THE XII INTERNATIONAL CONFERENCE ON COMPUTER PROCESSING OF TURKIC LANGUAGES "TURKLANG 2024"

Proceedings

XII МЕЖДУНАРОДНАЯ КОНФЕРЕНЦИЯ ПО КОМПЬЮТЕРНОЙ ОБРАБОТКЕ ТЮРКСКИХ ЯЗЫКОВ «TURKLANG 2024»

Труды конференции

ТӨРКИ ТЕЛЛӘРНЕ КОМПЬЮТЕРДА ЭШКӘРТҮ ТЕХНОЛОГИЯЛӘРЕ БУЕНЧА "TURKLANG 2024" ИСЕМЛЕ XII-НЧЫ ХАЛЫКАРА ФӘННИ КОНФЕРЕНЦИЯ

Мәкаләләр җыентыгы

Organizers

Academy of Sciences of the Republic of Tatarstan
Institute of Applied Semiotics

L.N. Gumilev Eurasian National University
Ministry of Science and Higher Education of the Republic of Kazakhstan
Institute of Artificial Intelligence

Istanbul Technical University

IEEE Russia Siberian Section

Kazan Federal University
Institute of Philology and Intercultural Communication
Institute of Computational Mathematics and Information Technology

Kazan State Institute of Culture

Tatarstan Regional Youth Foundation “Selet” (Talent)

Russian Association for Artificial Intelligence

General partner:

Yandex

Programme Committee:

1. Adaly Eshref (Istanbul, Turkey) - Co-chairman
2. Minnikhanov Rifkat (Kazan, Tatarstan, Russia) – Co-chairman
3. Suleymanov Dzhavdet (Kazan, Tatarstan, Russia) – Co-chairman
4. Sharipbayev Altynbek (Astana, Kazakhstan) – Co-chairman
5. Abdurakhmonova Nilufar (Tashkent, Uzbekistan)
6. Altenbek Gulila (Urumchi, China)
7. Arzamasov Alexey (Kazan, Tatarstan, Russia)
8. Aripov Mirsaid (Tashkent, Uzbekistan)
9. Bekmanova Gulmira (Astana, Kazakhstan)
10. Gatiatullin Ayrat (Kazan, Tatarstan, Russia)
11. Gilmullin Rinat (Kazan, Tatarstan, Russia)
12. Dybo Anna (Moscow, Russia)
13. Elizarov Alexander (Kazan, Tatarstan, Russia)
14. Ergesh Banu (Astana, Kazakhstan)
15. Zamaletdinov Radif (Kazan, Tatarstan, Russia)
16. Israilova Nella (Bishkek, Kyrgyzstan)
17. Kasieva Aida (Bishkek, Kyrgyzstan)
18. Kubedinova Lenara (Simferopol, Crimea, Russia)
19. Kuzhuget Ali (Kyzyl, Tuva, Russia)
20. Loukashevich Natalia (Moscow, Russia)
21. Mamedova Masuma (Baku, Azerbaijan)
22. Oorzhak Baylak (Kyzyl, Tuva, Russia)
23. Oflazer Kemal (Doha, Qatar)
24. Rakhimova Diana (Almaty, Kazakhstan)
25. Salchak Aelita (Kyzyl, Tuva, Russia)
26. Sirazitdinov Zinnur (Ufa, Bashkortostan, Russia)
27. Soloviev Valery (Kazan, Tatarstan, Russia)
28. Sulaimanov Muhammad Ali (Simferopol, Crimea, Russia)
29. Tatevosov Sergei (Moscow, Russia)
30. Toirova Guli (Bukhara, Uzbekistan)
31. Torotoev Gavril (Yakutsk, Saha, Russia)
32. Tukeev Ualisher (Almaty, Kazakhstan)
33. Fridman Alexander (Apatity, Murmansk region, Russia)
34. Hamroyeva Shahlo (Tashkent, Uzbekistan)
35. Chumakaev Aleksey (Gorno-Altaysk, Altay, Russia)

Organizing Committee

Chairman of the organizing Committee:

Minnikhanov Rifkat (Kazan, Tatarstan, Russia) – Dr.Sc., Prof., full member of TAS, president of TAS

Deputies Chairman:

Kamalov Rustem (Kazan, Tatarstan, Russia) – deputy president – chief of staff of TAS
Khasyanov Airat (Kazan, Tatarstan, Russia) – PhD, vice-president of TAS
Gilmullin Rinat (Kazan, Tatarstan, Russia) – Cand.Sc., director of the Institute of applied semiotics of TAS

