде зорлық-зомбылық жасауды жоспарлап отырғанын анықтауға болады.

Hadoop/HBase

Hadoop-Apache Software Foundation компаниясының ашық көзі. Ол бірнеше кластерлік түйіндерде параллель есептеулер үшін тиімді негіз береді. Apache Hadoop-тің бүкіл "платформасы" қазіргі уақытта Hadoop ядросынан, таратылған MapReduce және Hadoop файлдық жүйесінен (HDFS), сондай-ақ Apache Hive, Apache HBase және басқаларын қоса бірқатар жобалардан тұрады деп саналады. Hadoop "жетекші-құл" режимінде жұмыс істейді. Негізгі түйін бар және бағынышты түйіндердің n саны бар. Баяндамашы бағынышты құрылғыларды басқарады, қызмет етеді және бақылайды, ал бағыныштылар нақты жұмыс түйіндері болып табылады. Шебер жай метадеректерді (деректер деректерін) сақтайды, ал бағыныштылар деректерді сақтайтын түйіндер болып табылады. Деректер кластерде таратылады. Клиент кез-келген тапсырманы орындау үшін негізгі түйінге қосылады.

HBase-ашық бастапқы код; Java-да модельделген реляциялық емес таратылған мәліметтер базасы. Ол Apache Hadoop Apache Software Foundation жобасының бөлігі ретінде жасалған және HDFS (Hadoop таратылған файлдық жүйесі) үстінде жұмыс істейді, Bigtable сияқты мүмкіндіктерді және сирек кездесетін деректердің үлкен көлемін сақтаудың қателікке төзімді әдісін ұсынады.

Әлеуметтік желілерге майнинг жасау

Әлеуметтік медианы талдау және бақылау кезінде талдаушыларға көмектесетін жүйелер құру үшін мазмұнды талдау үшін табиғи тілді өңдеудің әртүрлі әдістері (NLP) бар. Мұнда біз мәтінді талдаудың бірнеше әдістеріне шолу жасаймыз:

Жазбаша аударма қызметтері

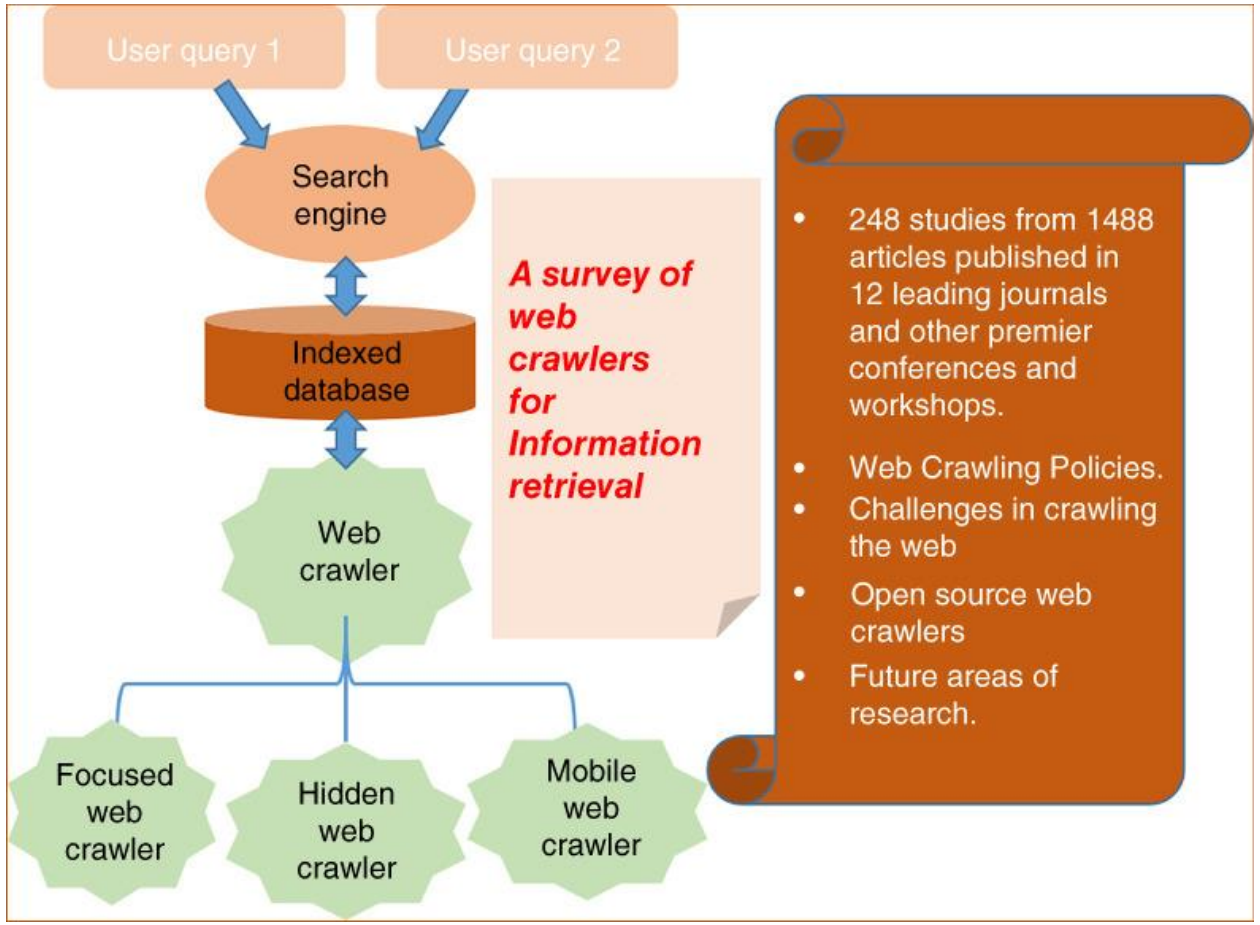
Көптеген адамдар Google сияқты ақысыз қызметтерді қолдана отырып, машиналық аударма саласындағы прогресті байқады. Бұл дамуға айтарлықтай ықпал ететін фактор дәстүрлі лингвистикалық әдістерді статистикаға негізделген әдістермен үйлестіру болып табылады. Машиналық аударманың бұл түрі экстремистік веб-сайттардан мәтіндерді аудару үшін өте пайдалы болуы мүмкін және аналитикке кез-келген тілде жазылған мәтінді өңдеуге мүмкіндік береді.

