EURASIAN NATIONAL UNIVERSITY

ҚАЗАҚСТАН РЕСПУБЛИКАСЫ
БІЛІМ ЖӘНЕ ҒЫЛЫМ МИНИСТРЛІГІ
Л.Н. ГУМИЛЕВ АТЫНДАҒЫ ЕУРАЗИЯ
ҰЛТТЫҚ УНИВЕРСИТЕТІ

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
РЕСПУБЛИКИ КАЗАХСТАН
ЕВРАЗИЙСКИЙ НАЦИОНАЛЬНЫЙ УНИВЕРСИТЕТ
ИМЕНИ Л.Н. ГУМИЛЕВА

MINISTRY OF EDUCATION AND SCIENCE
OF THE REPUBLIC OF KAZAKHSTAN
L.N. GUMILYOV EURASIAN NATIONAL UNIVERSITY

16-18 маусым
Нұр-Сұлтан, 2022

«TURKLANG 2022»

«Түркі тілдерін компьютерлік өңдеу»
атты X халықаралық конференция
ЕҢБЕКТЕРІ

ТРУДЫ
X Международной конференции
«Компьютерная обработка тюркских языков»

«TURKLANG 2022»

PROCEEDINGS
of the X International Conference
on Computer processing of Turkic Languages

«TURKLANG 2022»

ҚАЗАҚСТАН РЕСПУБЛИКАСЫ
БІЛІМ ЖӘНЕ ҒЫЛЫМ МИНИСТРЛІГІ
Л.Н. ГУМИЛЕВ АТЫНДАҒЫ ЕУРАЗИЯ ҰЛТТЫҚ УНИВЕРСИТЕТІ

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
РЕСПУБЛИКИ КАЗАХСТАН
ЕВРАЗИЙСКИЙ НАЦИОНАЛЬНЫЙ УНИВЕРСИТЕТ
ИМЕНИ Л.Н. ГУМИЛЕВА

MINISTRY OF EDUCATION AND SCIENCE OF
THE REPUBLIC OF KAZAKHSTAN
L.N. GUMILYOV EURASIAN NATIONAL UNIVERSITY

«TURKLANG 2022»
«Түркі тілдерін компьютерлік өңдеу»
атты X халықаралық конференция
ЕҢБЕКТЕРІ
16-18 маусым 2022 ж.

ТРУДЫ
X Международной конференции
«Компьютерная обработка тюркских языков»
«TURKLANG 2022»
16-18 июня 2022 г.

PROCEEDINGS
of the X International Conference
on Computer processing of Turkic Languages
«TURKLANG 2022»
16-18 June 2022

Нұр-Сұлтан, 2022

УДК 80/81:004
ББК 81.2:32-973
Т 90

Техникалық редакция:
Ергеш Б.Ж.
Елибаева Г.К.
Турсынова Н.А.

Т 90 ТҮРКІ ТІЛДЕРІН КОМПЬЮТЕРЛІК ӨҢДЕУ. X халықаралық конференция: Еңбектері = КОМПЬЮТЕРНАЯ ОБРАБОТКА ТЮРКСКИХ ЯЗЫКОВ. X международная конференция: Труды. / - Нұр-Сұлтан: «Булатов А.Ж.» ЖК, 2022.= Нур-Султан: ИП «Булатов А.Ж.»

ISBN 978-601-326-645-9

Жинақта «Түркі тілдерін компьютерлік өңдеу» атты X халықаралық конференция қатысушыларының баяндамалары енген.

Компьютерлік лингвистика бағыты бойынша оқитын студенттерге, магистранттарға, докоранттарға және мамандарға арналған.

Жинақ «BR11765535» Қазақ тілі мәдениетін арттыру және функцияларды кеңейту бойынша ғылыми-лингвистикалық негіздер мен IT-ресурстарды әзірлеу» бағдарламасы есебінен жарияланды.

В сборнике представлены доклады участников X международной конференции «Компьютерная обработка тюркских языков».

