Проектът TRACES и неговият принос

26 януари 2023 г.

Огледална зала, Ректорат на СУ

Какво представлява проектът

Едногодишен проект

Финансиран от Европейската Комисия през AI4Media

Екип: организаторите на Семинара

Цел: за първи път в България да се изследва дали и как може да се създаде софтуер, който да разпознава:

1) Целенасочена дезинформация написана от хора
2) Текстови дийпфейкове

Приложение: в социалните медии, на български език

Как го постигаме:

- чрез психолигвистични характеристики на езика на съобщението,
- с помощта на журналисти
- с методи на изскуствения интелект


Приноси на проекта

Можем ли да разпознаваме лъжата на български език?

и как?

Можем ☺ Създали сме ресурси в начален вариант,

които могат да се използват

Какво сме създали и създаваме през този проект, което ще е от полза и на кого

- Списъци от езикови характеристики, издаващи лъжа и дезинформация, изрази на български език

- Текстове, които могат да бъдат използвани от други учени за анализ на този феномен и за създаване на софтуер


- Програмен код (Python)

- Прототип на софтуер, анализиращ текст и казващ с някаква вероятност дали текстът съдържа дезинформация и дали е автоматично генериран (дийпфейк)

Какво още сме открили?

Има ли следи на автоматично генерирани текстове в българските социални медии?

Има следи, но за момента ни трябват повече изследвания, за да кажем, дали тези следи наистина са на такива текстове

Как можем и можем ли да разпознавате автоматично генерирани съобщения?

Някои можем, за някои е много много трудно и за някои е невъзможно

Как можете да разпознавате автоматично генерирани съобщения (особено неверни)

Съвсем лошо написани автоматични съобщения

- Главни букви посрЕДАта на дУМата

- изречения, започващи с малка буква

- Внезапно прекъснати изречен

- Повтарящи се изрази и думи и думи

Една идея по-добре написани

- В текста започва да се говори за определен човек или място или време, и изведнъж вече се говори за друг човек, място или време, или са споменати несвързани събития:

Министърът на външните работи на Сърбия Ивица Дачич е на посещение в Брюксел. Ден по-късно Джайшанкар се срещна с други министри от БРИКС.

- Започва да се говори за определени цифри, и внезапно те се променят:

Българската AE Solar Horizon инвестира 20 милиона лева във фабрика край Кюстендил и разкрива над 100 работни места. Новите 500 работни места ще осигурят работа на висшисти в областта.

- Неправилно използвани местоимения:

Министърът на външните работи на Сърбия Ивица Дачич е на посещение в Брюксел. Днес тя разговаря с ръководителя на европейската дипломация Жозеп Борел.

Още по-добре написани

- В текста се описва повтаряща се сцена:

Министърът на външните работи на Сърбия Ивица Дачич е на посещение в Брюксел. Днес той разговаря с ръководителя на европейската дипломация Жозеп Борел. По-рано днес Дачич разговаря и с депутати от Европейския парламент. Дачич е заявил, че членството в ЕС остава стратегическа цел на Сърбия. Дачич разговаря в Брюксел с ръководителя на европейската дипломация Борел.

- Текстът съдържа противоречия:

Министърът на външните работи на Сърбия Ивица Дачич е на посещение в Брюксел. Дачич е заявил, че членството в ЕС остава стратегическа цел на Сърбия. Дачич отказа да разговаря в Брюксел с ръководителя на европейската дипломация Борел.