Автоматты аударма қызметтерінің осы түрімен алынған нәтижелер веб-сайттың мазмұнын маман-адам аударған сияқты сирек кездеседі, бірақ автоматты түрде аударудың үлкен артықшылығы-бұл процесс толығымен қолмен жасалғаннан гөрі көп веб-сайттарды талдауға мүмкіндік беретін жылдамдық.

Кибер барлау сарапшылары бұл қызметтерді күнделікті әлеуметтік желілердегі экстремистік мазмұнды талдау үшін немесе, мысалы, жиһадтық сайттарда пайдаланады.

Веб-сайттарды көрсету/карталау

Мәтінді іздеу ықтимал күдікті веб-сайттарды (мысалы, зорлық-зомбылық, экстремистік) автоматты түрде анықтау және осындай веб-сайттар желісін құру үшін де қолданыла алады. Компьютерлік алгоритмдерді радикалды мазмұнды анықтау үшін, аналитиктер-адамдар егжей-тегжейлі зерттеулер жүргізгісі келетін етіп білдіретін пайдаланушыларды (веб-бүркеншік аттар) анықтау үшін пайдалануға болады.


User query 1
User query 2
Search engine
Indexed database
Web crawler
Focused web crawler
Hidden web crawler
Mobile web crawler
A survey of web crawlers for Information retrieval
248 studies from 1488 articles published in 12 leading journals and other premier conferences and workshops.
Web Crawling Policies.
Challenges in crawling the web
Open source web crawlers
Future areas of research.

Авторлықты көрсету

Зерттеудің осы саласында қолданылатын алгоритмдер лексикалық ерекшеліктерді, сөз кластары мен синтаксистік ерекшеліктерді шығарып, оларды мәтін авторын анықтау үшін қолданады. Идея - әр адамның жазу стилі ерекше. Бүгінгі таңда бұл әдістер жеткілікті емес, оларды ықтимал авторлардың қайсысы мәтіннің бір немесе басқа бөлігін жазғанын анықтау үшін қолдануға болады. Агрессивті радикалды мінез-құлықты көрсететін адамды табу міндетті түрде полиция кімді әрі қарай тергеу керектігін анықтай алады дегенді білдірмейді. Пайдаланушылар веб-бүркеншік аттарды өздері бақылайтын байланыссыз құруға жақсы мүмкіндіктерге ие. Пайдаланушының жеке басы жарамды электрондық пошта мекенжайымен тексерілсе де, пайдаланушыларға тек осы мақсат үшін электрондық пошта мекенжайын жасауға ештеңе кедергі болмайды. Электрондық пошта тіркелгісін жасау үшін пайдаланылатын нақты IP мекенжайын сақтауға болатынына қарамастан, пайдаланушы анонимді серфинг қызметтерінің әртүрлі түрлерін қолдана алады.

Авторды тану немесе авторды сәйкестендіру болашақта бүркеншік атыңызды жеке тұлғаға қосу үшін маңызды болуы мүмкін. Алайда, осы саладағы зерттеулер тез дамып келеді және біз бұл әдістер алдағы бірнеше жылда әлдеқайда пайдалы болады деп ойлаймыз. Нараянанның соңғы мақаласы және т.б. интернеттегі масштабта авторларды анықтаудың перспективалы нәтижелерін көрсетеді.

