Евразийский национальный университет им. Л.Н.Гумилева

Анализ тональности комментариев в социальных сетях на основе правил

Докладчик: Ергеш Бану Жантуғанқызы

b.yergesh@gmail.com

Актуальность

- широкое распространение интернета;
- Популярность социальных сетей, блог-платформ и форумов (Facebook, Twitter, VK, Instagram, TripAdviser и др.);
- необходимость автоматической обработки и анализа мнений пользователей Интернета.

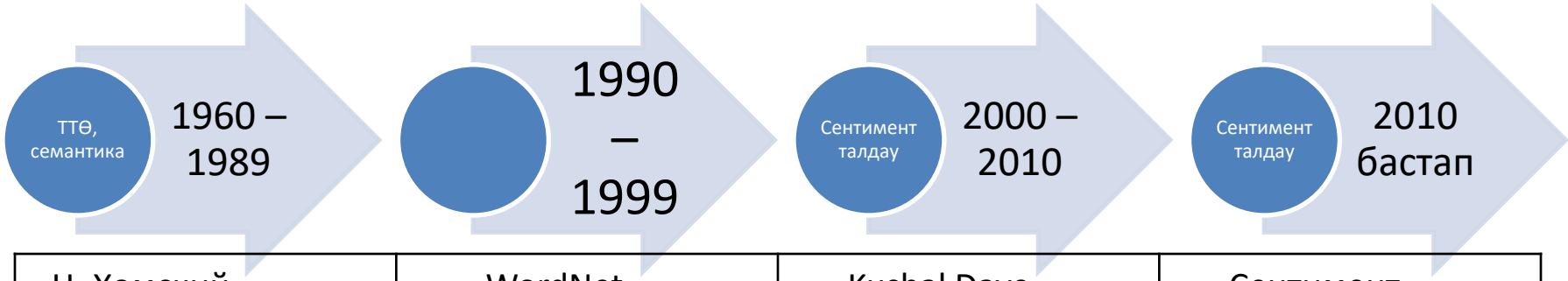
Цель исследования

Изучить тексты комментариев в социальных сетях и определить их тональности на основе правил


Обработка естественных языков(NLP)
Сентимент талдау
Искусственный интеллект(AI)
Text Data Mining

Мәтіндерді сентимент талдауға шолу

- **Қысқаша тарихы**



- Н. Хомский (фомалды грамматика) -Ахо, Ульман (тілдер, автоматтар теориясы) -Семантика (Ч.Филлмор, И.Мельчук, А.Жолковский, Р.Шенк,Д.Кнут)	- - WordNet - V. Hatzivassiloglou, J.M. Wiebe зерттеулері (сын есім) - Субъективтілікті анықтау - Dictionary based	- Kushal Dave (opinion mining) - Das and Chen (sentiment analysis) - P.Turney (Thumbs up or thumbs down?) - Bo Pang , B. Liu. - Аспектіні анықтау - Сентимент талдау қосымшалары - Corpus based, ML	- Сентимент талдауға қатысты зерттеулер тез өсті - Эмоцияны тану - Сарказмды анықтау - Видео, аудио, кескіндерден эмоцияны анықтау

Сентимент талдау мәселелері

Деңгейі:

•Құжат деңгейі;

•Сөйлем деңгейінде талдау;

•Болмыс және аспект деңгейі.

Есептері:

• Пікірлерді талдау (Тікелей ой-пікірі, Салыстырмалы пікір);

•Мәтіннің субъективтілік/объективтілігін анықтау;

•Тоналдылық (реңкі) бойынша сыныптау (жағымды, бейтарап, жағымсыз);

•Пікірде аталған кілттік объектілерді(аспекттерді) анықтау;