Предназначен для студентов, магистрантов, докторантов и специалистов специализирующихся в областях компьютерной лингвистика.

Сборник издан за счет средств программы BR11765535 «Разработка научно-лингвистических основ и IT-ресурсов по расширению функций и повышению культуры казахского языка».

УДК 80/81:004
ББК 81.2:32-973

ISBN 978-601-326-645-9

ӘОК 004.8
[1]Оралбекова І.Т., [2]Ергеш Б.Ж.
Л.Н. Гумилев атындағы Еуразия ұлттық университеті
Нұр-Сұлтан, Қазақстан
[1]iinkar9822@gmail.com, [2]b.yergesh@gmail.com

ҚАЗАҚ ТІЛІНДЕГІ ҚОНАҚҮЙЛЕР ТУРАЛЫ ПІКІРЛЕРДІҢ РЕҢКІН АСПЕКТІЛЕРГЕ ТАЛДАУ

Аңдатпа. Интернетте пайдаланушыларға өз пікірлерімен алмасуға және тауарлар мен қызметтердің барлық түрлері туралы пікірлер қалдыруға мүмкіндік беретін көптеген платформалар бар. Бұл пікірлер басқа пайдаланушылар үшін ғана емес, сонымен қатар өздерінің беделін қадағалап, өнімдері мен қызметтері туралы дер кезінде кері байланыс алғысы келетін компаниялар үшін де пайдалы болуы мүмкін. Бұл саладағы мәселенің ең егжей-тегжейлі тұжырымы пайдаланушының жалпы объектіге ғана емес, сонымен қатар оның жеке аспектілеріне деген қатынасын анықтайтын аспектілі-бағытталған сентимент талдауда қойылады. Мақалада машиналық оқыту негізінде қонақүй туралы пікірлерді аспектілерге бөлу мәселесін шешу қарастырылады.

Түйін сөздер: машиналық оқыту, сентимент талдау, қонақүй туралы пікілер, Naive Bayes, SVM.

УДК 004.8
[1]Оралбекова І.Т., [2]Ергеш Б.Ж.
Евразийский национальный университет им. Л. Н. Гумилева
Нур-Султан, Казахстан
[1]iinkar9822@gmail.com, [2]b.yergesh@gmail.com

АНАЛИЗ ТОНАЛЬНОСТИ КОМЕНТАРИЕВ ОБ ОТЕЛЯХ НА КАЗАХСКОМ ЯЗЫКЕ

Аннотация. В Интернете существует множество площадок, которые позволяют пользователям делиться своим мнением и оставлять комментарии обо всех видах товаров и услуг. Эти комментарии могут быть полезны не только другим пользователям, но и компаниям, которые хотят следить за своей репутацией и своевременно получать отзывы о своих продуктах и услугах. Наиболее детальная постановка проблемы в этой области производится при анализе аспектно-ориентированных настроений, определяющих отношение пользователя не только к общему объекту, но и к его отдельным аспектам. В статье

рассматривается решение задачи разделения мнений об отеле на аспекты на основе машинного обучения.

Для определения эффективности программы было проведено экспериментальное исследование и проанализированы результаты. В ходе эксперимента использовались методы обучения учителей машинному обучению: метод наивного Байеса, линейный классификатор SVM и классификаторы логистической регрессии.

Результаты могут быть использованы для мониторинга общественного мнения, проведения маркетинговых кампаний, оценки новостных событий, прогнозирования мнений на основе проанализированных текстов, выявления эмоционального насилия. Анализ настроений позволяет компаниям или любому предпринимателю изменить комплекс маркетинговых мероприятий для улучшения положения продукта на рынке, выявить сильные и слабые стороны своих продуктов и услуг конкурентов и фирм. Анализ настроений на уровне аспектов обычно представляет собой подробный (конкретный) уровень, необходимый для практического применения. На этом основаны многие промышленные системы. Несмотря на большую работу в исследовательском сообществе и создание множества систем, проблема до сих пор решается. Каждая внутренняя задача остается очень сложной задачей.