Көңіл-күйді талдау

Көңіл-күйді талдау немесе пікірлерді талдау ұйымдар үшін әлеуметтік желілерде қандай тақырыптар бойынша пікірлер айтылғанын анықтайтын танымал әдіс болды. Көңіл-күйді талдаудың маңызды бөлігі - тиісті хабарламаларды анықтау және олардың қызығушылық тақырыбы бойынша оң, теріс немесе бейтарап пікірлері бар-жоғын жіктеу. Табиғи тілдерді өңдеудің бұл саласы келесі бөлімде талқыланады.

Лингвистикалық идентификаторлар

Белгілі бір экстремистік көзқарастар немесе көзқарастар субъектінің әлеуметтік желіде өзін қалай білдіретіндігінде анықталуы мүмкін. Бұл қарым-қатынас немесе ойлау өрнектері лингвистикалық идентификаторлар деп аталады. Мұндай идентификаторларды радикалды зорлық-зомбылық белгілерін тану үшін компьютерлік алгоритмдер үшін кіріс ретінде пайдалануға болады. Facebook-тегі хабарламалар алынып, сөздердің жалпы негізгі формасы қайтарылғаннан кейін зорлық-зомбылық туралы сөздер тізімімен салыстырылатын негізгі тәсіл. Ол үшін біз синонимдер жиынтығы арасындағы семантикалық қатынасты нақты көрсететін белгілі лексикалық дерекқорды қолдана аламыз.

Үлкен деректерді шығаруға арналған әлеуметтік платформа/кейс

Біздің қарастыратын мысалымыз келесі төрт негізгі кезеңнен тұрады:

А. Интернеттегі әлеуметтік желілерден мәліметтер алу

Facebook-тің әлеуметтік графигі бойынша қозғалатын Facebook API қосымшасын құру арқылы Facebook-тің көпшілік парақтарының өкілдік үлгісінен әлеуметтік медиа ағындарын жинау мақсатты түйіннен мәтіндік жариялау нысанын сұрайды және өңдеу үшін деректерді талдайды.

Facebook парақтарының үш санаты радикалды және зорлық-зомбылық мазмұнын анықтауға ең қолайлы: себеп пен қауымдастық категориясы, суретші немесе қоғам қайраткері категориясы және ойын-сауық.

Ағымдағы процесс келесі төрт кезеңнен тұрады:

- Аутентификация үшін кіру белгілері мен басқа тіркелгі деректерін алу үшін Facebook қосымшасын Facebook-те тіркеу.
- Қосымшаның аутентификациясы: OAuth (OAuth - ашық авторизация стандарты) арқылы жасалатын Facebook әлеуметтік графигіне кіру үшін қосымшаны аутентификациялау керек.
- Әлеуметтік графикалық түйіндермен өзара әрекеттесу және JSON форматында жалпыға қол жетімді бет деректерін алу үшін Facebook API графигіне ағынды сұрауларды жіберу.
- Нәтижені талдау: JSON форматында алынған нәтиже айналма деректерден шикі мәтіндік деректерді сүзу және алу үшін талдануы керек.

В. көңіл-күйді талдау және экстремистік мазмұнды анықтау

1) көңіл-күйді талдау

Көңіл-күйді талдау немесе пікірді талдау келесі алты негізгі міндеттен тұрады:

Нысандарды шығару және жіктеу: мәтіндік деректердегі барлық өрнектерді нысан кластары бойынша шығару және жіктеу.

Аспектілерді шығару және санаттау: кластерлерге субъект аспектілерінің барлық өрнектерін шығару және жіктеу.

Пікір тасымалдаушыларын шығару және санаттау: мәтіннен пікір тасымалдаушыларын шығарып, оларды жіктеу.

Уақытты шығару және стандарттау: стандартты уақыт форматында пікір алу.

Көңіл-күйді аспектілер бойынша жіктеу: пікірдің белгілі бір мақсатты аспект бойынша оң, теріс немесе бейтарап екенін анықтаңыз немесе көңіл-күйді сандық бағалауды қызығушылық аспектісіне жатқызыңыз.

Бес пікірді құру: жоғарыда аталған тапсырмалардың нәтижелері негізінде мәтіндік мәліметтерде көрсетілген барлық пікір атрибуттарын жасаңыз.

