УДК - 004.8

Қ.С. Оразбек*, Б.З. Ернияз
магистр, КазНИТУ им.К. Сатпаева, Алматы, Казахстан
магистр, КазНИТУ им.К. Сатпаева, Алматы, Казахстан
***Автор для корреспонденции:** kadyrzhan111@gmail.com

МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ ДЛЯ АНАЛИЗА ТЕКСТОВ НА КАЗАХСКОМ ЯЗЫКЕ

Аннотация

В статье представлен обзор применения методов машинного обучения для анализа текстов на казахском языке. Рассматриваемые методы включают автоматическое исправление орфографии, анализ тональности текстов, машинный перевод и классификацию текстов. Особое внимание уделяется адаптации алгоритмов к специфическим лингвистическим особенностям казахского языка. Обсуждаются перспективы развития специализированных методов извлечения признаков, необходимых для повышения точности и производительности моделей. Одним из перспективных направлений является использование трансформеров для анализа текстов. Эти модели, благодаря своему механизму внимания, способны выделять ключевые элементы текста, что особенно важно для агглютинативных языков, таких как казахский. Наивный байесовский классификатор – это вероятностный метод, основанный на теореме Байеса. Он предполагает независимость признаков и вычисляет вероятность того, что текст принадлежит определённой категории. Его преимущество — простота и высокая скорость работы, однако он может страдать от недостаточной точности при сложных зависимостях между словами. При этом важно учитывать сложные морфологические структуры слов. Например, нейронные сети на основе архитектуры LSTM могут успешно выявлять скрытые эмоции даже в сложных предложениях.

Ключевые слова: машинное обучение, анализ текста, обработка естественного языка, алгоритмы машинного обучения, методы классификации, казахский язык, машинный перевод.

Введение

В эпоху цифровизации текстовая информация стала основным источником данных для анализа, что требует эффективных методов обработки естественного языка (NLP). Лингвистические особенности казахского языка, такие как агглютинация, гармония гласных и уникальная структура предложений, представляют собой вызов для стандартных моделей машинного обучения. Тем не менее, развитие технологий открывает новые возможности для анализа текстов, включая автоматическое исправление ошибок, определение тональности и разработку систем перевода [1]. Настоящая статья посвящена изучению возможностей применения машинного обучения для решения этих задач.

Современные методы обработки текстов базируются на использовании больших языковых моделей, таких как GPT, BERT и их модификации. Эти модели обучаются на больших корпусах данных, что позволяет им учитывать сложные языковые закономерности. Однако для казахского языка создание подобных корпусов требует значительных усилий, так как доступных текстов недостаточно [2]. Решением может стать создание национальных текстовых баз данных, которые будут учитывать региональные и культурные особенности.

Одним из перспективных направлений является использование трансформеров для анализа текстов. Эти модели, благодаря своему механизму внимания, способны выделять ключевые элементы текста, что особенно важно для агглютинативных языков, таких как казахский [3].

Методы и материалы

Раздел обзора литературы в этой статье содержит всесторонний обзор существующих

исследований методов машинного обучения для анализа текстов на казахском языке. В литературе были предложены различные алгоритмы и методы классификации.

Автоматическое исправление орфографии

Для казахского языка орфографические ошибки часто связаны с неправильным использованием суффиксов или изменением порядка букв. Применение моделей на основе глубоких нейронных сетей, таких как трансформеры, позволяет эффективно исправлять ошибки, учитывая контекст. Примеры таких подходов включают использование моделей BERT и GPT, адаптированных к казахскому языку [4]. Кроме того, можно внедрить гибридные системы, которые сочетают правила и машинное обучение для обеспечения более высокой точности.

Анализ тональности текста

Анализ тональности предполагает классификацию текстов как позитивных, негативных или нейтральных. Алгоритмы, такие как Naïve Bayes, SVM и рекуррентные нейронные сети (RNN), могут быть адаптированы для обработки казахских текстов.

Наивный байесовский классификатор (Naïve Bayes) – это вероятностный метод, основанный на теореме Байеса. Он предполагает независимость признаков (слов) и вычисляет вероятность того, что текст принадлежит определённой категории. Его преимущество — простота и высокая скорость работы, однако он может страдать от недостаточной точности при сложных зависимостях между словами [5].

Метод опорных векторов (SVM) строит гиперплоскость, разделяющую данные на классы (например, позитивные и негативные тексты). Для казахского языка его эффективность может быть улучшена за счёт использования специализированных функций ядра, учитывающих морфологические особенности [6].

Рекуррентные нейронные сети (RNN) учитывают последовательность слов в тексте, что особенно важно для анализа тональности. Архитектуры, такие как LSTM или GRU, хорошо справляются с задачей, учитывая контекст даже на больших расстояниях [7].

При этом важно учитывать сложные морфологические структуры слов. Например, нейронные сети на основе архитектуры LSTM могут успешно выявлять скрытые эмоции даже в сложных предложениях.