Ключевые слова: машинное обучение, сентимент анализ, коментарии о гостиницах, Naive Bayes, SVM.

UDC 004.8

[1]Oralbekova I., [2]Yergesh B.
L. N. Gumilyov Eurasian National University
Nur-Sultan, Kazakhstan
[1]iinkar9822@gmail.com , [2]b.yergesh@gmail.com

ANALYSIS OF THE TONALITY OF COMMENTS ABOUT HOTELS IN THE KAZAKH LANGUAGE

Abstract. There are many platforms on the Internet that allow users to share their opinions and leave comments about all kinds of goods and services. These comments can be useful not only to other users, but also to companies that want to monitor their reputation and receive feedback on their products and services in a timely manner. The most detailed statement of the problem in this area is made when analyzing aspect-oriented moods that determine the user's attitude not only to the general object, but also to its

individual aspects. The article deals with the solution of the problem of dividing opinions about the hotel into aspects based on machine learning.

To determine the effectiveness of the program, a pilot study was conducted and the results analyzed. During the experiment, methods for teaching machine learning to teachers were used: the naive Bayes method, the SVM linear classifier, and logistic regression classifiers.

The results can be used to monitor public opinion, conduct marketing campaigns, evaluate news events, predict opinions based on analyzed texts, and identify emotional abuse. Sentiment analysis allows companies or any entrepreneur to change the marketing mix to improve the position of the product in the market, to identify the strengths and weaknesses of their products and services of competitors and firms. Aspect-level sentiment analysis is usually the detailed (specific) level required for practical application. Many industrial systems are based on this. Despite a lot of work in the research community and the creation of many systems, the problem is still being solved. Every internal task remains a very difficult task.

Keywords: Machine learning, sentiment analysis, hotel reviews, Naive Bayes, SVM.

Кіріспе

Интернет-ресурстар – қарым-қатынасқа, пікірталасқа және жаңа ақпаратты іздеуге ыңғайлы алаң. Белгілі бір реңкті қамтитын пайдаланушы пікірлері өздерінің беделін қадағалап, өз өнімдері мен қызметтері туралы дер кезінде кері байланыс алғысы келетін компаниялар үшін өте маңызды. Бұл саладағы мәселенің ең егжей-тегжейлі тұжырымы – пайдаланушының жалпы объектіге ғана емес, сонымен қатар оның жеке аспектілеріне деген қатынасын анықтайтын аспектіге бағытталған сентимент талдау (АБСТ). Мысалы, мейрамхананың шолуында «Менің қызмет көрсетуге ешқандай шағымым жоқ және маған мұндай интерьер ұнайды, бірақ француздық сиыр еті дәмсіз болды», үш аспектіні ажыратуға болады - қызмет көрсету, интерьер және тағам. Аспекттердің реңктері әртүрлі болуы мүмкін. Әрбір аспект әртүрлі сөздер немесе сөз тіркестері арқылы көрсетіыледі, олар аспектілік терминдер (АТ) деп аталады. Келтірілген мысалда АТ «қызмет», «интерьер» сөздері және «француздық сиыр еті» тіркесі болып табылады [1].

АБСТ саласындағы зерттеулерге шолу

Жұмыс [1] АБСТ-дың барлық ішкі міндеттерін шешуге арналған тәсілдер мен әдістердің егжей-тегжейлі шолуын ұсынады. Осы зерттеу нысаны болып табылатын АТ алудың қосалқы міндетін шешу үшін келесі тәсілдер қолданылады:

– бақыланатын машиналық оқыту әдістерін пайдаланып аспектілерді алу [3], [4];

– жиі кездесетін зат есімдер мен есімді тіркестерді (requent nouns and noun phrases) іздеу [2], [7], [10];

– пікір мен объектінің арақатынасына негізделген аспектілерді шығару (relation-based methods) [2];

– тақырыптық модельдеу (topic modeling) арқылы аспектілерді шығару [6].