Бұл тапсырмалар өте қарапайым болып көрінеді, бірақ іс жүзінде көптеген жағдайларда олар өте күрделі және маңызды нәтижелерге қол жеткізу үшін табиғи тілді статистикалық өңдеу алгоритмдерін және жақсы ойластырылған репрезентативті оқыту жиынтығын бөлісуді қамтиды. Көңіл-күйді талдау табиғи тілді өңдеудің көптеген басқа мәселелерін қамтиды, соның ішінде сөздердің мағынасының түсініксіздігін жою, сарказмды анықтау, метафораларды түсіндіру және аспектілерді шығару.

2) Экстремистік Мазмұнды Анықтау

Экстремистік мазмұнды анықтау мақсатында біз үш мінез-құлық лингвистикалық идентификаторларын іске асыруды қамтитын өңдеудің екінші деңгейін ұсынамыз:

Ағып кету: ағып кету - бұл мақсатты тұлғаға немесе ұйымға зиян келтіру ниеті туралы хабарлама, мақсатқа деген қызығушылықты көрсетуі мүмкін және зорлық-зомбылық әрекетін зерттеу, жоспарлау және орындау туралы ескертуі мүмкін.

Бекіту: бекітілген адам белгілі бір жауды зерттеуге көп уақыт жұмсап, субъектінің наразылығына жауап беретін топқа немесе адамға терең қызығушылық танытады.

Сәйкестендіру: сәйкестендіру-бұл қару-жарақпен байланысты мінез-құлықтың алаңдатарлық сипаты. Тақырып өзін бұрынғы белгілі шабуылдаушылармен немесе өлтірушілермен немесе кез-келген істің фанаты және қорғаушысы ретінде анықтайды. Нарциссизм мен қиялдар да қауіпті мінез-құлықтың осы тобына тән белгілер болып табылады.

C. Hbase базасын құру

Facebook ағынды мәтіндік деректерінде экстремистік лингвистикалық маркерлердің көңіл-күйін талдау және анықтау алгоритмдерін өңдегеннен кейін аналитикалық деректерді HBase-де сақтау. Жолдардың деректер базасымен барлық байланыс өкілдік күйді (REST) беру қоңыраулары арқылы жүзеге асырылады.

D. құқық қорғау органдарының талдаушысы үшін бақылау интерфейсін құру

Күдікті мазмұнды анықтау үшін Java платформасы арқылы HBase-ті сұрайтын және қажет болған жағдайда одан әрі сандық сот-тергеу жүргізетін құқық қорғау органдарының бақылау интерфейсін құру.

Java платформасына қосылған Java сценарийінде интерфейс құру, ол өз кезегінде аналитикалық деректерді алу және талдау үшін HBase-ке қосылған.

Құқық қорғау органдарының криминалистикалық деректерін талдаушыға арналған графикалық ұсыну: аналитикалық деректер ақпараттың қауіп-қатер деңгейін бағалау үшін одан әрі киберин-тергеу жүргізе алатын барлау талдаушысына графикалық түрде ұсынылады. Егер кибер-тергеу тергеуді ашу үшін құқық қорғау органдарының сандық сот-медициналық тобына жіберілген сандық мәліметтерге қарағанда алдын-ала анықталған қауіптің жоғары деңгейін анықтаса, әйтпесе деректер контекст туралы хабардарлық базасында сақталады, оны мінез-құлықты талдау және анықтау және бақылау үшін пайдалануға болады .әсер ету дәрежесіне (ұнатулар, Пікірлер) байланысты зерттелуі және жіктелуі мүмкін...).

Экстремистер туралы мақсатты деректерді қолмен жинау мәселесін шешу үшін бұл зерттеу веб-тексергішті, сөйлеу бөліктерінің тегтерін (POS) біріктіру әдісін және Интернеттен жүктелген веб-беттерді жеке-жеке сыныптарға тиімді ұйымдастыра алатын шешім ағашын жасау үшін көңіл-күйді талдау әдісін көрсетуге тырысады.олардың нақты көңіл-күйлері. Релеванттықтың төрт бөлек кластары анықталды: экстремистік көңіл-күйді білдіретін мазмұн (теріс класс деп аталады), жаңалықтар көздері (бейтарап класс деп аталады), үкіметтік немесе экстремизмге қарсы ұйымдар (оң класс ретінде белгіленеді) және экстремизмге қатысы жоқ басқа мазмұн (басқа класс ретінде белгіленеді).