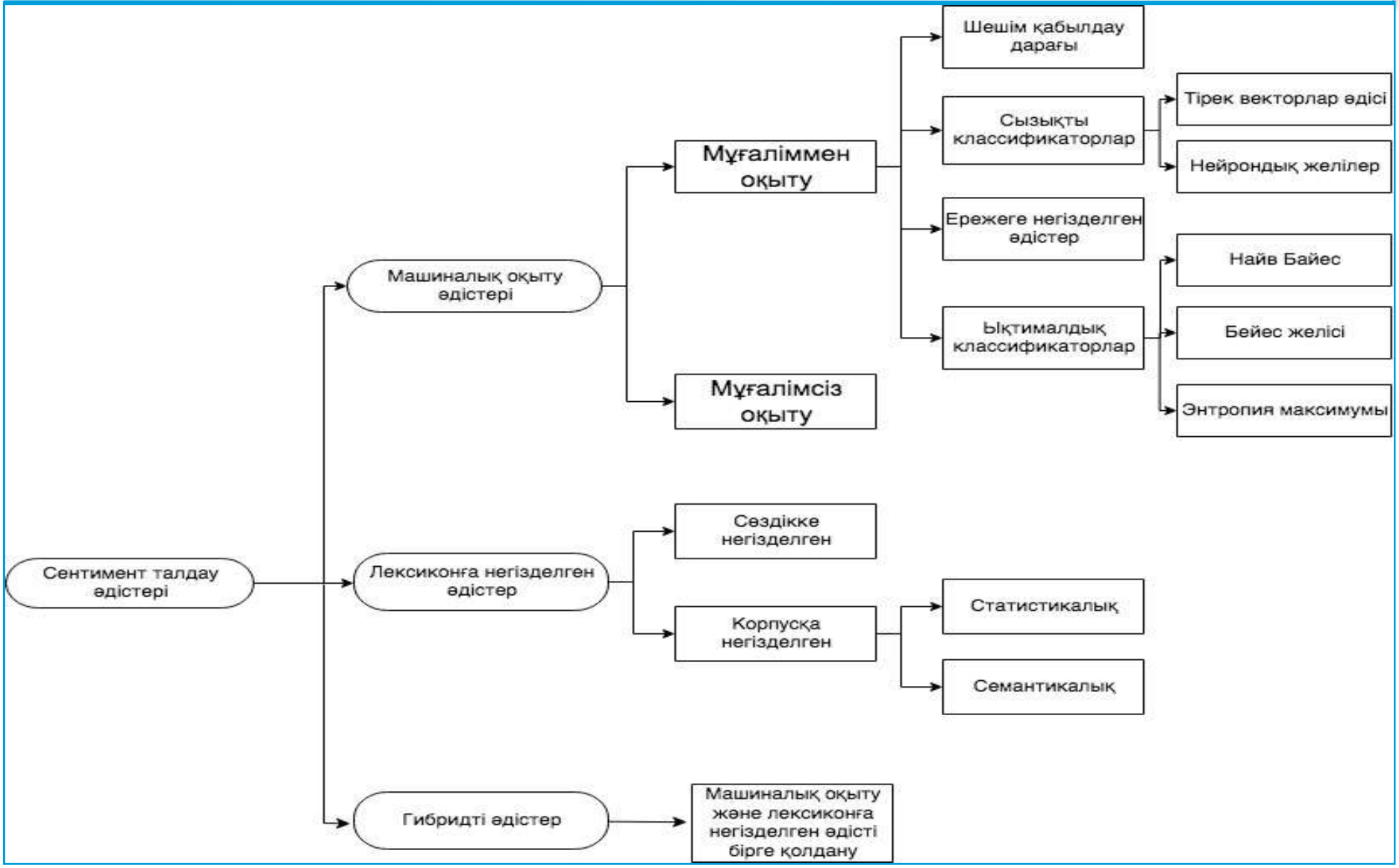
•Пікірлерді автоматты мазмұндау;

•Эмоцияны тану (қуаныш, ренжу, жындану және т.б.);

•Мысқылдауды(сарказм) тану.

Сентимент талдау әдістері

Сентимент талдау әдістері



Басқа тілдер үшін шешімдер

Қосымшалар

- ✓Google cloud;
- ✓Social Mention;
- ✓Sentiment140;
- ✓«SentiStrength» ;
- ✓Semantria;
- ✓SentiFinder;
- ✓Microsoft Azure Text Analytics API;
- ✓Microsoft Azure Emotion API.

Лексикалық ресурстар

- ✓WordNet-Affect;
- ✓ SentiWordNet;
- ✓SenticNet;
- ✓MPQA Opinion Corpus;
- ✓РуСентиЛекс.

Зерттеу есептері

- Қазақ тіліндегі мәтінді сентимент талдауға әсер ететін белгілерді (features, признаки) анықтау;
- Қазақ тіліндегі сентимент белгісі қойылған лексикалық бірліктердің семантикалық базасын құру;
- Сентимент талдау есебін шешуде қолданылатын қазақ тілінің грамматикалық ережелерінің метатілін әзірлеу;
- Қазақ тіліндегі мәтіндерді сентимент талдау үшін қазақ тілінің грамматикалық (морфологиялық және синтаксистік) ережелерін формалдау және талдау;
- Қазақ тіліндегі мәтіндерді сентимент талдау моделі мен әдісін құру және оны программалық жүзеге асыру.

Қорғауға шығарылатын негізгі нәтижелер

- Қазақ тіліндегі реңктік лексикалық бірліктердің семантикалық базасы;
- Қазақ тіліндегі мәтіндерді сентимент талдау моделдері;
- Қазақ тіліндегі мәтіндерді сентимент талдау әдісі;
- Қазақ тіліндегі мәтіндерді сентимент талдау алгоритмі және программасы.

Қазақ тіліндегі мәтіндердің сентиментін анықтау

Қазақ тіліндегі мәтіннің сентимент бағытын анықтауға әсер ететін белгілер:

•сөз таптары:

-зат есім (жауыздық, соғыс),

-етістік (тұтқындау, қуанды, ашуланды),

-сын есім (әдемі/көріксіз, жақсы/жаман),

-үстеу (нағыз, ең, өте),

-одағай (алақай!, бәрекелді!, әтеген-ай).