Классификаторлар

Белгілі болғандай, мәтіннің тоналдылығын талдау үшін көбінесе Naive Bayes классификаторы және SVM әдісі қолданылады. Бұл ретте мәселенің типтік тұжырымы – фильмдер немесе тауарлар туралы, саяси тұлғалар немесе оқиға туралы (блогта, твиттерде және т.б.) пікірлерді оң және теріс болып жіктеу қажет. Naive Bayes моделі әдетте негізгі, ең қарапайым үлгі ретінде пайдаланылады, ал SVM әдісі күрделірек, өйткені ол тиімдірек деп саналады. Дегенмен, Бермингем мен Смитон [1] және Ванг пен Мэннинг [9] Naive Bayes классификаторы твиттер сияқты қысқа мәтіндерде SVM-ге қарағанда жақсырақ жұмыс істейтінін көрсетті. Сонымен қатар, кейбір жұмыстарда сөздердің векторлық көрінісі бар нейрондық желілерді белгілер ретінде қолдануға негізделген тәсіл ұсынылады. Бұл тәсіл тілге тәуелді емес және семантикалық тезаурусты қажет етпейді. Жұмыстың бөлігі ретінде Naive Bayes классификаторы, логистикалық регрессия және Linear SVM сияқты классификаторларымен эксперименттер жүргізілді.

Пікірлер корпусы

Пікірлер (www.TripAdvisor.com) және (www.booking.com) сайттарынан алынды.

Пікірледің ұзақтығы 1-ден 10 сөйлемге дейін өзгереді, орташа есеппен 5 сөйлемді құрайды. Корпустың шағын бөлігі кейінгі машиналық оқыту үшін белгіленеді. Бұл корпусқа Нұр-Сұлтан қаласындағы 5 қонақүйдің 100 пікірі кіреді.

Қонақүй аспектілерін анықтау

Бұл жұмыста объектілердің аспектілерінің тізімі пікірден автоматты түрде алынбайды, бірақ көңіл-күйді талдау модулі болатын жүйені пайдаланушылардың қажеттіліктері негізінде қолмен құрастырылатын тәсіл қолданылады. Пікірлерден алынған аспектілер тізімінде мекеме түрі мен асханасы, тағам мен қызмет көрсету сапасы, жайлылықтың болуы, романтикалық атмосфера, бардың, би алаңының, балалар бөлмесінің болуы, жақын жерде сауда үйінің болуы, автотұрақ және т.б кіреді. Осы жұмыста қарастырылатын қонақүйдің аспектілері кестеде көрсетілген (Кесте 1). Әр аспекті үшін кестеде осы аспект

қабылдай алатын мәндер жиынтығы көрсетілген. Қонақүйдің объективті аспектілері де (мысалы, би алаңының, бардың, балалар бөлмесінің және т.б. болуы) және субъективтік аспектілері бар . Объективті аспектілер үшін аспектілерді жіктеу тапсырмасы ақпаратты алу тапсырмасы болып табылады, ал субъективті аспектілер үшін бұл сезімді талдау тапсырмасы.

Бұл жұмыста 3 баллдық шкаламен бағаланатын { -1; 0; 1 } мәндерімен белгіленетін аспектілер қарастырылады және реңкті талдау мәселесі шешіледі.

Кесте 1

Қонақүй аспектілері (Аспекты гостиниц / Aspects of hotels)

Аспект	**Мәндер жиыны**
Food	{-1; 0; 1;}
Location	{-1; 0; 1;}
Room	{-1; 0; 1;}
Service	{-1; 0; 1;}
General	{-1; 0; 1;}

– Әрбір пікірге сәйкес реңк беріледі: Оң,теріс және бейтарап;
– Сөйлем немесе пікір бірнеше аспектілерден тұруы мүмкін;
– Тапсырма: пікірлерді аспектіге бағытталған сентимент талдау.

