АНАЛИЗ МИРОВОГО ОПЫТА (СУЩЕСТВУЮЩИХ МОДЕЛЕЙ) МОНИТОРИНГА ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ВЕБ-ПРОСТРАНСТВА (ТЕКСТОВОГО КОНТЕНТА) С ТОЧКИ ЗРЕНИЯ РЕАЛИЙ И ИНТЕРНЕТ-СРЕДЫ БЕЛАРУСИ

تحليل التجربة العالمية (النماذج الحالية) لرصد التحليل الذكي للويب (محتوى النص) من وجهة نظر الواقع وبيئة الإنترنت في بيلاروسيا

Атрошкина А. Д.
أناستاسيا دميتريفنا أتروشكينا
Магистрант БГЭУ
طالبة ماجستير في جامعة بيلاروسيا الحكومية للاقتصاد

Романовская П. Д.
بولينا دميتريفنا رومانوفسكايا
Специалист ОАО «Гипросвязь»,
اخصائية في مصنع جيبروسفياز بمينسك

Юневич Н. Г.
نيكول جورجيفنا يونيفيتش
Аспирант БНТУ
طالبة دكتوراه في جامعة بيلاروسيا الوطنية التقنية

Аннотация: в этом работе анализируется возможность применения существующих моделей мониторинга интеллектуального анализа веб-пространства в Республике Беларусь. В качестве примера рассматривается идентификация деструктивного поведения в социальных сетях, с помощью применения технологии больших данных и машинного обучения.

Ключевые слова анализ текстового контента, безопасное веб-пространство, модели мониторинга, инструменты автоматизированного мониторинга.

الملخص : في هذا العمل يتم تحليل إمكانية استخدام النماذج الحالية لرصد التعدين على شبكة الإنترنت في جمهورية بيلاروسيا فعلى سبيل المثال يتم تحديد السلوك المدمر على الشبكات الاجتماعية باستخدام تكنولوجيا البيانات الضخمة والتعلم الآلي.

الكلمات المفتاحية : تحليل محتوى النص، مساحة ويب آمنة، نماذج المراقبة، أدوات المراقبة الآلية

Введение

Данное исследование проводится в рамках белорусско-индийского проекта «Система мониторинга и интеллектуального анализа веб-пространства "Безопасное веб-пространство"».

Исследования и разработка

Гипотеза проекта: исследование существующих моделей мониторинга и интеллектуального анализа текстового контента и разработка методов (алгоритмов) сбора, хранения, обработки и лингвистического анализа текста, что ранее подробно не рассматривалось в отношении русского языка и хинди в корреляции с негативным контентом для уязвимых групп (женщины, дети, молодежь). Актуальность проекта дополнительно подтверждается вовлеченностью в интернет-пространство всех возрастных групп граждан Республики Беларусь [1].

Особенностями интернет-среды Республики Беларусь в отношении к текстовому контенту выступает: многоязычность и сленг, эмоциональность (эмодзи, gif), хештеги и метаинформация, ирония и сарказм, аббревиатуры и акронимы, гипертекстовая структура, коллективные языковые нормы, краткость и низкое качество текста и т. п. [2]. Данные особенности зависят от возраста и социальных групп, в рамках которых осуществляется формирование контента. Они напрямую влияют на процесс обучения моделей и требуют более детального выбора датасетов.

В рамках выбора модели для реализации проекта проведен анализ механизмов и инструментов автоматизированного мониторинга веб-пространства (текстового контента). Изучение моделей включало в себя анализ механизмов обработки естественного языка (NLP) в части возможности распознавания текста, синтеза речи, морфологического анализа и канонизации, синтаксического разбора и токенизации предложений, извлечения отношений, определения языка, анализа эмоциональной окраски и т.д. Дополнительным критерием выступала возможность обучения моделей на русском языке или их изначальная адаптация к языку. Предварительно рассмотрены стадии анализа текстового контента, включая маркировку, предварительную обработку, удаление шума, нормализацию, токенизацию, поиск стоп-слов, стеминг, извлечение функций и построение моделей.

Актуальные исследования демонстрируют, что в рамках классификации и анализа текстового контента (естественного языка) целесообразно использовать методы глубокого обучения (LSTM, GRU, CNN, BERT, USE и др.). При этом более эффективны гибридные Q.модели (BiLSTM+CNN+Self-Attention, SVM+Naïve Bayes+Logistic, др. комбинации) [1, 3]. В дополнении, построение моделей осуществляется с использованием различных инструментов в зависимости от стадии анализа и конкретных задач (для классификации – LSTM, SVM, NB , LR, KNN, BERT, для извлечения признаков – Word2vec, Bag-of-words и TF-IDF, для интерпретации решений, принятых другими моделями, и для обнаружения влияния конкретных слов или фраз могут быть полезны модели LIME, для маркировки – LSTM, GRU, CNN и другие глубокие нейронные сети и т.п.) [1, 3].

В качестве метрик для оценки производительности моделей рассмотрены BLEU (двуязычная оценка) или GLUE (девять различных наборов задач), а также F-оценка, Matthews correlation coefficient (MCC) [1, 3].

В ходе исследования наибольшую эффективность продемонстрировала модель для анализа русскоязычного текста RuBERT (RuBERT-NLI, RuBERT-tiny и др.), которая использовалась для схожих исследований (русскоязычный контент) и проявила наибольшую эффективность.

Заключение

В ходе исследования определены основные инструменты, которые способны решить вопрос анализа негативных комментариев в социальных сетях от пользователей Республики Беларусь. Следующий шагом является сбор данных для формирования датасета. Тестовый датасет будет использован для формирования классификатора текстового контента, а также последующего обучения модели.