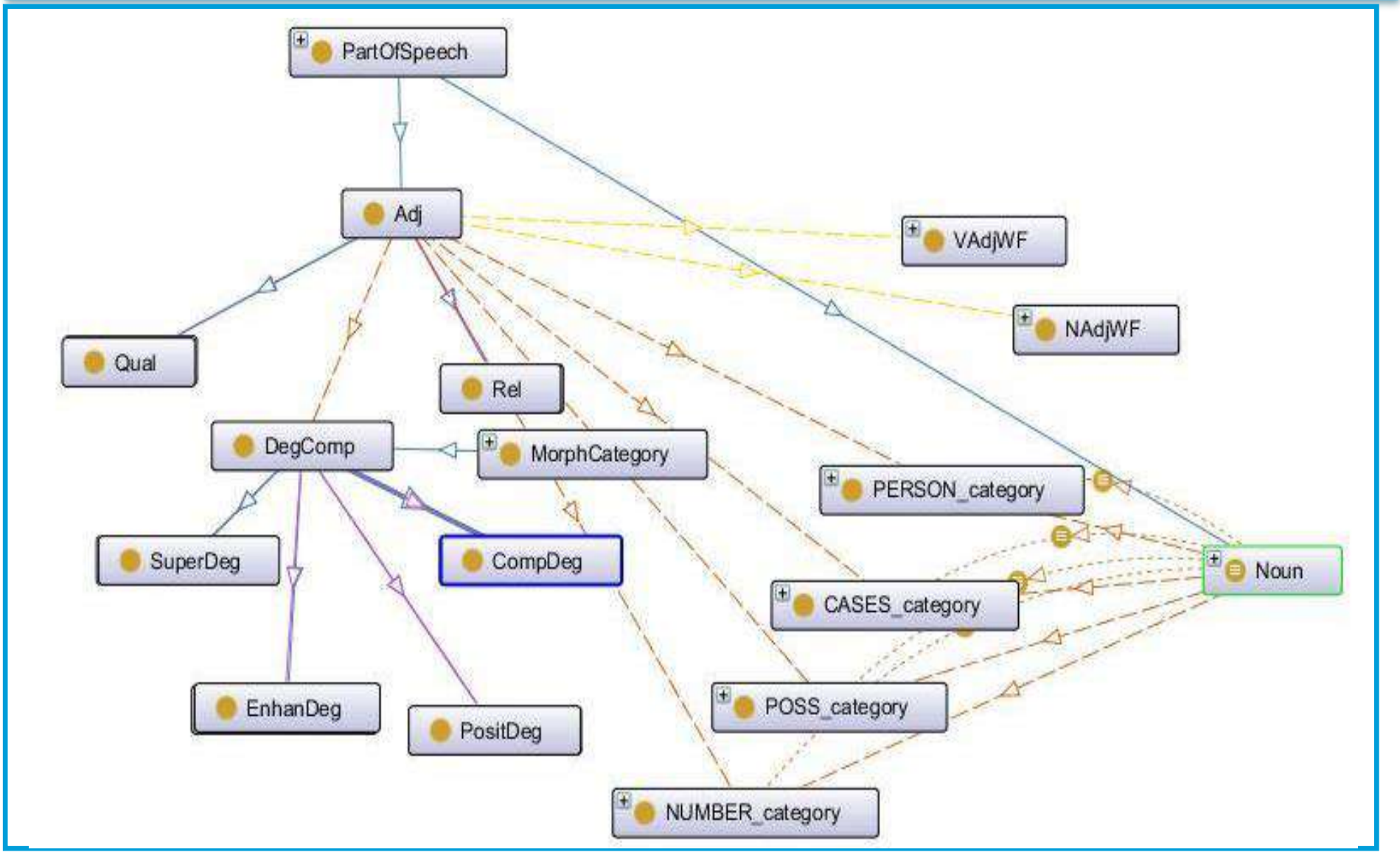
•терістеу сөздері(жоқ, емес, еш, ешқашан)

•эмотикондар

Қазақ тілінің грамматикалық ережелерінің метатілі

Tag	Name_English	Kazakh
N	Noun	Зат есім
Adj	Adjective	Сын есім
Adj_P	PositiveAdjective	Жағымды сын есім
Adj_N	Negative Adjective	Жағымсыз сын есім
V	Verb	Етістік
FW	Function words	Шылау
Intrj	Interjection	одағай
Emtn	Emotional	Көңіл-күй одағайлары
Pstv	Positive emotion	Жағымды көңіл күйді білдіретін

