Дослідження методів інтелектуального аналізу текстових даних у режимі реального часу

Винонав: Спітковський В. І., ДА-62
Науковий керівник: Яременко В. С.

Об’єкт та предмет дослідження

- Об’єкт дослідження: методи інтелектуального аналізу тестових даних
- Предмет дослідження: методи машинного навчання для вирішення задач класифікації текстових даних у режимі реального часу

Постановка проблеми

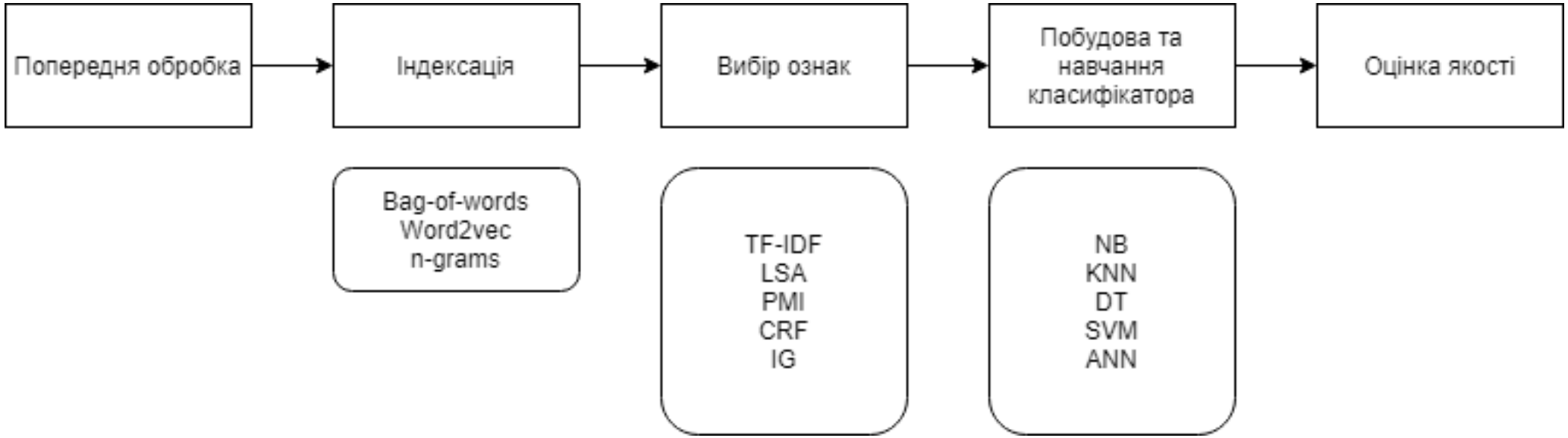
Для класифікації потокових даних в режимі реального часу ми повинні мати справу з трьома основними вимірами:

- Точність классифікованих даних.
- Час потрібний для класифікації певної порції даних.
- Об’єм даних.