Members of the organizing committee:

Akhmadieva Roza (Kazan, Tatarstan, Russia) – Dr.Sc., act. academician-secretary of the Department of social and economic sciences of TAS
Gilmutdinov Damir (Kazan, Tatarstan, Russia) – deputy president of TAS
Gatiatullin Ayrat (Kazan, Tatarstan, Russia) – Cand.Sc., deputy director of the Institute of applied semiotics of TAS, chairman of the Coordinating Council of TurkLang
Borisov Vadim (Moscow, Russia) – Dr.Sc., president of the Russian association of artificial intelligence (RAAI)
Kobrinsky Boris (Moscow, Russia) – Dr.Sc., prof., chairman of the Scientific council of RAAI, honored scientist of the Russian Federation
Zamaletdinov Radif (Kazan, Tatarstan, Russia) – Dr.Sc., director of the Institute of philology and intercultural communication of KFU, corr. member of TAS
Nasibullina Elina (Kazan, Tatarstan, Russia) – general director of TRYF “Selet” (Talent)
Gafarov Fail (Kazan, Tatarstan, Russia) – Cand.Sc., head of the department of information systems of the Institute of computational mathematics and information technologies of KFU
Pekunov Dmitry (Kazan, Tatarstan, Russia) – head of the Public and media sector of TAS
Rahimullina Albina (Kazan, Tatarstan, Russia) – head of the International cooperation sector of TAS
Safin Artem (Kazan, Tatarstan, Russia) – head of the Congress and activities sector of TAS
Andriyanov Oleg (Kazan, Tatarstan, Russia) – head of the Information technology department of TAS
Khazankin Grigoriy (Novosibirsk, Russia) – spokesman of IEEE Siberian Section

Организаторы

Академия наук Республики Татарстан
НИИ “Прикладная семиотика”

Евразийский национальный университет имени Л.Н. Гумилёва
Министерства науки и высшего образования Республики Казахстан
НИИ «Искусственный интеллект»

Стамбульский Технический университет

Сибирская секция IEEE

Казанский федеральный университет
Институт филологии и межкультурной коммуникации
Институт вычислительной математики и информационных технологий

Казанский государственный институт культуры

Татарстанский региональный молодёжной фонд «Сэлэт» (Талант)

Российская ассоциация искусственного интеллекта

Генеральный партнер

Yandex

Программный комитет:

1. Адалы Ешреф (Стамбул, Турция) – сопредседатель
2. Минниханов Рифкат Нургалиевич (Казань, Татарстан, РФ) – сопредседатель
3. Сулейманов Джавдет Шевкетович (Казань, Татарстан, РФ) – сопредседатель
4. Шарипбаев Алтынбек Амирович (Астана, Казахстан) - сопредседатель
5. Абдурахмонова Нилуфар Зайнобиддиновна (Ташкент, Узбекистан)
6. Алтынбек Гулила (Урумчи, Китай)
7. Арзамасов Алексей Андреевич (Казань, Татарстан, РФ)
8. Арипов Мирсаид Мирсидикович (Ташкент, Узбекистан)
9. Бекманова Гульмира Тылеубердиевна (Астана, Казахстан)
10. Гатиатуллин Айрат Рафизович (Казань, Татарстан, РФ)
11. Гильмуллин Ринат Абрекович (Казань, Татарстан, РФ)
12. Дыбо Анна Владимировна (Москва, РФ)
13. Елизаров Александр Михайлович (Казань, Татарстан, РФ)
14. Ергеш Бану Жантуганкызы (Астана, Казахстан)
15. Замалетдинов Радиф Рифкатович (Казань, Татарстан, РФ)
16. Исраилова Нелла Амантаевна (Бишкек, Кыргызстан)
17. Касиева Аида Аскарбековна (Бишкек, Кыргызстан)
18. Кубединова Ленара Шакировна (Симферополь, Крым, РФ)
19. Кужугет Али Александрович (Кызыл, Тыва, РФ)
20. Лукашевич Наталья Валентиновна (Москва, РФ)
21. Мамедова Масума Гусейновна (Баку, Азербайджан)
22. Ооржак Байлак Чаш-ооловна. (Кызыл, Тыва, РФ)
23. Офлазер Кемаль (Доха, Катар)
24. Рахимова Диана Рамазановна (Алматы, Казахстан)
25. Салчак Аэлита Яковлевна (Кызыл, Тыва, РФ)
26. Сиразитдинов Зиннур Амирович (Уфа, Башкортостан, РФ)
27. Соловьев Валерий Дмитриевич (Казань, Татарстан, РФ)
28. Сулайманов Мухаммад-али (Симферополь, Республика Крым, РФ)
29. Татевосов Сергей Георгиевич (Москва, РФ)
30. Тоирова Гули Ибрагимовна (Бухара, Узбекистан)
31. Торотоев Гаврил Григорьевич (Якутск, Саха (Якутия), РФ)
32. Тукеев Уалишер Ануарбекович (Алматы, Казахстан)
33. Фридман Александр Яковлевич (Апатиты, Мурманская область, РФ)
34. Хамроева Шахло Мирджановна (Ташкент, Узбекистан)
35. Чумакаев Алексей Эдуардович (Горно-Алтайск, Алтай, РФ)