المقدمة

يتم إجراء هذا البحث في إطار المشروع البيلاروسي الهندي "نظام المراقبة والتحليل الفكري لمساحة الويب "مساحة الويب الآمنة"".

النتائج والمناقشة

فرضية المشروع: البحث في النماذج الحالية للرصد والتحليل الفكري لمحتوى النص وتطوير أساليب (خوارزميات) لجمع وتخزين ومعالجة وتحليل النص اللغوي، والتي لم يتم فحصها بالتفصيل من قبل فيما يتعلق باللغتين الروسية والهندية وارتباطه بالمحتوى السلبي للفئات الضعيفة (النساء، الأطفال، الشباب). يتم تأكيد أهمية المشروع بشكل أكبر من خلال إشراك جميع الفئات العمرية لمواطني جمهورية بيلاروسيا في فضاء الإنترنت [1].

خصوصيات بيئة الإنترنت في جمهورية بيلاروسيا فيما يتعلق بمحتوى النص هي: التعددية اللغوية والعامية، والعاطفية (الرموز التعبيرية، GIF)، وعلامات التصنيف والمعلومات الوصفية، والسخرية والسخرية، والاختصارات والمختصرات، وبنية النص التشعبي، ومعايير اللغة الجماعية، الإيجاز وانخفاض جودة النص، الخ. البند [2]. تعتمد هذه الميزات على العمر والفئات الاجتماعية التي يتم إنشاء المحتوى ضمنها. إنها تؤثر بشكل مباشر على عملية التدريب النموذجي وتتطلب مجموعة أكثر تفصيلاً من مجموعات البيانات.

كجزء من اختيار نموذج لتنفيذ المشروع، تم إجراء تحليل لآليات وأدوات المراقبة الآلية لمساحة الويب (محتوى النص). وتضمنت دراسة النماذج تحليل آليات معالجة اللغة الطبيعية (NLP) من حيث إمكانية التعرف على النص، وتركيب الكلام، والتحليل الصرفي والتقنين، وإعراب الجمل وترميزها، واستخلاص العلاقات، والكشف عن اللغة، وتحليل التلوين العاطفي، إلخ. كان المعيار الإضافي هو إمكانية تدريب النماذج باللغة الروسية أو تكييفها الأولي مع اللغة. تتم تغطية مراحل تحليل محتوى النص، بما في ذلك وضع العلامات، والمعالجة المسبقة، وتقليل الضوضاء، والتطبيع، والترميز، وإيقاف البحث عن الكلمات، والأصل، واستخراج الميزات، وبناء النماذج بشكل مبدئي.

توضح الأبحاث الحالية أنه من المستحسن استخدام أساليب التعلم العميق (LSTM، GRU، CNN، BERT، USE، وما إلى ذلك) في إطار تصنيف وتحليل محتوى النص (اللغة الطبيعية). في الوقت نفسه، تعد نماذج Q الهجينة (BiLSTM+CNN+Self-Attention، و SVM+Naïve Bayes+Logistic، ومجموعات أخرى) أكثر فعالية [1، 3]. بالإضافة إلى ذلك، يتم بناء النماذج باستخدام أدوات مختلفة اعتمادًا على مرحلة التحليل والمهام المحددة (للتصنيف - LSTM، SVM، NB، LR، KNN، BERT، لاستخراج الميزات - Word2vec، Bag-of-words و يمكن أن تكون نماذج TF-IDF وLIME مفيدة لتفسير القرارات التي تتخذها النماذج الأخرى ولاكتشاف تأثير كلمات أو عبارات معينة؛ LSTM، GRU، CNN وغيرها من الشبكات العصبية العميقة، وما إلى ذلك يمكن أن تكون مفيدة في وضع العلامات) [1، 3].

تعتبر BLEU (تقييم ثنائي اللغة) أو GLUE (تسع مجموعات مهام مختلفة)، بالإضافة إلى درجة F ومعامل ارتباط ماثيوز1] (MCC) ، [3، بمثابة مقاييس لتقييم أداء النماذج.

أثناء الدراسة، أظهر نموذج تحليل النص باللغة الروسية RuBERT (RuBERT-NLI، RuBERT-tiny، وما إلى ذلك)، والذي تم استخدامه لدراسات مماثلة (محتوى اللغة الروسية) وأظهر أكبر قدر من الكفاءة، أكبر قدر من الكفاءة.

الخاتمة

حددت الدراسة أهم الأدوات التي يمكنها حل مشكلة تحليل التعليقات السلبية على شبكات التواصل الاجتماعي من مستخدمي جمهورية بيلاروسيا. الخطوة التالية هي جمع البيانات لتشكيل مجموعة بيانات. سيتم استخدام مجموعة بيانات الاختبار لتشكيل مصنف محتوى النص، بالإضافة إلى التدريب اللاحق للنموذج.

Литература

1. Digital 2023: Belarus [Electronic resource] // DataReportal – Global Digital Insights. URL: https://datareportal.com/reports/digital-2023-belarus (дата обращения: 20.02.2023).
2. Бутуханова З.А. Влияние «Русинглиша» на языковую коммуникацию и культуру / З.А. Бутуханова, Г.А. Желаев, Н.А. Янькова // Вестник науки. – 2023. – №12 (69). – С. 795–801.
3. Musleh, D.A. Arabic Sentiment Analysis of YouTube Comments: NLP-Based Machine Learning Approaches for Content Evaluation / D.A. Musleh, I. Alkhwaja, A. Alkhwaja, M. Alghamdi, H. Abahussain, F. Alfawaz, N. Min-Allah, M.M. Abdulqader // Big Data Cogn. Comput. – 2023 – №7 – C.127.