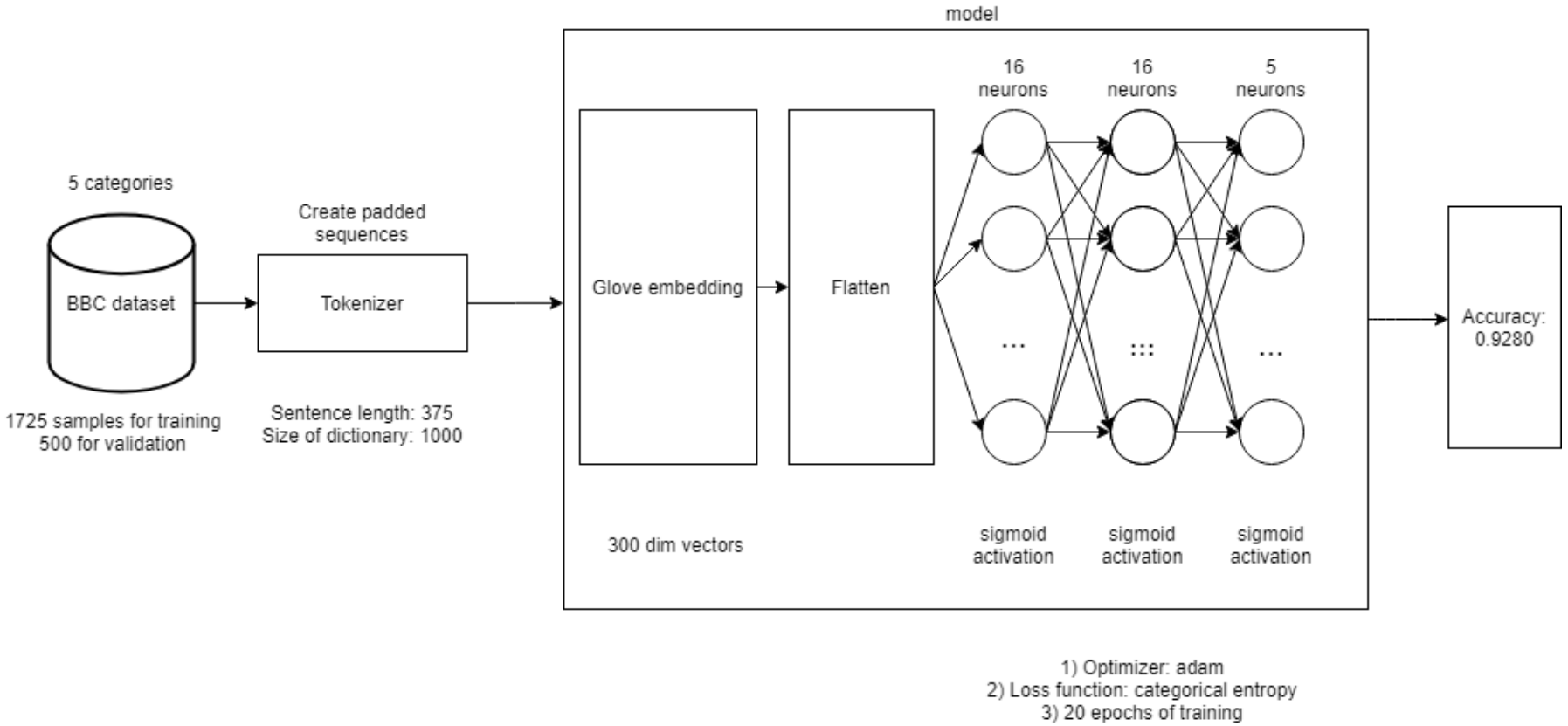
Етапи дослідження

Назва етапу	Результат
Аналіз статей та матеріалів що відносяться до предметної області.	Схема етапів процесу класифікації текстів.
Проектування, реалізація моделей декількох типів нейронних мереж та регрессивних класифікаторів.	Моделі класифікаторів на основі нейронних мереж або регресій (грунт для аналізу).
Аналіз отриманих класифікаторів з різними параметрами.	Порівняльна характеристика різних класифікаторів.

Схема етапів процессу классифікації текстів.



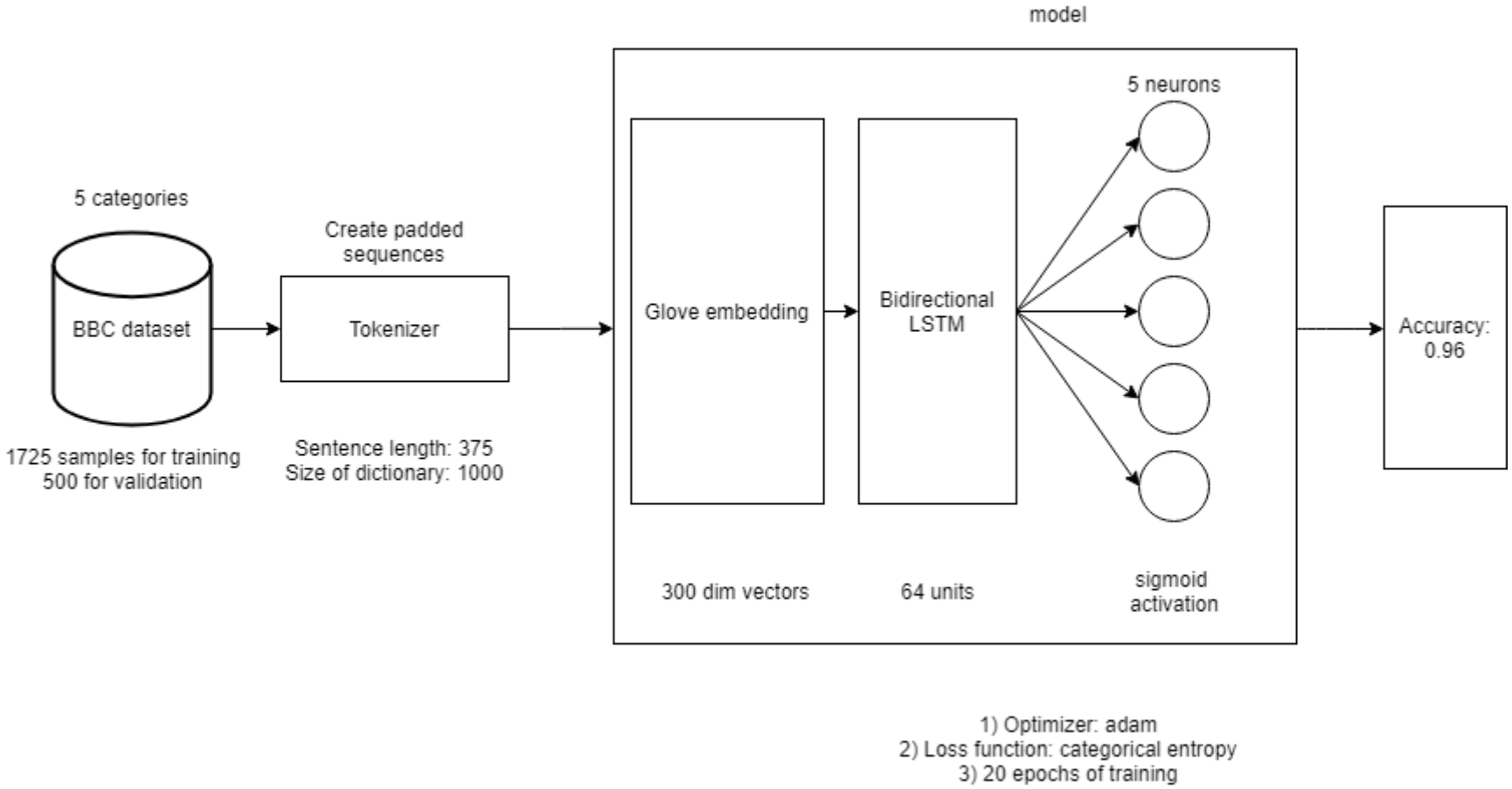
Модель роботи DNN (BBC)



