Технически Университет Варна

МЕСЕЦ НА НАУКАТА 2023

ФАКУЛТЕТ ПО ИЗЧИСЛИТЕЛНА ТЕХНИКА И АВТОМАТИЗАЦИЯ

КЛАСИФИКАЦИЯ НА ТЕКСТ НА БЪЛГАРСКИ ЕЗИК ЧРЕЗ МЕТОДИ НА МАШИННОТО ОБУЧЕНИЕ

Ръководител на проекта:
доц. д-р инж. Нели Калчева, СИТ

Участници:
доц. д-р инж. Виолета Божикова, СИТ
гл. ас. д-р Мая Тодорова, СИТ
гл. ас. д-р инж. Гинка Маринова, КНТ
ас. Даниела Петрова – докторант, СИТ
инж. Димитър Кръстев, СИТ
инж. Йордан Станчев, СИТ
Викторио Николаев – студент, спец. СИТ
Иван Яръмов – студент, спец. СИТ
Мустафа Мустафов – студент, спец. СИТ
Никола Лазаров – студент, спец. СИТ
Димитър Линов – студент, спец. СИТ
Алекс Орозов – студент, спец. СИТ
Георги Соколов – студент, спец. КСТ

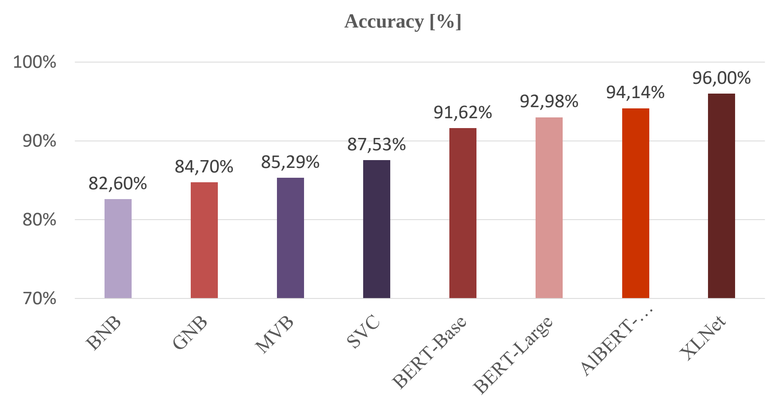
Въведение

Класификацията на текст на български език е сложна задача, поради специфичните езикови характеристики на българския език, предопределящи трудностите при компютърната обработка.

Основна научно-изследователска цел на проекта е класификация на текстове на български език, чрез методи, алгоритми и съвременни модели на машинното обучение.

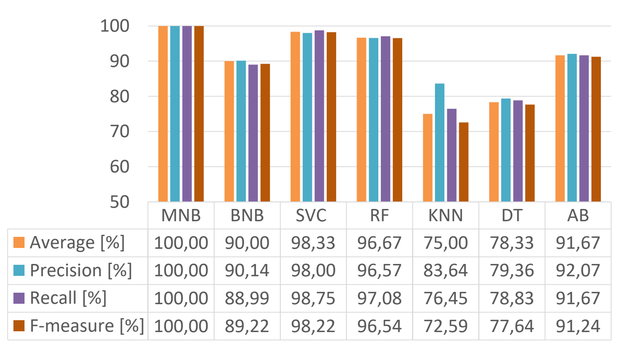
Резултати

Реализирано е сравняване на точността на моделите Bidirectional Encoder Representations from Transformers (BERT), Generalized Autoregressive Pretraining for Language Understanding (XLNet) при класификацията на текст с точността на класическите методи и алгоритми за машинно обучение: Bernoulli Naive Bayes classifier (BNB), Gaussian Naive Bayes classifier (GNB), Multinomial Naive Bayes classifier (MNB), Support Vector Machines (SVC). Резултатите показват, че при класифициране на 50 000 рецензии, XLNet класира с най-висока точност – 96%, което е с близо 8% повече от най-добре представящия се класически класификатор Support Vector Machines (Фиг. 1).