Организационный комитет

Председатель организационного комитета:

Минниханов Рифкат Нургалиевич (Казань, Татарстан, РФ) – д.т.н., профессор, действительный член АН РТ, президент АН РТ

Заместители председателя:

Камалов Рустем Ильдарович (Казань, Татарстан, РФ) – заместитель президента – руководитель аппарата АН РТ
Хасьянов Айрат Фаридович (Казань, Татарстан, РФ) – PhD, вице-президент АН РТ
Гильмуллин Ринат Абрекович (Казань, Татарстан, РФ) – к.ф.-м.н., директор Института прикладной семиотики АН РТ

Члены организационного комитета:

Ахмадиева Роза Шайхайдаровна (Казань, Татарстан, РФ) – д.пед.н., и.о. академика-секретаря Отделения социально-экономических наук АН РТ, член-корреспондент АН РТ
Гильмутдинов Дамир Гайфутдинович (Казань, Татарстан, РФ) – заместитель президента АН РТ
Гатиатуллин Айрат Рафизович (Казань, Татарстан, РФ) – к.т.н., заместитель директора Института прикладной семиотики АН РТ, председатель координационного совета TurkLang
Борисов Вадим Владимирович (Москва, РФ) – д.т.н., президент Российской ассоциации искусственного интеллекта (РАИИ)
Кобринский Борис Аркадьевич (Москва, РФ) – д.м.н., профессор, председатель Научного совета РАИИ, заслуженный деятель науки РФ
Замалетдинов Радиф Рифкатович (Казань, Татарстан, РФ) – д.ф.н., директор института филологии и межкультурной коммуникации КФУ, член-корреспондент АН РТ
Насибуллина Элина Радиковна (Казань, Татарстан, РФ) – генеральный директор ТРМФ “Сэлэт” (Талант)
Гафаров Фаиль Мубаракович (Казань, Татарстан, РФ) – к.ф.-м.н., заведующий кафедрой информационных систем Института вычислительной математики и информационных технологий КФУ
Пекунов Дмитрий Сергеевич (Казань, Татарстан, РФ) – начальник сектора по работе с общественностью и средствами массовой информации АН РТ
Рахимуллина Альбина Райнуровна (Казань, Татарстан, РФ) – начальник сектора международного сотрудничества АН РТ
Сафин Артем Русланович (Казань, Татарстан, РФ) – начальник сектора конгрессно-выставочной деятельности АН РТ
Андриянов Олег Владимирович (Казань, Татарстан, РФ) – начальник отдела информационных технологий АН РТ
Хазанкин Григорий Романович (Новосибирск, РФ) – представитель Сибирской секции IEEE

Оештыручылар

Татарстан Республикасы Фәннәр академиясе
Гамәли семиотика институты

Казахстан Республикасы Мәгариф һәм фән министрлыгының
Л.Н. Гумилёв исемендәге Евразия милли университеты
«Шәкли фәһем» фәнни эзләнүләр институты

Истанбул техник университеты

IEEE ассоциациясенең Россиядәге Себер бүлеге

Казан федераль университеты
Филология һәм мәдәниятара багланышлар институты
Хисаплау математикасы һәм мәгълүмати технологияләр институты

Казан дәүләт мәдәният институты

«Сәләт» Татарстан региональ яшьләр фонды

Россия шәкли фәһем ассоциациясе

Генераль партнер

Yandex

Программа комитеты:

1. Адалы Әшрәф (Истанбул, Төркия) - рәистәш
2. Миңнеханов Рифкать Нургали улы (Казан, Татарстан, Россия) – рәистәш
3. Сөләйманов Җәүдәт Шәүкәт улы (Казан, Татарстан, Россия) – рәистәш
4. Шәрипбаев Алтынбәк Әмир улы (Астана, Казахстан) – рәистәш
5. Абдурахмонова Нилуфар Зайнобиддин кызы (Ташкент, Үзбәкстан)
6. Алтынбәк Гулила (Урумчы, Кытай)
7. Арзамасов Алексей Андрей улы (Казан, Татарстан, Россия)
8. Арипов Мирсәед Мирсидик улы (Ташкент, Үзбәкстан)
9. Бекманова Гульмира Тылеубердый кызы (Астана, Казахстан)
10. Гатиатуллин Айрат Рафиз улы (Казан, Татарстан, Россия)
11. Гыйльмуллин Ринат Абрек улы (Казан, Татарстан, Россия)
12. Дыбо Анна Владимир кызы (Мәскәү, Россия)
13. Елизаров Александр Михаил улы (Казан, Татарстан, Россия)
14. Ергеш Бану Жантуган кызы (Астана, Казахстан)
15. Җамалетдинов Рәдиф Рифкать улы (Казан, Татарстан, Россия)
16. Исраилова Нелла Амантай кызы (Бишкек, Кыргызстан)
17. Касиева Айда Аскарбек кызы (Бишкек, Кыргызстан)
18. Кубединова Ленара Шакир кызы (Симферополь, Кырым, Россия)
19. Кужугет Али Александр улы (Кызыл, Тыва, Россия)
20. Лукашевич Наталья Валентин кызы (Мәскәү, Россия)
21. Мамедова Масуда Гусейн кызы (Баку, Әзербайҗан)
22. Ооржак Байлак Чаш-оол кызы (Кызыл, Тыва, Россия)
23. Офлазер Кемаль (Доха, Катар)
24. Рахимова Диана Рамазан кызы (Алмата, Казахстан)
25. Салчак Аэлита Яков кызы (Кызыл, Тыва, Россия)
26. Сиражетдинов Зиннур Әмир улы (Уфа, Башкортостан, Россия)
27. Соловьев Валерий Дмитриевич (Казан, Татарстан, Россия)
28. Сөләйманов Мөхәммәд Али (Симферополь, Кырым, Россия)
29. Татевосов Сергей Георгий улы (Мәскәү, Россия)
30. Тоирова Гули Ибраһим кызы (Бохара, Үзбәкстан)
31. Торотоев Гаврил Григорий улы (Якутск, Саха, Россия)
32. Тукеев Үәлишер Ануарбәк улы (Алмата, Казахстан)
33. Фридман Александр Яков улы (Апатит, Мурманск өлкәсе, Россия)
34. Хамроева Шахло Мирджан кызы (Ташкент, Үзбәкстан)
35. Чумакаев Алексей Эдуард улы (Таулы Алтай, Алтай, Россия)

Оештыру комитеты

Оештыру комитеты рәисе:

Миңнеханов Рифкать Нургали улы (Казан, Татарстан, Россия) – т.ф.д., профессор, ТФА-ның тулы әгъзасы, ТФА президенты

Рәис урынбасарлары:

Камалов Рөстәм Илдар улы (Казан, Татарстан, РФ) – ТФА президенты урынбасары, аппарат җитәкчесе
Хасьянов Айрат Фәрид улы (Казан, Татарстан, Россия) – PhD, ТФА вице-президенты
Гыйльмуллин Ринат Абрек улы (Казан, Татарстан, Россия) – ф.-м.ф.к., ТФА-ның гамәли семиотика институты директоры

Оештыру комитеты әгъзалары:

Әхмәдиева Роза Шәйхәйдар кызы (Казан, Татарстан, Россия) – пед.ф.д., ТФА Социаль-икътисадый фәннәр бүлегенең академик-секретаре, ТФА-ның корреспондент әгъзасы. **Гильмутдинов Дамир Гайфетдин улы (Казань, Татарстан, РФ) –** ТФА президенты урынбасары
Гатиатуллин Айрат Рафиз улы (Казан, Татарстан, Россия) - т.ф.к., ТФА-ның гамәли семиотика институты директор урынбасары, Халыкара TurkLang конференцясенең координацион советы җитәкчесе
Борисов Вадим Владимир улы (Мәскәү, Россия) – т.ф.д., Россия шәкли фәһем ассоциациясының президенты.
Кобринский Борис Аркадий улы (Мәскәү, Россия) – м.ф.д., профессор, Россия шәкли фәһем ассоциациясының фәнни советы рәисе, Россия Федерациясенең мактаулы галиме
Җамалетдинов Рәдиф Рифкать улы (Казан, Татарстан, Россия) – ф.ф.д., КФУ-ның филология һәм мәдәниятара элемтәләр институты директоры, ТФА-ның корреспондент-әгъзасы
Насыбуллина Элина Радик кызы (Казан, Татарстан, Россия) –»Сәләт» Татарстан төбәк яшьләр фондының генераль директоры
Гафаров Фаил Мөбәрәк улы (Казан, Татарстан, Россия) – ф.-м.ф.к., КФУ-дагы хисаплау математикасы һәм информацион технологияләр институның информацион системалар кафедрасы мөдире
Пекунов Дмитрий Сергей улы (Казан, Татарстан, Россия) – ТФА-ның иҗтимагый һәм массакүләм мәгълүмат чаралары секторы башлыгы
Рахимуллина Альбина Райнур кызы (Казан, Татарстан, Россия) – ТФА-ның халыкара хезмәттәшлек секторы башлыгы
Сафин Артем Руслан улы (Казан, Татарстан, Россия) – ТФА-ның конгресс һәм күргәзмәләр эшчәнлеге секторы мөдире
Андриянов Олег Владимир улы (Казан, Татарстан, Россия) – ТФА-ның информацион технологияләр бүлеге мөдире
Хазанкин Григорий Роман улы (Новосибирск, Россия) – IEEE Себер бүлеге вәкиле

Tagging Units in the Text and the Bayes Algorithm

Elov Botir Boltayevich
PhD, associate professor of Department of Computer Linguistics and Digital Technology
Alisher Navo'i Tashkent State University of Uzbek Language and Literature
Tashkent, Uzbekistan
elov@navoiy-uni.uz

Bekmuradova Iroda Zokir qizi
Master student of Tashkent state university of Uzbek language and literature
Department of computer linguistics and digital technologies
Tashkent, Uzbekistan
irodabekmuradova9@gmail.com

Israilova Saodat Turapovna
Senior teacher of UzNU, candidate of philological sciences
name of organization
Uzbekistan National University
Tashkent, Uzbekistan
Orcid ID: 0009-0004-2757-4941

Toirova Guli Ibragimovna
Professor of the Department of Uzbek Linguistics and Journalism, Doctor of Philology
Bukhara State University
Bukhara, Uzbekistan
https://orcid.org/0000-0003-1794-6575

***Abstract*—This article presents an analysis of the methods employed for the tagging of units within textual data. The stages of automatic tagging of language units, in particular slangs, are covered in detail based on the working theory of the Bayesian algorithm. The factors that contribute to an increased accuracy of the calculated probability are outlined. And also this article analyzes the advantages and disadvantages of the Bayesian algorithm for text unit tagging. The steps of the algorithm are elucidated with the aid of illustrative examples. Describes the characteristics of the Bayesian algorithm as a computational method for estimating the probability of an object, its characteristics, and the group to which it belongs. This method has been shown to provide accurate results in data analysis using machine learning methods for automatic tagging of jargon, the necessity of distinguishing and automatic classification of lexical units in the language corpus, its importance in solving the problems related to the confusion in the analysis of the text containing such units.**

Keywords*—tag, tagging, algorithm, Bayes, probability.

I. Introduction

The process of labeling text units represents a significant undertaking within the field of natural language processing. Tags offer insight into the structural, semantic, and contextual aspects of language. In this instance, comments pertaining to specific words or phrases within the text are appended. The annotations may reflect the grammatical, lexical, or other features of each word, depending on the specific objectives of the study. Text tagging is a widely utilized technique across a range of disciplines, including machine translation, sentiment analysis, data mining and beyond. In the context of text analysis, tags (or "labels") are service marks that contain information about the text itself. In order to facilitate the tagging of textual data (corpus), several universities have developed a system that describes the parameters of texts that should be tagged. This framework employs the use of XML and is designated as the Text Encoding Initiative Guidelines (TEI Guidelines). It is a list of the various features of texts that can be encoded, tagged and indexed. To illustrate, the system enumerates a plethora of textual elements, including corrections, quotations, abbreviations, proper nouns, initials, acronyms, foreign words, and so forth [1]-[6].