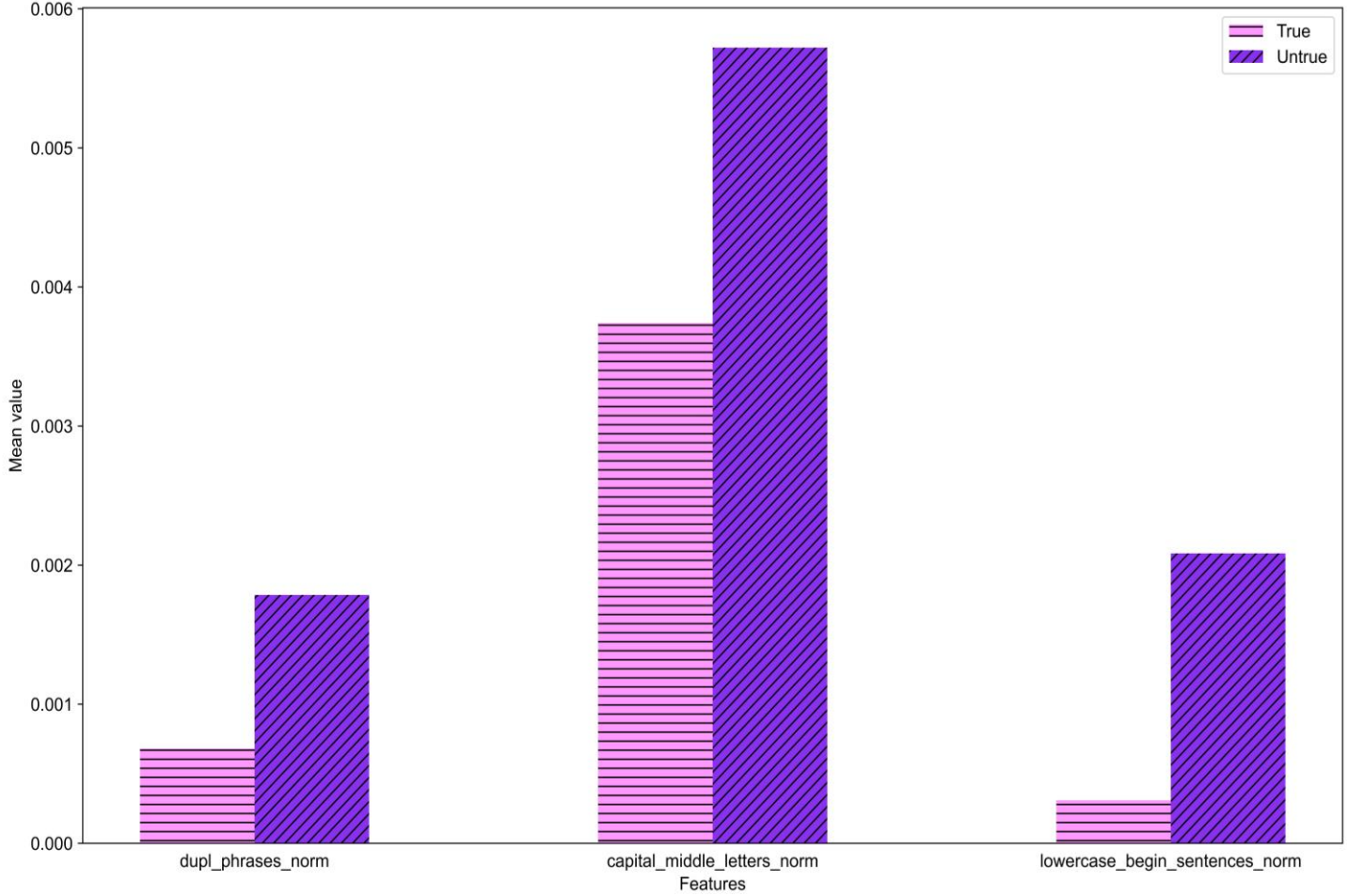
Съвсем добре написани
(изискват се усилия и замисляне)