Осы мақсатқа жету үшін ережелер негізінде басқарылатын деректерді жинау процесін жүзеге асыруға мүмкіндік беретін құрылымды жасау қажет болды. Бұл процесс белгілі бір веб-домендерден веб-сайттарды іріктеуді анықтауды қамтыды, соның негізінде жүйенің қалған бөлігі шешім қабылдау ережелерін жасайды. Содан кейін осы веб-домендерден веб-беттерді алу үшін қолданыстағы веб-іздеуші қолданылды. Содан кейін бұл веб-беттер алдымен сөйлеу бөліктерінің тегімен (POS) талданды тоналдылық есептелетін нысандарды анықтау үшін. Барлық осы компоненттер тональды қабық деп аталатын компонентте өзара әрекеттеседі, әрбір базалық бағдарламалық жасақтамамен өзара іс-қимыл жасайтын және соңында жіктеуге арналған бағдарламалық қамтамасыз етуді негізге алатын бастапқы деректерді жасайтын кез-келген веб-бетте қандай мазмұн класы бар екенін анықтауға көмектесетін жіктеу ережелерін құру үшін пайдаланылуы мүмкін.

Бұл процестің түпкі нәтижесі-төрт сыныпты бір-бірінен ажырату үшін қолдануға болатын ережелер жиынтығы. Осы ережелерді веб-парақты айналып өту процесіне біріктіру арқылы веб-іздеуші веб-бетті жүктеу және Ережелерді қолдану арқылы бұл веб-парақтың тақырыпқа сәйкес келетіндігін немесе сәйкес келмейтінін, деректерді жинау процесін тиімді басқара алатындығын анықтай алады. Осы процестің нәтижесінде жасалған ережелер көңіл-күйді ескере отырып талданған айналып өту деректері сайттың белгілі бір сыныптың идеяларына сәйкес келетіндігін тиімді көрсете алатындығын көрсетеді.

А. зерттелетін веб-сайттар

Бұл зерттеу үшін экстремистік мазмұнға бағытталған немесе кездейсоқ экстремистік мазмұнды қамтитын веб-сайттарды пайдалану қажет болды, тіпті егер веб-сайттың негізгі бағыты табиғатта экстремистік болмаса да. Теңшелген көңіл-күйді талдау бағдарламаларымен бірге жұмыс істейтін іздеуші ақпарат категориялары арасындағы айырмашылықты тиімді тани алатындығын анықтау үшін белгілі бір кластардың белгілі бір жиынтығынан деректерді шығару қажет деп шешілді. Бұл төрт сынып оң, теріс, бейтарап және басқа сыныптар ретінде анықталды және әр сынып үшін жіктеу ережелеріне үйретілген веб-парақтардың үлгілері жиналды.