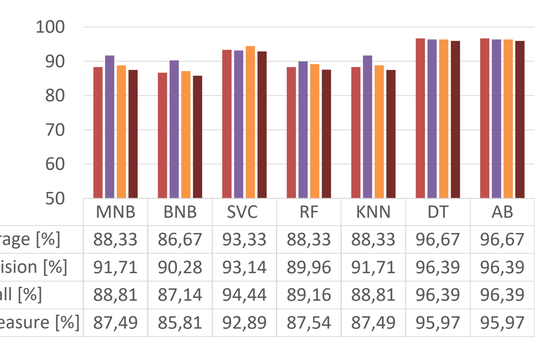
Фиг.1 Точност на алгоритми на машинното обучение при класификация на текст

Генерирани са модели за класификация и е осъществен сравнителен анализ на точността, прецизността, чувствителността и F-мярка на следните алгоритми на машинното обучение: Multinomial Naive Bayes classifier (MNB), Bernoulli Naive Bayes classifier (BNB), Support Vector Machines (SVC), Random Forest (RF), AdaBoost (AB) , Decision Tree (DT) и K-Nearest Neighbosr (KNN) при класификация на произведения на английски автори и класификация на авторите на същите произведения в превод на български език. Проведените изследвания показват, че при класификация на английски автори с равен брой произведения на английски език с най-високи стойности на изследваните показатели са Support Vector Machines и Multinomial Naive Bayes classifier, докато при текстове на български език най-добри резултати се получават в зависимост от конкретните автори (Фиг. 2 и Фиг. 3).



	MNB	BNB	SVC	RF	KNN	DT	AB
Average [%]	100,00	90,00	98,33	96,67	75,00	78,33	91,67
Precision [%]	100,00	90,14	98,00	96,57	83,64	79,36	92,07
Recall [%]	100,00	88,99	98,75	97,08	76,45	78,83	91,67
F-measure [%]	100,00	89,22	98,22	96,54	72,59	77,64	91,24

Фиг. 2 Класификация на произведения на английски автори



	MNB	BNB	SVC	RF	KNN	DT	AB
Average [%]	88,33	86,67	93,33	88,33	88,33	96,67	96,67
Precision [%]	91,71	90,28	93,14	89,96	91,71	96,39	96,39
Recall [%]	88,81	87,14	94,44	89,16	88,81	96,39	96,39
F-measure [%]	87,49	85,81	92,89	87,54	87,49	95,97	95,97

Фиг. 3 Класификация на произведения в превод на български език на английски автори

Създадени са собствени бази от данни с коментари от две различни сфери - от клиенти на хотели и на магазини, заведения, културни мероприятия, козметични и здравни центрове, както и са предложени подобрения на предварителната обработка на данните на текстове на български език. Осъществена е класификация на текст чрез прилагане на най-често използваните методи за машинно обучение, подходящи за анализ на настроенията, а именно Naïve Bayes, Support Machine Vectors, Logistic Regression. Резултатите от тези изследвания, се оказват противоречиви, тъй като се получава съществена разлика в точността на предсказванията за двете бази данни. Приложени са методите Bidirectional RNN и Random Forest, в търсене на по-високи крайни резултати и по-малка разлика между двете бази.

Заключение

Някои от постигнатите резултати са следните:

- Осъществен е сравнителен анализ на точността на различни методи и алгоритми за класификация на текстове на български език и на английски език.
- Осъществен е сравнителен анализ на точността, прецизността, чувствителността и F-мярка на 7 алгоритми на машинното обучение при класификация на автори на произведения на английски автори и класификация на авторите на същите произведения в превод на български език.
- Предложени са нови модели на преподаване на методи и алгоритми на машинното обучение за студентите от техническите университети.

Избрани публикации по проекта

1. Kalcheva N., Karova M., A Comparison of Machine Learning Classification Algorithms and Methods for English Author's Works and their Translations into Bulgarian, 57th International Scientific Conference on Information, Communication and Energy Systems and Technologies, Ohrid, North Macedonia, 16-18 June, 2022, pp. 1-4
2. Petrova, D., Bozhikova V., Random forest and recurrent neural network for sentiment analysis on texts in Bulgarian language, International Conference on Biomedical Innovations and Applications, Varna, Bulgaria, 2-4 June 2022, vol.1, pp.66-69
3. Kalcheva N., Kovachev I., Comparison of BERT and XLNet accuracy with classical methods and algorithms in text classification, International Conference on Biomedical Innovations and Applications, Varna, Bulgaria, 2-4 June 2022, vol.1, pp.74-76
4. Kalcheva N., Teaching of Bayes formula, 57th International Scientific Conference on Information, Communication and Energy Systems and Technologies, Ohrid, North Macedonia, 16-18 June, 2022, pp. 1-3