Результати роботи DNN

activation	epochs	time per iter	optimizer	dense layers	neurons	val_accuracy
relu, relu, sigmoid	20	4s	adam	3	32, 16, 5	0.934
sigmoid x3	20	4s	adam	3	32, 16, 5	0.948
linear x3	20	4s	adam	3	32, 16, 5	0.252
sigmoid	20	3s	adam	1	5	0.942
sigmoid x2	20	4s	adam	2	32, 5	0.958
relu, sigmoid	20	4s	adam	2	32, 5	0.938
sigmoid x3	20	3s	adam	3	16, 16, 5	0.96
sigmoid x3	20	3s	adam	3	16, 10, 5	0.952
sigmoid x3	20	2s	sgd	3	16, 16, 5	0.238

Модель роботи RNN (BBC)



1) Optimizer: adam
2) Loss function: categorical entropy
3) 20 epochs of training

Результати роботи RNN

RNN (GRU)

activation	epochs	time per epoch	optimizer	GRU outputs	simple RNN	val_accuracy
relu	20	12s	adam	32	16	0.186
sigmoid	20	12s	adam	32	16	0.162
linear	20	12s	adam	32	16	0.142
relu	20	13s	adam	16	16	0.232
relu	10	8s	adam	32	16	0.192
relu	10	10s	adam	10	16	0.164
sigmoid	20	12s	adam	16	16	0.148

RNN (LSTM)

activation	epochs	optimizer	LSTM dim	time per epoch	val_accuracy
softmax	20	adam	300	120-160s	0.95
softmax	20	adam	64	40-50s	0.932
relu	20	adam	64	10s	0.22
sigmoid	20	adam	64	21s	0.96

Результати роботи CNN

activation	epochs	time per iter	optimizer	dense layers	conv layers	kernel_sz	batch_sz	val_accuracy
relu	20		adam	1	1	5	5	0.192
sigmoid	20		adam	1	1	5	5	0.19
linear	20		adam	1	1	5	5	0.182
relu	20	26s	adam	1	1	300	5	0.19
relu	20	12	adam	2	1	32	5	0.246
relu	20	13	adam	2	2	32	5	0.23

NB
val_accuracy
0.98

- Розмір статті: 10000 сим.
- К-сть статей: 7000

- Розмір статті: 375 сим.
- К-сть статей: 2225

Результати роботи DNN (Reuters)

DNN						
activation	epochs	time per iter	optimizer	dense layers	neurons	val_accuracy
sigmoid x3	20	2s	adam	3	64, 64, 46	0.7964
sigmoid x5	20	83	adam	5	64, 64, 64, 64 46	0.78

Результати роботи RNN, NB

RNN (LSTM)					
activation	epochs	time per iter	optimizer	LSTM dim	val_accuracy
sigmoid	3	3408s	adam	32	0.3315

NB
val_accuracy
0.64

Використані бібліотеки та набори даних

Висновки

Алгоритм	Точність	Швидкість	Об'єм
DNN	0.96	<1 ms	BBC
	0.79	<1 ms	Reuters
RNN	0.96	<1 ms	BBC
	0.3315	<1 ms	Reuters
CNN	0.246	<1 ms	BBC
	--	--	Reuters
NB	0.98	>2s	BBC
	0.64	>4s	Reuters

Дякую за увагу!