Бағалау өлшемдері

Accuracy (Дәлдік) – кең таралған және түсінуге оңай метрика. Бұл барлық дұрыс болжамдардың барлық болжанған үлгілердің жалпы санына қатынасы. Бірқатар тапсырмаларда дәлдік ақпаратсыз болуы мүмкін.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Дәлдік (precision) – бұл барлық оң болжамды объектілер үшін шын мәнінде оң нәтиже болып табылатын болжамды оң нәтижелердің үлесі. *Precision* формула (2) арқылы есептелінеді.

$$precision = \frac{TP}{TP+FP} \quad (2)$$

Толықтық (recall) – барлық шынайы-оң болжанған объектілердің шын мәнінде оңды объектілердің жалпы санына пропорциясы. Яғни, толықтық барлық оң мысалдардың қанша үлгі дұрыс жіктелгенін көрсетеді. Ол формула (3) арқылы есептелінеді.

$$recall = \frac{TP}{TP+FN} \quad (3)$$

F-өлшем – толықтық пен дәлдіктің өлшенген гармоникалық ортасы. Бұл көрсеткіш модель қанша жағдайды дұрыс болжайтынын және үлгінің қанша шынайы дананы өткізіп жібермейтінін көрсетеді.

Precision және recall, неғұрлым жоғары болса, соғұрлым жақсы екені анық.

Бірақ нақты өмірде максималды дәлдік пен толықтыққа бір уақытта қол жеткізу мүмкін емес, белгілі бір тепе-теңдікті іздеу керек.

Сондықтан, біз алгоритмнің дәлдігі мен толықтығы туралы ақпаратты біріктіретін белгілі бір метрикаға ие болғымыз келеді. Бұл жағдайда бізге қандай жүзеге асыруды өндіріске енгізу туралы шешім қабылдау оңайырақ болады. F-өлшемі дәл осындай метрика [10].

F-өлшем пайдаланылатын модельдің толықтығы мен дәлдігі туралы ақпаратты біріктіреді. F-өлшем формула (4) арқылы есептелінеді.

$$F = 2\frac{Precision * Recall}{Precision + Recall} \quad (4)$$

Тәжірибе нәтижесі

Бағалау шаралары precision, recall, f1-score және accuracy қамтиды. Нәтижелер төменде кесте түрінде (Кесте 2) берілген.

Кесте 2

Тәжірибе нәтижесі (Результат эксперимента / Experiment result)

Classifier		**Naive bayes**	**Logistic regression**	**Linear SVM**
Accuracy		0.7	0.7	**0.75**
Service	Precision	0.75	0.60	0.60
	Recall	1.00	1.00	1.00
	F1- score	**0.86**	0.75	0.75
General	Precision	0.67	0.67	0.67
	Recall	0.80	0.80	0.80
	F1- score	0.73	0.73	**0.75**
Food	Precision	0.67	0.67	1.00
	Recall	0.67	0.67	0.67
	F1- score	0.67	0.67	**0.80**
Room	Precision	0.67	0.75	0.75
	Recall	1.00	0.75	0.75
	F1- score	**0.80**	0.75	0.75
Location	Precision	1.00	1.00	1.00
	Recall	0.20	0.40	0.60
	F1- score	0.33	0.57	**0.75**

Linear SVM классификаторын қолдану әдісі ең тиімді екені анықталды. Оның accuracy мәні 0,75-ке тең болды. Яғни, ол 75 % дәлдікпен жұмыс жасайды. Naive bayes әдісі мен Logistic regression әдісі де аса төмен нәтиже көрсеткен жоқ. Олардың accuracy мәндері 0,7-ге тең. Linear SVM әдісі ‘General’ , ‘Food’ және ‘Location’ аспектілері үшін жақсы мән көрсетті. General – ‘0,75’-ке , Food – ‘0.8’-ге , Location – ‘0.75’-ке тең. Ал Naive bayes әдісі ‘Service’ және ‘Room’ аспектілерінде жоғары нәтиже берді. Service – ‘0.86’-ға, Room – ‘0,80’-ге тең болды.