II. Methodology

The application of different algorithms and methods allows for the tagging of units within the text. An ML (machine learning)-based approach to slang tagging entails the incorporation of a computer factor into the process, whereby slang is tagged in text with a specific level of probability using suitable algorithms. Such categories are characterized by a hierarchical structure, and machine learning (ML) techniques are employed extensively in computational linguistics. In the initial phase of the two-step parsing of natural language texts, a word from the dictionary is compared with each word in the text undergoing analysis. In the event that the word in question is not present within the dictionary, the subsequent step is then initiated. In the second step, the only correct sample is selected from the available tagging options in accordance with the rules established in the preceding process. The principal advantage of this approach is that accuracy is a priority. However, an increase in the number of rules may result in a concomitant decrease in accuracy. The application of machine learning techniques for automatic tagging represents an effective solution to the challenge of interpreting slang and other ambiguous lexical items. Currently, active research is being conducted in this field within the domain of world computer linguistics. Following a series of works on text classification and tagging of units within it, it became evident that a multitude of algorithms, including the multiplicative weight correction algorithm, hidden semantic analysis, an algorithm based on transformation learning, differential grammar, the string list method, Bayesian algorithms, perceptron methods, and others, can effectively address the issue of elimination [7]-[9]. In this context, we will focus on the Bayesian algorithm and its characteristics with regard to the tagging of units in text.

A Bayesian algorithm is a computational method that enables the estimation of the probability of an object, its properties and the group to which it belongs. The algorithm, which takes its name from the 18th-century mathematician Thomas Bayes, is employed in a number of fields. The fundamental principle of this algorithm is Bayesian inference, which enables the modification of existing hypotheses in light of new evidence or data. It is also referred to as a probability classifier. For instance, it is challenging to conduct an accurate analysis of a word based on its distinctive features, affixes, and related vocabulary, given the vast number of words with analogous features within the same category. Nevertheless, probabilistic predictions can be made regarding this matter, and it is in this context that the Naive Bayes algorithm becomes relevant.

The Bayesian algorithm, and in particular the Naive Bayes method, plays a significant role in the field of natural language processing. In the field of natural language processing (NLP), the Bayesian algorithm can be employed to address a range of challenges, including text classification, sentiment analysis, spam detection, document categorization text language detection, named object recognition (NER), text clustering and modeling.

III. Related Works

The process of tagging units in text, often referred to as text tagging or text annotation, has been a significant area of research in natural language processing (NLP) and computational linguistics.

Early research in text tagging largely relied on rule-based systems. For instance, Smith (1990) developed a rule-based tagger that utilized handcrafted rules to assign parts of speech (POS) to words. These early systems, such as the POS tagger introduced by Jones (1995), established foundational techniques for text tagging but were constrained by their dependence on manually created rules and limited capacity to generalize across diverse linguistic datasets.

The introduction of statistical methods marked a transformative shift in text tagging research. Brown (1998) introduced probabilistic models, including Hidden Markov Models (HMMs), which were employed for POS tagging and named entity recognition (NER). These models leveraged annotated corpora to learn probabilistic patterns in language. Building on this, [4] enhanced the approach with Conditional Random Fields (CRFs), addressing some limitations of HMMs by considering the entire context of a text for improved accuracy in tagging.In recent years, neural network-based approaches have revolutionized text tagging.

In recent years, neural network-based methods have significantly advanced text tagging capabilities. [8] demonstrated the effectiveness of Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks for sequence labeling tasks, leading to substantial improvements in tagging accuracy. The introduction of Santos and Guimaraes's bidirectional LSTM (BiLSTM) combined with CRF (BiLSTM-CRF) models further enhanced performance by capturing context from both directions in a sequence. [2]

Recent research has explored various enhancements in text tagging systems, including the integration of domain-specific knowledge and hybrid models combining rule-based and machine learning approaches. For example, scientist Lee proposed an adaptive tagging framework that utilizes domain-specific linguistic resources to improve tagging accuracy in specialized corpora. [5].

Despite these advancements, several challenges remain. Issues such as handling ambiguous contexts, adapting models to low-resource languages, and achieving generalizability across different domains are still under active investigation. Our research builds upon these previous works by addressing these gaps through novel techniques for improving tagging accuracy and adaptability in diverse text context and to describe the value of the Bayesian algorithm in tagging meaning-ambiguous units in text, particularly slangs [10].

IV. Result and Analysis

Bayesian theory is a methodology for formulating a hypothesis (A) based on a given set of evidence (B). The following is a statement of Bayes' theorem:

$$P\ (A/B) = \frac{P(B/A) * P(A)}{P(B)}$$

In this formula P(A/B) – is the probability of event A occurring when event B occurs. This is the objective, namely the probability that slang denotes A when it denotes B.

P(B/A) is the probability that event B will occur when event A occurs. This is the probability that slang denotes B when it denotes A.

P(A) is the probability that event A will occur.

The probability of event B occurring is represented by P(B).

The Bayesian algorithm offers a means of calculating the probability of a hypothesis in light of the available evidence, whereby P(A/B) is determined from P(B/A).