Машинный перевод

Машинный перевод казахского языка представляет значительные трудности из-за его уникальной грамматики. Нейронные сети с механизмом внимания (attention) и модели, такие как Seq2Seq, уже продемонстрировали успехи в переводе между казахским и другими языками. Тем не менее, требуется дополнительное обучение моделей с использованием специализированных корпусных данных, чтобы повысить качество перевода. Также возможно использование многомодальных моделей, которые объединяют текстовую и визуальную информацию для улучшения понимания контекста.

Классификация текстов

Классификация текстов на казахском языке, включая определение категорий новостей или тем обсуждений, может быть выполнена с использованием таких алгоритмов, как логистическая регрессия, KNN и глубокие сверточные сети (CNN).

Логистическая регрессия - алгоритм прост в реализации и интерпретации. Он вычисляет вероятность принадлежности текста к определённому классу на основе линейной комбинации признаков. Для казахского языка логистическая регрессия может быть дополнена специфическими признаками, связанными с морфологией [8].

Метод k-ближайших соседей (KNN) - этот метод классифицирует текст на основе сходства с другими текстами в обучающей выборке. Его основное преимущество — интуитивность, однако при больших объёмах данных он становится вычислительно затратным [9].

Сверточные нейронные сети (CNN) эффективны для анализа текстов благодаря способности выявлять ключевые фразы. Они особенно полезны для задач, где важен

локальный контекст [10].

Особое внимание необходимо уделить предварительной обработке данных, включая токенизацию и стемминг, адаптированные к казахскому языку. Для достижения высоких результатов требуется учитывать частотность употребления слов и синтаксические связи между ними.

Результаты

Различные алгоритмы машинного обучения показали свою эффективность при решении задач анализа текстов. Например, использование трансформеров для исправления орфографии улучшает точность (accuracy) на 15–20% по сравнению с традиционными методами, что означает увеличение доли правильно исправленных слов. Анализ тональности текстов демонстрирует точность классификации до 90% при достаточной обучающей выборке, то есть алгоритмы корректно определяют тональность в 9 из 10 случаев. Машинный перевод, несмотря на сохраняющиеся трудности, демонстрирует значительный прогресс благодаря использованию Bilingual Evaluation Understudy (BLEU) метрики, которая отражает качество перевода на основе схожести с эталонным текстом. Классификация текстов с применением сверточных сетей позволяет достичь точности около 85%, что открывает перспективы для разработки специализированных приложений.

Использование машинного обучения для анализа текстов на казахском языке имеет широкий спектр применения. Это включает автоматизацию работы с документами, улучшение пользовательского опыта в поисковых системах, создание интеллектуальных помощников и разработку образовательных платформ. Например, автоматическое исправление орфографии может быть интегрировано в текстовые редакторы, а анализ тональности может использоваться для оценки общественного мнения.

Кроме того, машинный перевод на казахский язык и обратно открывает возможности для международного сотрудничества, образовательных обменов и культурной интеграции. Применение моделей машинного обучения в журналистике и маркетинге позволяет ускорить обработку информации и повысить эффективность работы.

Выводы

Применение машинного обучения для анализа текстов на казахском языке открывает новые горизонты для исследований и разработок. Однако для достижения высокой производительности требуется дальнейшая адаптация моделей и развитие специализированных методов обработки данных. Внедрение таких технологий может значительно способствовать цифровизации казахского языка и созданию удобных инструментов для его использования. Более того, создание национальных программ поддержки разработок в области NLP может ускорить процесс интеграции казахского языка в глобальное цифровое пространство.

Список литературы

1. Radford A., Narasimhan K., Salimans T., Sutskever I. Improving Language Understanding by Generative Pre-training. OpenAI, 2018
2. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser Ł., Polosukhin I. Attention Is All You Need. Advances in Neural Information Processing Systems, 2017.
3. Devlin J., Chang M.W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT, 2019.
4. M. RAZA, N.D. Jayasinghe, and M. M. A. Muslam. A comprehensive review on email spam classification using machine learning algorithms. 2021 International Conference on Information Networking (ICOIN), pages 327–332, Jan. 2021. URL https://doi.org/10.1109/ICOIN50884.2021.9334020.

5. McCallum A., Nigam K. A comparison of event models for Naive Bayes text classification. AAAI Workshop on Learning for Text Categorization, 1998
6. Niken L.O., Eko H.R., Christy A.S., DRIM Setiadi. Comparison of Multinomial Naïve Bayes Classifier, Support Vector Machine, and Recurrent Neural Network to Classify Email Spams // In International Seminar on Application for Technology of Information and Communication (iSemantic), September 19-20, 2020.
7. Bahdanau D., Cho K., Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate. ICLR, 2015.
8. Cover T., Hart P. Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 1967
9. Kim Y. Convolutional Neural Networks for Sentence Classification. EMNLP, 2014
10. Jurafsky D., Martin J.H. Speech and Language Processing. Prentice Hall, 2009.