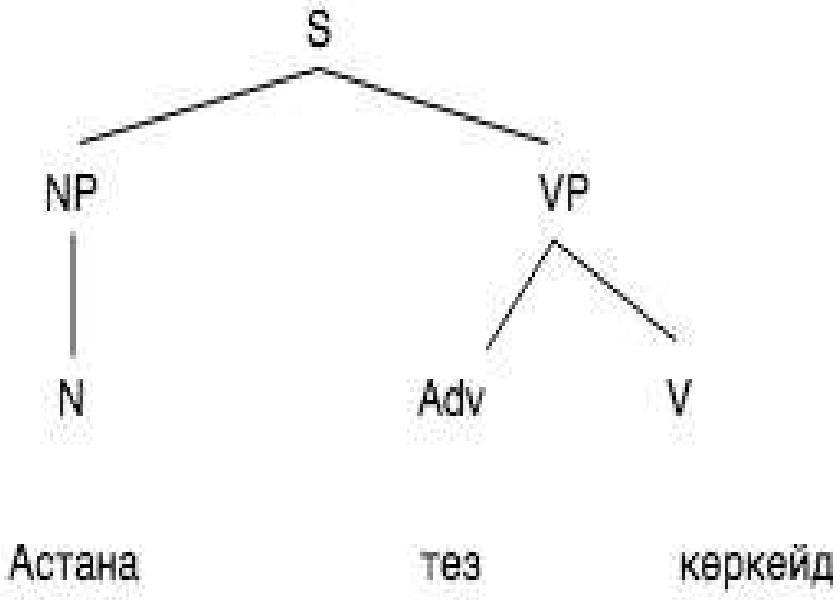
Қазақ тілінің морфологиялық ережелерін моделдеу

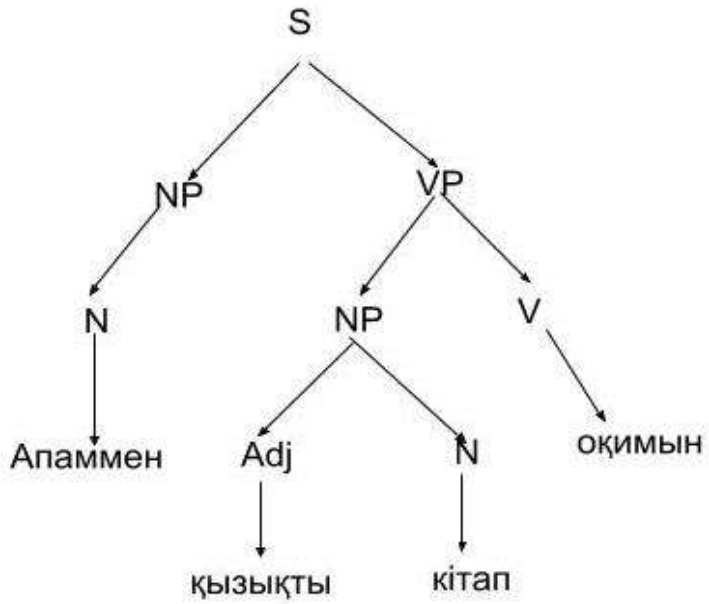


Қазақ тілінің синтаксистік ережелерін формалдау

$G = < N_s, S, T_s, R >,$

S → NP VP|Imit NP VP| Intrj VP| Intrj NP





S(NP(N), VP(Adv,V)) құраушы дарағы

S(NP(N), VP(NP(Adj,N),V)) құраушы дарағы

Қазақ тілінде реңк беретін сөздер мен сөз тіркестері

- [N] · [V]
- [N] · [V] · [Negation]
- [ADJ] · [N]
- [ADJ] · [Negation] · [N]
- [ADJ] · [V]
- [ADJ] · [V] · [Negation]
- [ADV] · [ADJ]
- [ADV] · [N];
- [Intrj];

Реңктік бірліктердің атауы, өлшемі, мәндері

№	Реңктік бірліктің атауы	Реңктік бірліктің өлшемі	Реңктік бірліктің мәндері
1	өте жағымсыз	бүтін сан	-2
2	жағымсыз	бүтін сан	-1
3	бейтарап	бүтін сан	0
4	жағымды	бүтін сан	1
5	өте жағымды	бүтін сан	2

Сентимент талдау моделінің белгілері

Белгілеуі	Түсіндірмесі
$\alpha, \beta, \gamma, \ldots, \zeta, \xi$	Тілдегі сөздер – айнымалылар
ω	$\omega = \zeta \cdot \alpha \cdot \beta \cdot \xi$ – лексикалық бірлік (бос емес сөз немесе сөз тіркесі)
L	Тілдегі сөйлемдер жиыны
N	Зат есімдер жиыны
Adj	Сын есімдер жиыны
$Pron$	Есімдік сөздер жиыны
$V_{\downarrow}Post$	Болымды етістіктер жиыны
$V_{\downarrow}Negt$	Болымсыз етістіктер жиыны
$AdvIntens$	Күшейтпелі үстеулер жиыны
$sent$	Сентимент -Предикат
@	Терістеу сөздері емес/жоқ - Тұрақтылар
¬	Болымсыз түрге айналу - Операция
.	Конкатенация - Операция

Мәтіндерді сентимент талдау моделі

Егер сентимент талдаудың лексикалық бірлігінде жағымды реңкті зат есім болса және одан кейін бейтарап реңкті болымды етістік болса, онда осы бірліктің сентименті жағымды болады.

$$\frac{\omega \in L,\ \omega = \zeta \cdot \alpha \cdot \beta \cdot \xi\ ,\ \alpha \in N,\ sent(a) = 1,\ \beta \in V\_Post, sent(\beta) = 0}{sent(\omega) = 1}$$

Мысалы, той болды, қуанышқа толды.