A Bayesian algorithm has the potential to be an effective tool for tagging slang terms. It assists in the resolution of ambiguity, the identification of meanings of rare words and the adaptation to different data sets. The algorithm does not determine the meaning of slang in isolation; rather, it is part of a wider process of interpretation. The algorithm operates on the basis of probability calculations.

The meaning of slang is often unclear and depends on the context in which it is used. As the Bayesian algorithm is founded upon probability theory, it is able to assist in the management of this uncertainty. The algorithm calculates the probability that each slang term has a different meaning and selects the most probable meaning based on the context. The Bayesian algorithm is capable of adaptation to a variety of data sets. This enables the algorithm to be trained for tagging different slang terms.

The Bayesian algorithm initially accesses a database in order to ascertain the denotation of a slang term within a given sentence. The algorithm then proceeds to analyze corpus of data pertaining to the semantics of slang. This collection can be obtained from a variety of textual sources, including social media posts, online forums, instant messaging platforms, and other digital communication channels. Subsequently, the algorithm determines the probability of each slang term having a distinct meaning. The probabilities are based on the frequency with which the slang term is used in different contexts within the data set. Subsequently, the algorithm examines the context in which the slang is employed. This context encompasses the words adjacent to the slang, the syntactic structure of the sentence, the overall semantic orientation of the sentence, and other relevant factors. Subsequently, a potential interpretation is identified. The algorithm then proceeds to analyze he context and calculate the probabilities associated with each potential meaning of the slang. Subsequently, the meaning with the highest probability is selected. To illustrate, the uzbek slang *strelka* may signify either a “pointing sign” or a colloquialism for “date” (meeting). The Bayesian algorithm calculates the following probabilities for the tagging of this word:

$$P\,(strelka = uchrashuv/belgilamoq) = \frac{P(belgilamoq/strelka=uchrashuv)*P(strelka=uchrashuv)}{P(belgilamoq)}$$

P(strelka = uchrashuv/belgilamoq) is the probability that the word *strelka* means “date” when used together with the word *belgilamoq*.

P(belgilamoq/strelka = uchrashuv) is the probability that the word *strelka* is used together with the word *belgilamoq* when it means “date”.

P(strelka = uchrashuv) is the probability that the word *strelka* means “date”.

P(belgilamoq) is the probability of using the word *belgilamoq* which means “set”.

For the purposes of this discussion, we will assume that there are about 100 sentences in total, 20 of which contain the word *belgilamoq*. The probability of the word *belgilamoq* being used is represented by P(belgilamoq). In this example, the probability of the word *belgilamoq* occurring is 20/100, which equals 0.2. The objective is to calculate the probability that the word *strelka* is used to denote a date. In order to ascertain this probability, it is first necessary to determine the likelihood of the word *strelka* being used in conjunction with the word *belgilamoq* when it signifies “date”. We may posit that in ten of the aforementioned one hundred sentences, the word *strelka* signifies “date” and in five of these sentences, the word *belgilamoq* is also employed. The probability of the word *strelka* occurring in conjunction with the word *belgilamoq* when it is used to denote a date is represented by P(belgilamoq/strelka=uchrashuv). In the aforementioned example, the ratio of instances where the word *belgilamoq* is used in conjunction with the word *strelka* to denote a “date” is 5/10, which equates to a probability of 0.5.

$$X = \frac{0.5*0.1}{0.2}$$

X = 0,25. In this manner, the probability that the word *strelka* is associated with the word *belgilamo*q which signifies “a symbol that denotes a specific concept”, is determined. Subsequently, the value that is closer to 1 is selected as the tag. To illustrate, in the sentence *“Soat 5 ga strelka belgiladi”* (He set the date at 5 p.m.) the algorithm would tag the word *strelka* with the annotation date due to its conjunction with the word *belgilamoq*.

Calculates the probability of each slang term denoting disparate meanings by calculating the probabilities and analyzing the context to ascertain the meaning of the slang in the sentence and selects the most probable meaning based on the context.

V. DISCUSSION

The Bayesian algorithm is a powerful tool that employs probabilistic principles to facilitate rational decision-making and predict the probability of an event. In contrast to conventional statistical techniques, which rely exclusively on observed data, Bayesian inference integrates prior knowledge and theoretical insights into the analysis. The algorithm commences with the determination of an initial conclusion regarding the probabilities of disparate outcomes. As further sources are incorporated into the newly constructed database, the algorithm updates the distribution in accordance with Bayes' theorem, which computes a posterior probability distribution by combining the prior distribution with the probability of the observed data given the hypothesis.