References

1. Radford A., Narasimhan K., Salimans T., Sutskever I. Improving Language Understanding by Generative Pre-training. OpenAI, 2018
2. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser Ł., Polosukhin I. Attention Is All You Need. Advances in Neural Information Processing Systems, 2017.
3. Devlin J., Chang M.W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT, 2019.
4. M. RAZA, N.D. Jayasinghe, and M. M. A. Muslam. A comprehensive review on email spam classification using machine learning algorithms. 2021 International Conference on Information Networking (ICOIN), pages 327–332, Jan. 2021. URL https://doi.org/10.1109/ICOIN50884.2021.9334020.
5. McCallum A., Nigam K. A comparison of event models for Naive Bayes text classification. AAAI Workshop on Learning for Text Categorization, 1998
6. Niken L.O., Eko H.R., Christy A.S., DRIM Setiadi. Comparison of Multinomial Naïve Bayes Classifier, Support Vector Machine, and Recurrent Neural Network to Classify Email Spams // In International Seminar on Application for Technology of Information and Communication (iSemantic), September 19-20, 2020.
7. Bahdanau D., Cho K., Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate. ICLR, 2015.
8. Cover T., Hart P. Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 1967
9. Kim Y. Convolutional Neural Networks for Sentence Classification. EMNLP, 2014
10. Jurafsky D., Martin J.H. Speech and Language Processing. Prentice Hall, 2009.

Қ. Оразбек*, Б.З. Ернияз
магистр, Қ. Сәтбаев атындағы ҚазҰТЗУ., Алматы, Қазақстан
магистр, Қ. Сәтбаев атындағы ҚазҰТЗУ, Алматы, Қазақстан
***Корреспондент авторы:** kadyrzhan111@gmail.com

ҚАЗАҚ ТІЛІНДЕГІ МӘТІНДЕРДІ ТАЛДАУҒА АРНАЛҒАН МАШИНАЛЫҚ ОҚЫТУ ӘДІСТЕРІ

Түйін

Мақалада қазақ тіліндегі мәтіндерді талдау үшін машиналық оқыту әдістерін қолдануға шолу берілген. Қарастырылып отырған әдістерге емлені автоматты түрде түзету, мәтіннің кілтін талдау, машиналық аударма және мәтіндерді жіктеу кіреді. Алгоритмдерді қазақ тілінің ерекше лингвистикалық ерекшеліктеріне бейімдеуге ерекше көңіл бөлінеді. Модельдердің дәлдігі мен өнімділігін арттыру үшін қажетті белгілерді алудың мамандандырылған әдістерін дамыту

перспективалары талқыланады. Перспективалы бағыттардың бірі-мәтіндерді талдау үшін трансформаторларды қолдану. Бұл модельдер назар аудару механизмінің арқасында мәтіннің негізгі элементтерін ажырата алады, бұл әсіресе қазақ сияқты агглютинативті тілдер үшін өте маңызды. Аңғал Байес классификаторы-Байес теоремасына негізделген ықтималдық әдісі. Ол белгілердің тәуелсіздігін болжайды және мәтіннің белгілі бір санатқа жату ықтималдығын есептейді. Оның артықшылығы-жұмыстың қарапайымдылығы мен жоғары жылдамдығы, бірақ сөздер арасындағы күрделі тәуелділіктерде дәлдіктің жеткіліксіздігінен зардап шегуі мүмкін. Сөздердің күрделі морфологиялық құрылымдарын ескеру маңызды. Мысалы, lstm архитектурасына негізделген нейрондық желілер күрделі сөйлемдерде де жасырын эмоцияларды сәтті анықтай алады.

Кілттік сөздер: машиналық оқыту, мәтінді талдау, табиғи тілді өңдеу, машиналық оқыту алгоритмдері, жіктеу әдістері, қазақ тілі, машиналық аударма.

K.S. Orazbek*, B.Z. Yeryaz
Master's degree, KazNTU named after K. Satpayev, Almaty, Kazakhstan
Master's degree, KazNTU named after K. Satpayev, Almaty, Kazakhstan
***Corresponding author's email:** kadyrzhan111@gmail.com

MACHINE LEARNING METHODS FOR ANALYZING TEXTS IN KAZAKH

Abstract

The article provides an overview of the application of machine learning methods for analyzing texts in the Kazakh language. The methods considered include automatic spelling correction, text tonality analysis, machine translation, and text classification. Special attention is paid to the adaptation of algorithms to the specific linguistic features of the Kazakh language. The prospects for the development of specialized feature extraction methods necessary to improve the accuracy and performance of models are discussed. One of the promising directions is the use of transformers for text analysis. These models, due to their attention mechanism, are able to identify key elements of the text, which is especially important for agglutinative languages such as Kazakh. The naive Bayes classifier is a probabilistic method based on Bayes' theorem. It assumes the independence of features and calculates the probability that the text belongs to a certain category. Its advantage is simplicity and high speed of operation, however, it may suffer from insufficient accuracy with complex dependencies between words. It is important to take into account the complex morphological structures of words. For example, neural networks based on the LSTM architecture can successfully identify hidden emotions even in complex sentences.

Keywords: machine learning, text analysis, natural language processing, machine learning algorithms, classification methods, Kazakh language, machine translation.