Мәтіндерді сентимент талдау моделі

Егер сентимент талдаудың лексикалық бірлігінде жағымды реңкті зат есім және одан кейінгі бейтарап реңкті етістікке теріcтеу сөзі жалғасса, онда осы бірліктің сентименті жағымсыз болады.

$$\frac{\omega \in L,\ \omega = \zeta \cdot \alpha \cdot \beta \cdot \gamma \cdot \xi\ ,\ \alpha \in N,\ sent(a) = 1,\ \beta \in V\_Post,\ sent(\beta) = 0,\ \gamma = @}{sent(\omega) = -1}$$

Мысалы, әділеттілік орнаған жоқ.

Мәтіндерді сентимент талдау моделі

Егер сентимент талдаудың лексикалық бірлігінде жағымды реңкті сын есім мен бейтарап реңкті зат есімнің арасында емес/жоқ сияқты терістеу сөздері кездессе, онда онда бірліктің сентименті жағымсыз болады.

$$\frac{\omega \in L,\ \omega = \zeta \cdot \alpha \cdot \beta \cdot \gamma \cdot \xi\ ,\ \alpha \in Adj,\ sent(a) = 1,\ \beta = @, \gamma \in N, sent(\gamma) = 0}{sent(\omega) = -1}$$

Мысалы, қызық емес кітап.

Мәтіндерді сентимент талдау моделі

Егер сентимент талдаудың лексикалық бірлігінде күшейтпелі үстеу сөздер жағымды сын есімдердің алдында тұрса және одан кейінгі сөз бейтарап реңкті зат есім болса, онда ол бірліктің сентименті өте жағымды болады.

$$\frac{\omega \in L,\ \omega = \zeta \cdot \alpha \cdot \beta \cdot \gamma \cdot \xi\ ,\ \alpha \in AdvIntens,\ \beta \in Adj,\ sent(\beta) = 1, \gamma = N, sent(\gamma) = 0}{sent(\omega) = 2}$$

Мысалы, өте әдемі қыз, ең әдемі қала.

Мәтіндерді сентимент талдау моделі

Мәтіннің реңктік бағасы лексикалық бірліктердің (сөз тіркестерінің) реңін өлшеудің орташа арифметикалық шамалары және олардың үйлесім ережелері ретінде анықталады:

$$sent(L) = \frac{\sum_{i=1}^{n} sent_i(\omega)}{n}$$

Лексикалық бірлік – қандай да бір затты, құбылысты, олардың қасиеттерін білдіретін сөз, тұрақты сөз тіркесі немесе тілдің басқа бірлігі.

Мәтіндерді сентимент талдау әдісі

Қазақ тіліндегі мәтіннің сентиментін анықтау үшін **сөздікке және ережеге негізделген әдіс** ұсынылады.

Сөздік ретінде осы жұмыста құрылған қазақ тіліндегі реңктік бояулары белгіленген лексикалық бірліктердің семантикалық базасы болады.

Ереже ретінде продукциондық модел қолданылып жасалған қазақ тіліндегі мәтіндердің сентиментін анықтайтын формалды ережелері алынды.

Мәтін бөлігінің сентиментін анықтау үшін қолданылатын әрбір ереже **«ЕГЕР шарт, ОНДА қорытынды»** түрінде берілген.

Қазақ тіліндегі реңктік лексикалық бірліктердің семантикалық базасы

База құрылымы:
!1. Лексикалық бірліктердің кілті,
!2. Сөз немесе сөз тіркесі,
!3. Синонимі,
!4. Сөз табы,
!5. Реңктік бағасы
!6. Комментарий.

Лексикалық бірліктердің реңкі 5 балдық шкаламен өлшенеді:
-2 – өте жағымсыз; -1 – жағымсыз; 0 – бейтарап; 1 –жағымды; 2 – өте жағымды.
Кейбір сөздер не белгілер мәтіннің мағынасына байланысты сентимент бағытын өзгертуі мүмкін («+-»).

Қазақ тіліндегі реңктік лексикалық бірліктердің семантикалық базасы

Қазақ тіліндегі реңктік лексикалық бірліктердің семантикалық базасы сын есім, зат есім, етістік, үстеу (күшейтпелі), одағай және эмотиконнан құралған.

База көлемі: 10 000-нан аса сөздер және сөз тіркестері және эмотикондар.

«Қазақ тілінің реңктік лексикалық бірліктер базасы» деп аталған деректер базасына авторлық құқықпен қорғалатын объектісін тіркеу куәлігі алынды (Куәлік № 8392, 26.02.2020).

Сентимент талдау алгоритмі

Мәтінді сентимент талдау Реңктік лексикалық бірліктердің семантикалық базасы (РЛБСБ) пайдаланылып, келесі алгоритммен жұмыс істейді:

1. *Сөздерге реңктік белгі беру* мәтінде кездескен реңктік мағынасы бар лексикалық бірліктерге РЛБСБ-ғы ақпаратқа сейкес сентимент бағытын меншіктейді.
2. *Сөз тіркестерінің реңктік бағасын беру* формалды ережелерге сәйкес лексикалық бірліктерінің реңктік бағасын береді.
3. *Мәтінге реңктік баға беру* мәтінде жалпы табылған реңктік лексикалық бірліктерінің жалпы бағасын шығарады.