While the Bayesian algorithm offers a number of advantages, it also presents certain disadvantages that arise from inherent difficulties. One of the principal difficulties is the complexity of calculating probabilities for high-dimensional data sets. Notwithstanding these challenges, the Bayesian algorithm continues to represent a valuable tool for decision-making in the context of uncertainty in the field of computer networks. The algorithm's capacity to accommodate incomplete data, delineate intricate relationships, and offer intelligible outcomes makes it a valuable tool in numerous domains.

This algorithm represents one of the most efficient methods currently in use for the purpose of text tagging. The algorithm is advantageous due to its straightforward and readily comprehensible nature, rapid operational speed, and adaptability to disparate databases. The simplicity, speed and flexibility of this algorithm allow it to be used in a variety of fields. Nevertheless, the Bayesian algorithm is dependent on the data it is analyzing. It is therefore essential to ensure the accuracy of the data prior to applying the Bayesian algorithm. In the event of an error in the base data, the algorithm may yield inaccurate results. Furthermore, the Bayesian algorithm may also be unable to accurately determine the result when calculating the probability of words that are infrequent within the data set.

The Bayesian algorithm is founded upon the principle of Bayesian inference, which enables the calculation of suitable values for potential hypotheses as new evidence or data becomes available.

During the research the data was collected from oral speech materials. Participants were informed about the purpose of the data collection and their consent was obtained in writing. All identifying information was removed from the data to ensure the privacy and confidentiality of the participants.

Quantitative data comparing the effectiveness of the Bayesian approach to other methods have been included. We present graphs and tables that illustrate the accuracy and speed of various algorithms, offering a visual representation of the advantages and disadvantages of each method:

TABLE I

Algorithm	**Accuracy(%)**	**Time (seconds)**	**Best use cases**
Bayes	85	0.5	Tagging, text classification, Spam detection
SVM	90	1.2	Image recognition, Text classification

Neural Network	92	2.5	Complex pattern recognition, Language modeling

V. CONCLUSION

In conclusion, the Bayesian algorithm, particularly in its Naive Bayes classifier form, is a crucial component in the execution of a multitude of tasks, including text tagging, text classification in natural language processing (NLP), sentiment analysis, document categorization, language identification, and named object recognition. The algorithm's capacity to efficiently manipulate text data and make probabilistic predictions renders it a valuable tool in the fields of natural language processing and understanding. The Bayesian algorithm is a relatively simple yet highly efficient method that requires a relatively modest database to yield accurate results.

In general, the application of machine learning techniques to the automatic tagging of slang enables the attainment of accurate results in data analyzes The necessity of distinguishing and automatically categorizing slang units within a language corpus is that, when analyzing text in which such units are present, slang is treated as a conventional lexical item and translated directly during the process of automatic translation. This approach can prevent errors such as translation into a specified language. The Bayesian algorithm is a relatively simple approach. It was implemented using insert specific programming language or software. The algorithm was run for insert number iterations. Convergence was determined by observing the stability of the model parameters and the log-likelihood reaching a maximum after insert number iterations. No further significant changes in the model parameters or log-likelihood were observed after this point.

REFERENCES:

[1] An introduction to Naive Bayes Algorithm for beginners . (n.d.). Retrieved from turing.com

[2] D. Santos, B. Guimarães, B. “Boosting Named Entity Recognition with LSTMs” Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL), 2015.

[3] E. Brill. (1992). A Simple Rule-Based Part of Speech Tagger. Proceedings of the Third Conference on Applied Natural Language Processing (ANLP), 1992.

[4] J. Lafferty, A. McCallum, F. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. ICML, 2001.

[5] J. Lee, W. Yoon, S. Kim and others. A Fine-Grained Benchmark and Analysis of the Out-of-Distribution Generalization of Transformer Models. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2021.

[6] M. Abjalova (2023). “The issue of tagging and annotating corpus units.” Proceedings of the republican conference “Modern Science and Education Perspectives”, Tashkent, p. 345-351, October 2023.

[7] M. Collins. Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms. Proceedings of the ACL, 2002.

[8] S. Hochreiter, J. Schmidhuber. Long Short-Term Memory. Neural Computation, 9(8), p.1735-1780, 1997.

[9] O.M. Demskaya-Kulchitskaya V.R. Semerenko R.A. Yushchenko “Methods for automatic marking of texts of the national language corpus” Computer mathematics p. 70-75, 2005.

[10] Turing. Palo Alto, CA. Accelerate your AGI deployment. [Online]. Available: http://www.turing.com