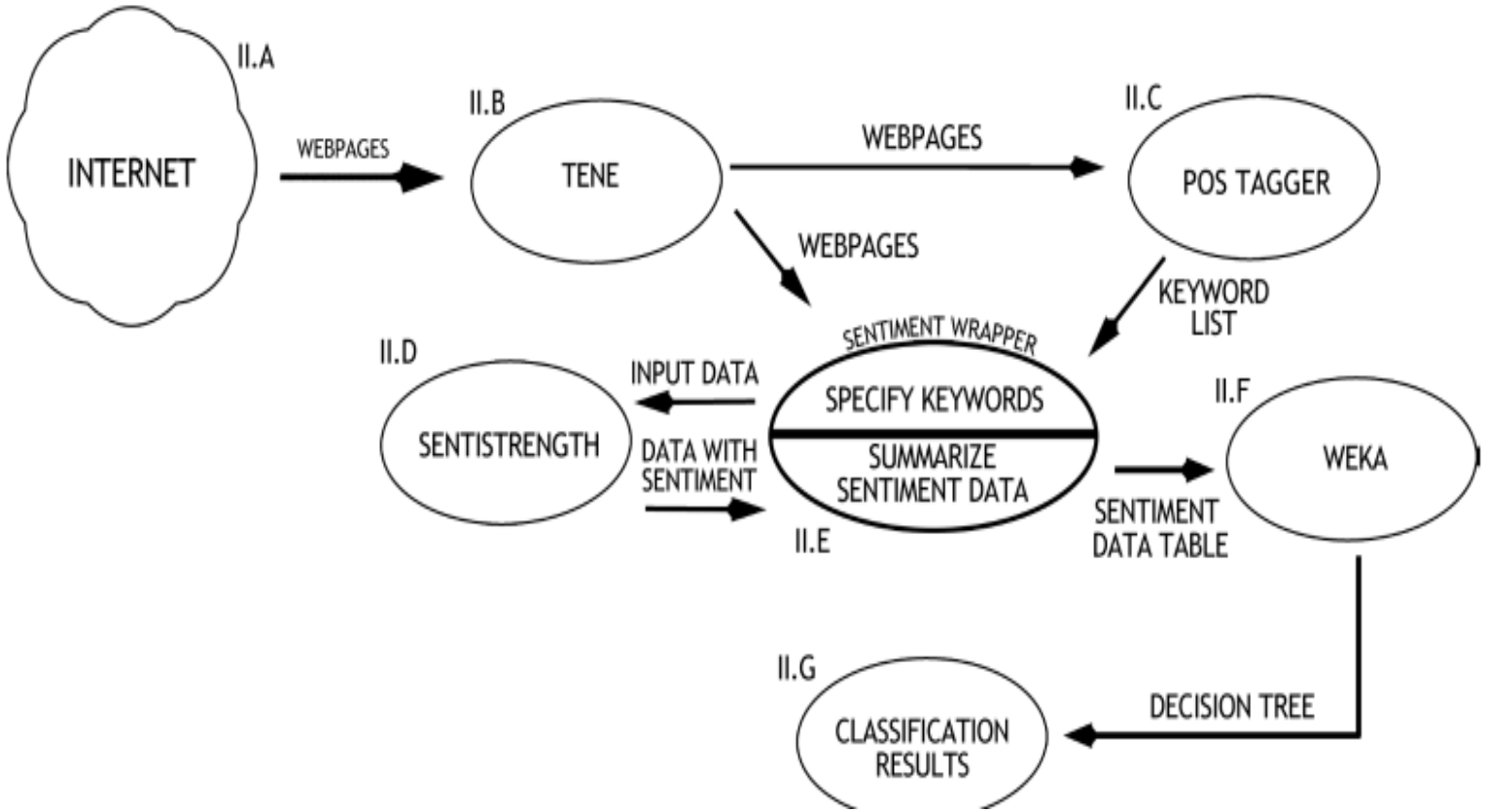
Қолмен іздеуден бастап, әр сынып үшін веб-домендердің тізімі анықталды. Оң сынып үшін деректер зорлық-зомбылық экстремистерінің сайттарына ұқсас тіркестерді (мысалы, қару-жарақ, кісі өлтіру, терроризм) қамтуы мүмкін, бірақ зорлық-зомбылық экстремизміне қатысты теріс көзқарастары бар веб-сайттардан жиналды. Оларға қоғамдық қауіпсіздік бюролары және терроризмге қарсы топтар сияқты сайттар кіреді, олар өздерінің экстремизмге қарсы саясаты мен іс-қимыл жоспарлары туралы мақалалар жариялады. Теріс үлгілер үшін веб-сайттар бұрын академиялық басылымдарда экстремистік мазмұнды немесе экстремистік пікірлерді орналастыру ретінде анықталған веб-сайттар жиынтығынан алынды. Үшінші класс, бейтарап, бұқаралық ақпарат құралдарынан алынды, олар экстремистік оқиғалар туралы есептерді қамтыды, бірақ оқиғаларды бейтарап жариялау тұрғысынан талқыланды. Сайттардың негізгі біріктіруші ерекшеліктері беттер арасындағы терминологияның ұқсастығы болды. Әр түрлі беттердің лексикасы мен тақырыбы арасындағы күрт айырмашылыққа байланысты айырмашылық болмауы үшін веб-сайттарда ұқсас мазмұн бар екеніне көз жеткізу керек болды. Ол сонымен қатар соңғы басқа сыныпты, мүлде басқа кілт сөздер жиынтығы мен фокустары бар топты қажет етті. Содан кейін осы төрт сынып веб-тексерушіге интернеттегі экстремистік мазмұнды анықтауға мүмкіндік береді, ал анықталған мазмұн үшін оның мотивациясын одан әрі анықтауға мүмкіндік береді, бұл оған тек экстремистік сипаттағы веб-сайттарға назар аударуға мүмкіндік береді.

B. Веб-іздеуші

Терроризм мен экстремизмнің желілік экстракторы (TENE)-бұл Интернетте үлкен көлемде деректерді жинау үшін халықаралық киберқылмыскерлерді зерттеу орталығында (ICCRC) құрылған веб-сканер. Бастапқыда балалардың операциялық желілерін зерттеу және картаға түсіру үшін жасалған хаттамаларды қолдана отырып, TENE веб-беттерді қарап, кілт сөздер немесе көрсетілген домендер негізінде веб-сайттар туралы ақпаратты жинай алады. Ол басталатын бастапқы домендерді және іздеудің қажеті жоқ домендерді енгізгеннен кейін, мысалы, нәтижелердің шатасуы немесе байланысты емес мәліметтер болуы мүмкін, бағдарлама әр бетте көрсетілген кілт сөздер бойынша іздейді, сонымен қатар ағымдағы парақтан басқа беттерге сілтемелер арқылы өтеді.

Бұл зерттеу үшін TENE II тарауда көрсетілген сайттардан барлық веб-беттерді жинау үшін пайдаланылды.А. белгілі бір класқа жататын домендерде сол доменде сол класқа жататын веб-беттер болады деп болжалды. Содан кейін іздеуші барлық веб-беттерді домендерден жинады, нәтижесінде әр сынып үшін мыңдаған беттер жасалды. Автоматтандырылған сканерлеу процестерінің көмегімен беттер HTML түзетулерінен тазартылды, осылайша талдау веб-беттің мәтіндік мазмұнына назар аудара алады және негізгі кодқа тәуелді болмайды. Содан кейін бұл беттер алдымен кілт сөздерді шығару үшін пайдаланылды.