Сентимент талдау алгоритмі

Мәтінге реңктік баға беру кезеңінің нақтыланған алгоритмі төмендегідей:

1. БАСЫ. Сентимент талдау кірісіне сөйлемді морфологиялық және синтаксистік талдаудан кейін алынған сөздер тізбегі анықталған сөз таптары бойынша құрылым ақпараттарымен беріледі.
2. Берілген сөздер тізбегінен цикл бойынша t-сөз бен t+1 сөздің реңктік бағасы РЛБСБ-нан ізделінеді. Егер табылса, онда 3- қадамға өтеді, әйтпесе 7-қадамға өтеді.
3. Берілген сөздер тізбегінен цикл бойынша t-сөз бен t+1 сөздің реңктік бағасы қойылады.
4. Сентимент бағытын анықтау ережелеріне тексеріледі. Егер ережеге сәйкес келсе, сөз тіркестерінің сентимент бағыты беріледі. Ережеге сәйкес тіркестер табылмаса, онда 7-қадамға өтеді.
5. Ережеге сәйкес әр сөйлемнің сентимент бағыты анықталады.
6. Жалпы мәтіннің сентимент бағыты анықталады.
7. СОҢЫ.

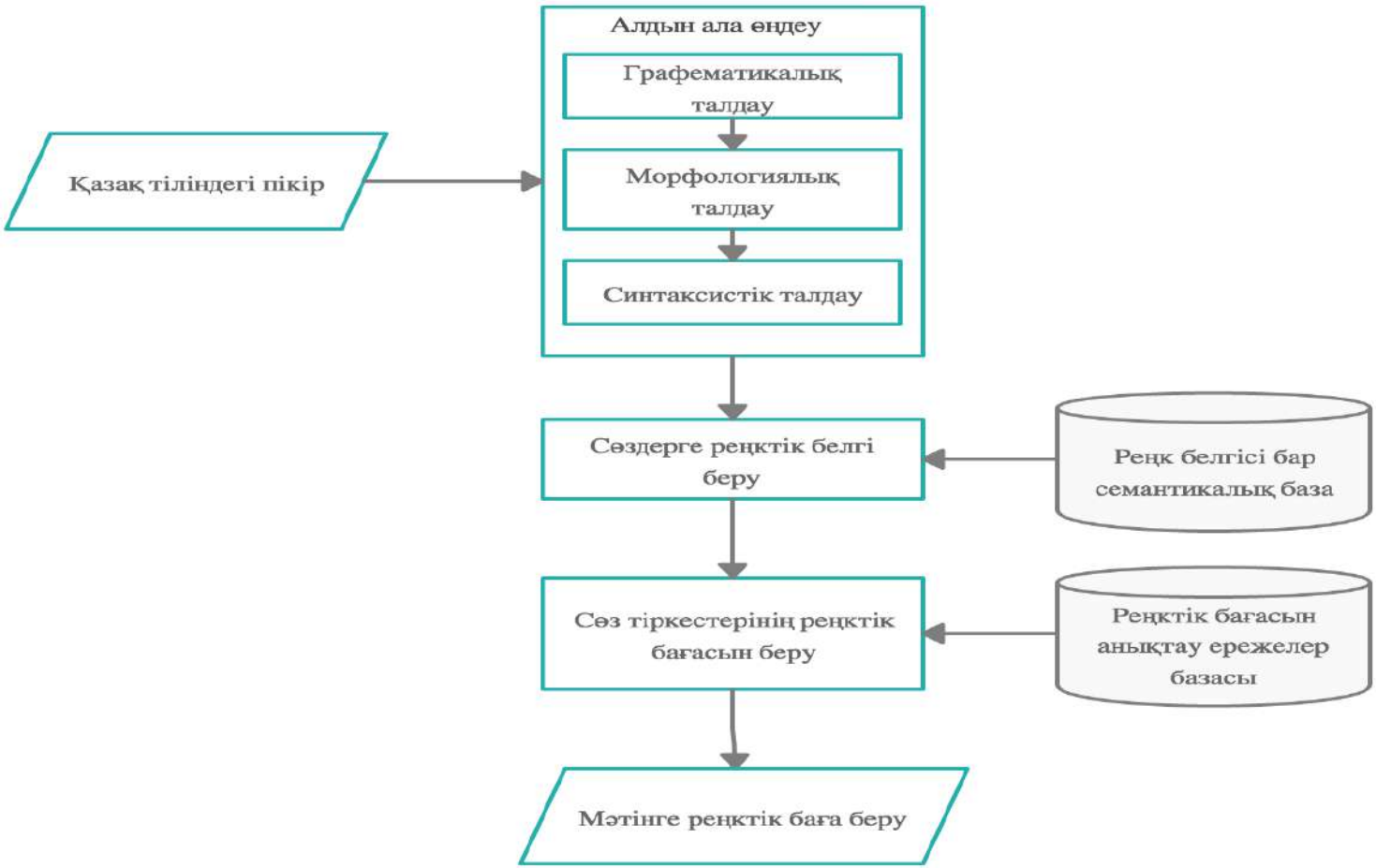
Программалық жүзеге асыру

Мақсаты:
Қазақ тіліндегі мәтіндерді сентимент талдау үшін ұсынылған модел, әдістің практикалық жүзеге асырылуын негіздеу.