Қорытынды

Программаның тиімділігін анықтау үшін эксперименттік зерттеу жүргізілді және соған сәйкес нәтижелерге талдау жасалды. Тәжірибе барысында Naive bayes әдісі, Linear SVM классификаторы және Logistic Regression классификаторлары қолданылды.

Нәтижелер мәліметтің көлеміне байланысты өзгеруі мүмкін. Деректер жиыны неғұрлым көбірек болса, соғұрлым дәлірек жұмыс істейді. Аспектілі бағытталған сентимент талдауды іске асыру үшін осы Naive bayes әдісін және SVM классификаторын қолдану тиімді екені анықталды. Бұл әдістер қарапайым және ыңғайлы болып табылады.

Алынған нәтижерлерді қоғамдық пікірді бақылау, маркетингтік науқандар жүргізу, жаңалықтар оқиғаларын бағалау, талданған мәтіндер негізінде пікірлерді болжау, эмоционалды теріс қылықтарды анықтау үшін қолдануға болады. Сентимент талдау компанияларға немесе кез келген кәсіпкерлерге өзінің және бәсекелес фирмалардың тауарларының немесе қызметтерінің күшті және әлсіз жақтарын анықтау үшін, нарықтағы өнімнің позициясын жақсарту үшін маркетингтік іс-шаралар кешенін өзгертуге мүмкіндік береді. Аспект деңгейіндегі сентиментті талдау, әдетте, практикалық қолдану үшін қажет егжей-тегжейлі(нақты) деңгей болып табылады. Көптеген өнеркәсіптік жүйелер осыған негізделген. Зерттеу қоғамдастығында көп жұмыс жасалып, көптеген жүйелер құрылғанына қарамастан, мәселе әлі де шешіліп жатыр. Әрбір ішкі тапсырма өте қиын міндет болып қала береді.

Әдебиеттер тізімі

1. Bermingham A., Smeaton A. Classifying Sentiment in Microblogs: Is Brevity an Advantage? // Proceedings of the International Conference on Information and Knowledge Management (CIKM), 2010.

2. Hu M., Liu B. Mining and summarizing customer reviews. Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining, 2004, pp. 168–177.

3. Ivanov V., Tutubalina E., Mingazov N., Alimova I. Extracting Aspects, Sentiment and Categories of Aspects in User Reviews about Restaurants and Cars.

Proceedings of the 21st International Conference on Computational Linguistics (Dialog-2015), 2015, pp. 46–57.

4. Jakob N., Gurevych I., Extracting opinion targets in a single-and cross-domain setting with conditional random fields, Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 2010, pp. 1035-1045.

5. Liu B., Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 2012, pp. 1–167.

6. Mukherjee A, Liu B. Aspect extraction through semi-supervised modeling. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, 2012, volume 1, pp. 339-348.

7. Popescu A. M., Nguyen B., Etzioni O. OPINE: Extracting product features and opinions from reviews. Proceedings of HLT/EMNLP on interactive demonstrations, 2005, pp. 32–33.

8. Scaffidi C., Bierhoff K., Chang E., Felker M., Ng H., Jin C. Red Opal: product-feature scoring from reviews. Proceedings of the 8th ACM conference on Electronic commerce, 2007, pp. 182–191.

9. Wang S., Manning Ch. D. Baselines and Bigrams: Simple, Good Sentiment and Topic Classification // Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL): Short Papers – vol. 2, pp. 90–94, 2012.

10. Дудченко П. В. Метрики оценки классификаторов в задачах медицинской диагностики / П. В. Дудченко // Молодежь и современные информационные технологии : сборник трудов XVI Международной научно-практической конференции студентов, аспирантов и молодых учёных, 3-7 декабря 2018 г., г. Томск. – Томск : Изд-во ТПУ, 2019. – [С. 164-165].