Әр сынып үшін бастапқы сайттар ретінде домендердің тең саны таңдалды. Алайда, барлық домендерде қол жетімді беттер болған жоқ, сондықтан әр домендегі беттер саны бірдей емес. Зерттеу мақсаттары үшін бұл проблема болған жоқ, өйткені веб-іздеу жүйесі әр домен үшін бірдей парақтардың санын жинау мүмкін емес және мүмкін емес. Себебі, әр доменде айналып өту кезінде іздеу шарттарына байланысты мазмұн көп немесе аз болуы мүмкін. Бұл проблема болуы мүмкін жалғыз уақыт - егер талдауды жеңілдету үшін жеткіліксіз қаралған беттер болса. Осы себепті, жиналған беттері жоқ домендер алынып тасталды.


II.A
INTERNET
WEBPAGES
II.B
TENE
WEBPAGES
II.C
POS TAGGER
WEBPAGES
KEYWORD
LIST
SENTIMENT WRAPPER
II.D
SENTISTRENGTH
INPUT DATA
DATA WITH
SENTIMENT
SPECIFY KEYWORDS
SUMMARIZE
SENTIMENT DATA
II.E
II.F
WEKA
SENTIMENT
DATA TABLE
II.G
CLASSIFICATION
RESULTS
DECISION TREE

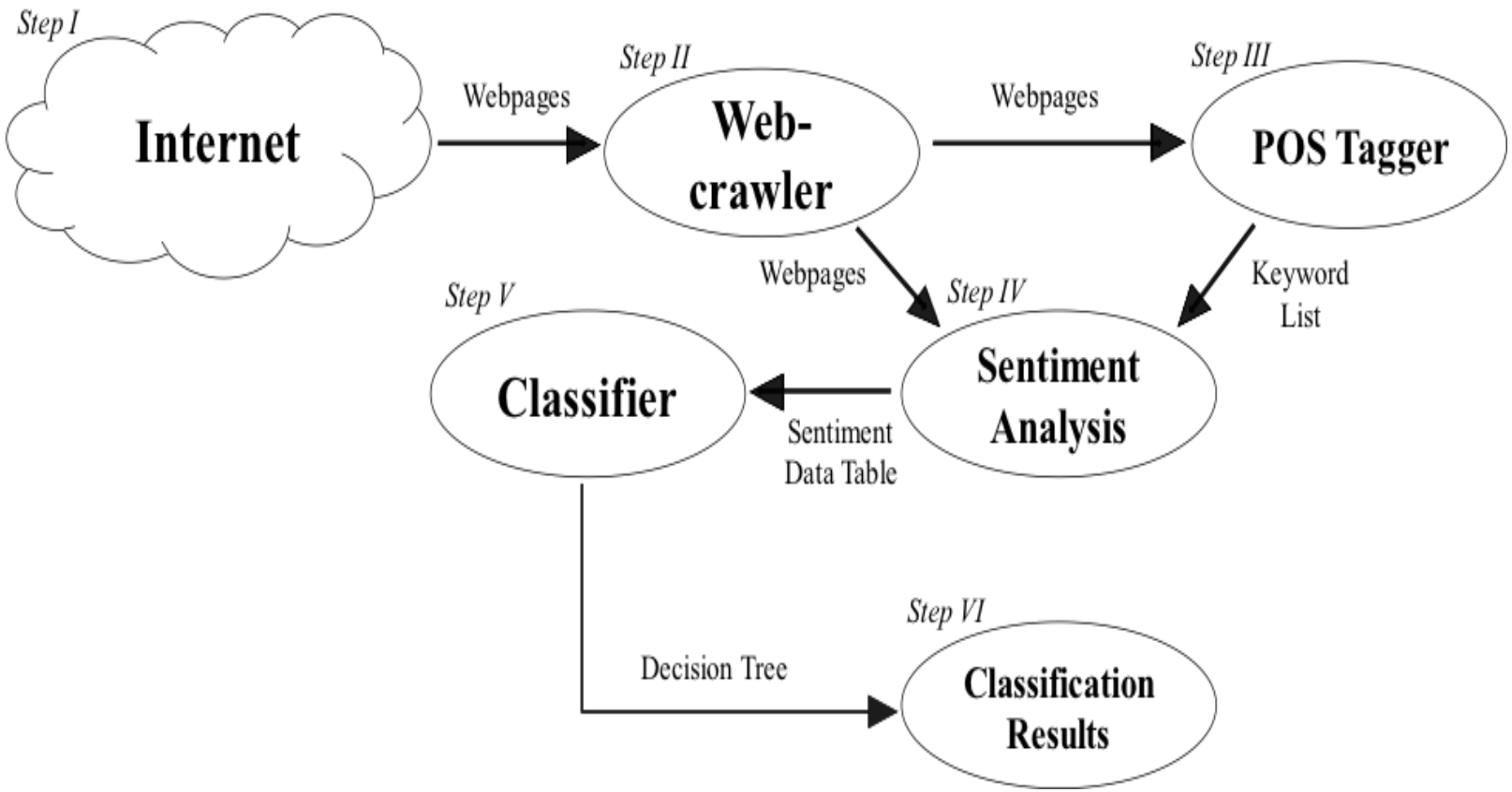
Name	Domain	Description
Stormfront	www.Stormfront.org	A white supremacist forum that involves many racially driven discussions with a distinctly right wing focus [25].
Hizb ut-Tahrir	hizbuttahrir.org/	An Islamic political party that seeks to reinstate the caliphate through the dissemination of propaganda [24].
Army of God USA	www.aogusa.org	An anti-abortion group labelled "radical" or called a "terrorist group" due to their extreme methods of protesting abortion [22, 23].
Islamic Awakening	www.islamicawakening.com	A web forum discussing Islam that has had incidental discussions involving extremism and hosted radical extremist content from other sites [31].
The American Nazi Party	www.americannaziparty.com	A neo-Nazi group identified as a "domestic extremist group". [25]
Life Site News	www.lifesitenews.com	A news site with strictly negative views towards, homosexuality, abortion and gay marriage [30].
Shahmat	shahamat-english.com	A news site that has hosted interviews from prominent Al-Qaeda leadership [28].
Kavkaz Center	www.kavkazcenter.com/eng/	Labelled an extremist site by Article 1 of Russia's federal law "On Countering Extremist Activities" [27]
The Muslim Brotherhood	www.ikhwanweb.com	A group that has been associated with radical Islamic nationalism. [26]