Ендірілулер:
-«Қазақстан жасанды интеллект академиясы» Қоғамдық бірлестігінде
-«GeoSeT» ЖШС
.

Программалық жүзеге асыру

Программа құрылымы



Нәтижелердің қолдану аясы

Интернетте күн сайын көлемі өсіп жатқан қазақ тіліндегі мәтіндерді сентимент талдау арқылы:

- білім, денсаулық және басқа әлеуметтік салаларға байланысты қоғамдық пікірді және халықтың көңіл-күйін анықтау жүйелері;
- экономика, қаржы, өндіріс және басқа салаларда мониторинг, ұсыныс, сараптау және болжау жасау жүйелері.

Келешегі

Түрлі пәндік салаларда пікірлердің объективтілік/субъективтілігін анықтау, аспектіге бағытталған сентимент талдау, эмоцияны тану, мысқылдауды анықтау, пікірлерді мазмұндау (opinion summarization), жалған пікірлерді анықтау сияқты есептерді шешуге болады.

Қорытынды

- Қазақ тіліндегі сентимент бағыты белгіленген реңктік лексикалық бірліктердің сенмантикалық базасы құрылды;
- Қазақ тіліндегі мәтіндерді сентимент талдау моделі мен әдісі әзірленді;
- Қазақ тіліндегі мәтіндерді сентимент талдау алгоритмі әзірленіп, программалық жүзеге асырылды.
- Жасалған ғылыми-зерттеу нәтижелері бойынша авторлық құқықпен қорғалатын объектісін тіркеу куәлігі алынды, ендірілу жасалды.

Ғылыми бағдарламалар мен жобалар

- «Қазақ тілінің жазбаша және ауызша сөйлеулерін тану және тудыруды автоматтандыру» (мем. Тіркеу № 0112РК02251)
- «AP05132249 Көптілді іздеу және білімдерді шығару жүйелерін құруға арналған түркі тілдерінің электрондық тезаурустарын әзірлеу»

Жарияланымдар

Жалпы – 25, оның ішінде:

•ҚР БҒМ Білім және ғылым саласында бақылау комитеті ұсынған журналдарда – 6;
•шетелдік журналдарда - 4, оның ішінде Web of Science және Scopus ДБ кіретін журналдарда – 3;
•халықаралық және басқа конференциялар жинақтарында – 15, оның ішінде төрт жұмыс Web of Science және Scopus БД индекстелген.

Назарларыңызға рахмет!

Авторлық құқықты тіркеу куәлігі


ҚАЗАҚСТАН РЕСПУБЛИКАСЫ

РЕСПУБЛИКА КАЗАХСТАН

АВТОРЛЫҚ ҚҰҚЫҚПЕН ҚОРҒАЛАТЫН ОБЪЕКТІЛЕРГЕ ҚҰҚЫҚТАРДЫҢ МЕМЛЕКЕТТІК ТІЗІЛІМГЕ МӘЛІМЕТТЕРДІ ЕНГІЗУ ТУРАЛЫ

КУӘЛІК

03.06.2019

Подписано ЭЦП

Нәтижелерді ендіру акті

УТВЕРЖДАЮ

Директор ТОО «GeoSet»
_____ А. Б. Амренова
«__» _____ 2019 г.
М.П.

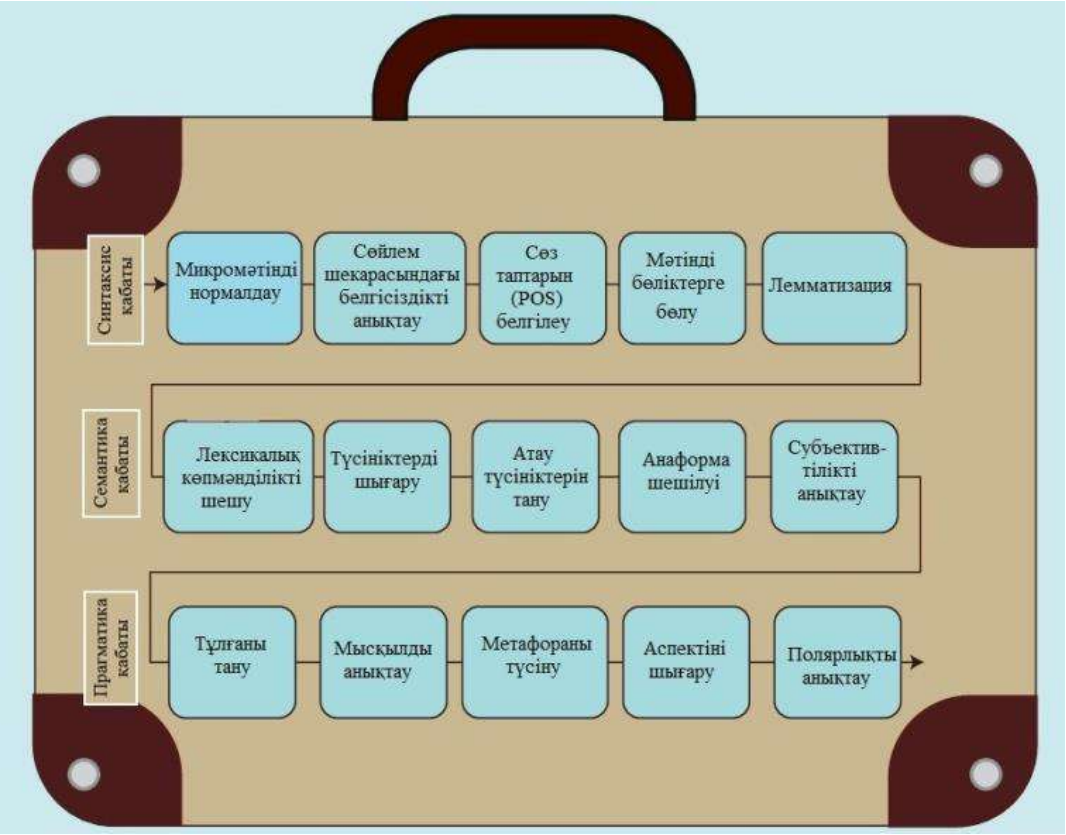
АКТ
о внедрении (использовании) результатов научно-исследовательской работы Ергеш Бану Жантуғанқызы