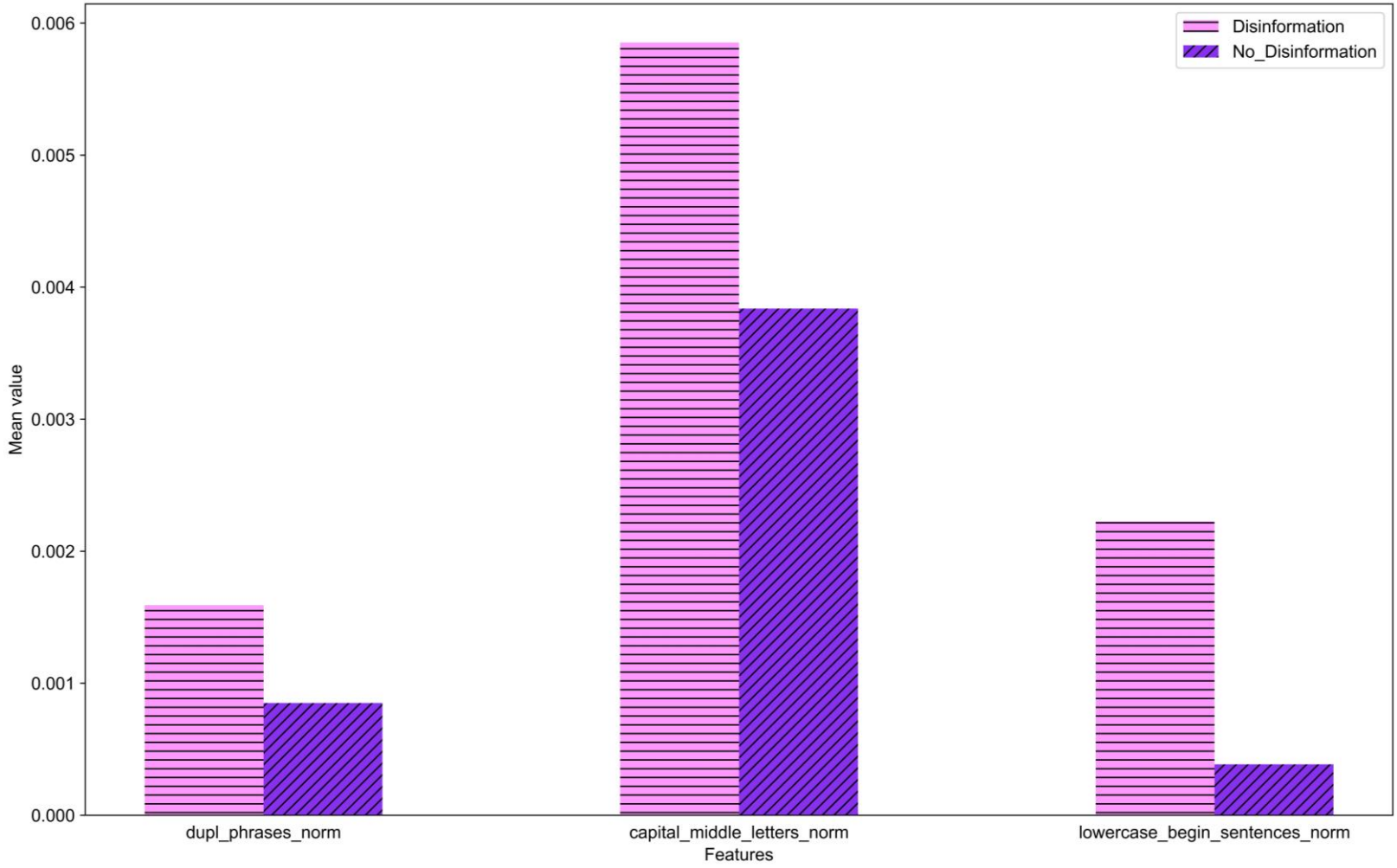
- Текстът е добре написан, няма никакви видими грешки, НО изказва нещо абсолютно нелогично, което никога не би било казано от човек

Обичам да прекарвам време вкъщи. Основно любимите ми занимания са да чета книги и да готвя вкуснотии. Обаче най-много ми харесва да гледам как работи пералнята.

Открития на проекта до този момент

- Съобщенията съдържащи главни букви посредата на изречението, изречения, започващи с малка буква, и повтарящи се изрази са двойно по-често отбелязвани като неверни и съдържащи дезинформация от журналистите, които работиха по проекта


0.006
0.005
0.004
0.003
0.002
0.001
0.000
Mean value
True
Untrue
dupl_phrases_norm
capital_middle_letters_norm
lowercase_begin_sentences_norm
Features


Disinformation
No_Disinformation
0.006
0.005
0.004
0.003
0.002
0.001
0.000
Mean value
dupl_phrases_norm
capital_middle_letters_norm
lowercase_begin_sentences_norm
Features

Примери на автоматично генерирани текстове в процес на генериране

Примери на автоматично генерирани съобщения от GPT

Здравейте Днес ми е последния ден от карантината след преболедуване от коронавирус. Това съобщиха от пресцентъра на Министерството на здравеопазването. През последното денонощие в страната са регистрирани повече от 100 случая на коронавиру

Здравейте драги сънародници.. Бях закупил тест за 2ри ден в България, който се оказа положителен за коронавирус. Това заяви премиерът Бойко Борисов по време на правителственото заседание, предаде репортер на Агенция „Фокус“. „Правителс

Племенника ми е от 12г.в Лондон има си ресторант пицария, където се събират повече хора, отколкото през последните години. Това е най-големият ресторант в Лондон, където се събират повече хора, отколкото през последните години. Това е ?

След пандемията, светът ще е различен, защото когато ти е трудно да разбереш какво се случва в света, тогава ще ти е много трудно да разбереш какво се случва в света и какво се случва в света и какво се случва в света и какво се слу?


Благодаря за вниманието ☺

Някои източници

1. Проблеми на автоматично генерираните текстове:

Dale, R. (2021). GPT-3: What's it good for?. Natural Language Engineering, 27(1), 113-118;
https://www.sigmoid.com/blogs/gpt-3-all-you-need-to-know-about-the-ai-language-model/

https://www.sigmoid.com/blogs/gpt-3-all-you-need-to-know-about-the-ai-language-model/

2. Примери на автоматично генерирани текстове:

https://www.newyorker.com/magazine/2019/10/14/can-a-machine-learn-to-write-for-the-new-yorker