Table 1 – Extremist (negative) class domains sampled with the Web-crawler

Name	Domain	Description
Secret Intelligence Service	sis.gov.uk	The British Secret Service, involved in counter terrorism, cyber security and various anti extremist activities.
The Global Counterterrorism Forum	www.thegctf.org	This group's Mission statement is "to reduce the vulnerability of people everywhere to terrorism" [29]
Australian Department of Foreign Affairs and Trade	http://dfat.gov.au/	This site discusses Australia's work with other countries to counter terrorism.
Public Safety Canada	www.publicsafety.gc.ca	A Nationally organized group that has anti-terrorism within its official mandate.
U.S. Department of State	http://www.state.gov/	A National organization that includes the Bureau of Counter Terrorism.
United Nations	www.un.org	An international peacekeeping organization that includes committees specifically tailored to combatting terrorism
NATO	www.nato.int	An international organization with their own anti-terrorism programs

Table 2 – Anti-extremist (positive) class domains sampled with the Web-crawler

Бұл жоба деректерді, атап айтқанда экстремистік веб-сайттардан жинау үшін дербес веб-сканер құру тәсілін ұсынды. Веб-іздеуші нұсқаулығының құрамдас бөлігі іздеушіге жүктелген веб-парақтың мазмұны туралы шешім қабылдауға мүмкіндік беретін көңіл-күйге негізделген жіктеу ережелерін қолдану арқылы қол жеткізілді. 2500 веб-парақтан тұратын алғашқы мазмұн сезімге негізделген төрт түрлі сыныптың әрқайсысына жиналды: экстремизмге қарсы веб-сайттар, экстремизмге қарсы веб-сайттар, экстремизмді талқылайтын жаңалықтар сайттары және соңында экстремизмді талқыламайтын сайттар. Содан кейін осы беттерде жиі кездесетін кілт сөздерді іздеу үшін сөйлеу бөліктерін белгілеу (POS) қолданылды. Көңіл-күйді анықтайтын бағдарламалық жасақтаманы жіктеу бағдарламалық жасақтамасымен бірге қолдана отырып, белгілі бір беттің қай сыныпқа кіретінін тиімді анықтай алатын шешім ағашы жасалды. Алынған ағаш төрт сынып арасындағы дифференциация кезінде 80% сәтті және экстремистік беттерді нақты жіктеуде 92% сәтті көрсетті.


Step I
Internet
Webpages
Step II
Web-
crawler
Webpages
Step III
POS Tagger
Webpages
Step IV
Keyword
List
Step V
Classifier
Sentiment
Analysis
Sentiment
Data Table
Step VI
Decision Tree
Classification
Results

Назарларыңызға рақмет!