1. Настоящий Акт свидетельствует, что результаты диссертационной работы Ергеш Б.Ж. на тему «Қазақ тіліндегі арнайы мәтіндерді семантикалық талдау моделдері мен әдістері» представленной на соискание степени доктора PhD внедрены в деятельности ТОО"GeoSet"
2. Форма внедрения: установка и настройка прототипа программы сентимент анализа текстов на казахском языке на серверах организации.
3. Сроки внедрения: с 9 по 20 декабря 2019 года.
4. Эксплуатационные характеристики объекта внедрения:
   - метод на основе формальных правил и размеченной по тональности базы данных позволяет определять тональность отзывов/мнений на казахском языке по 5-бальной шкале.
   - возможность пополнения базы данных казахских слов и словосочетаний, характерные для предметной области.
5. Эффективность внедрения: внедерение данной программы позволяет автоматический определять тональность отзывов/мнений пользователей социальных сетей, форумов, блогов написанных на казахском языке.
6. ТОО "GeoSet" обязуется не передовать разработку для использования в другие организации.

Директор ТОО «GeoSet» **А.Б. Амренова**

Мәтіндерді сентимент талдауға шолу

сентимент талдау, пікірлерді интеллектуалды талдау, пікірлерді табу, субъективтілікті табу, көңіл күйді талдау деген сияқтылар атаулары бар



Шолу. Қолдану аясы

- ✓ коммерциялық салада(өнімдер, қызметтер, сауда белгілерін бағалауда);
- ✓ қоғамдық/әлеуметтік салада(сайлау нәтижелерін болжауда, сол уақытта талқыланып жатқан тақырыптарды бағалауда);
- ✓ ойын-сауық(кино, кітап, жұлдыздарды бағалауда);
- ✓ денсаулық сақтау саласы;
- ✓ білім (университет, колледж, мектеп, балабақшаларды анықтауда);
- ✓ туризм;
- ✓ спорт....

Түсініктер

- Сентимент – мәтінде сипатталған кейбір объектіге (өнім, ұйым, тұлға және т.б.), оқиғаға (сайлау, көтеріліс, соғыс және т.б.), құбылысқа (ай тұтылу, су тасқыны және т.б.), үрдіске (білім беру, қызмет көрсету және т.б.) немесе оның қасиеттеріне қатысты пікірді білдірген автордың эмоционалдық қатынасы.

Есептің қойылуы

Семантика – тіл бірліктерінің мағыналық мәнін зерттейтін лингвистика бөлімі.

Сентимент талдаудың лексикалық бірлігі – эмотикондар, сөз, сөз тіркесі, фраза немесе жай сөйлем.

Лингвистикада семантикалық талдау дегеніміз контекстте сөздердің, тұрақты тіркестердің, сөйлемдердің және пікірлердің мағынасын талдау [4].

Есептің қойылуы

нормалданған сөйлем – қазақ тілінің жай сөйлемдерінің құрылым ережесіне сәйкес(19)

Түсініктер

•Пікір – фактілерді түсіндіру және оларға эмоционалдық қатынас бойынша толық объективті болмайтын қандай да бір тақырыпқа қатысты көзқарас, бағалау, сезім, әсер, ой қорыту, қорытындылау және шешім шығару туралы түсінік.

Түсініктер

- Мәтінді сентимент талдау – мәтінге реңдік бояу беріп тұрған сөздерді және мәтінде жазылған объектіге, оқиғаға, құбылысқа, үрдіске немесе оның қасиеттеріне қатысты пікірдің эмоционалды бағасын автоматты түрде анықтауға арналған мағыналық талдау (контент анализ) әдістерінің тобы.