



Универзитет „Св. Кирил и Методиј“ во Скопје
Факултет за информатички науки и компјутерско инженерство

ДИПЛОМСКА РАБОТА

Тема:

**Компаративна анализа на различни пристапи и модели
за препознавање на машински генериран текст**

Ментор:

Проф. д-р. Соња Гиевска

Студент:

Александар Стојанов - 211067

Скопје, 2025

Содржина:

1. Апстракт.....	3
2. Вовед.....	4
3. Сродни истражувања.....	5
3.1 StyloAI : Distinguishing AI-Generated Content with Stylometric Analysis.....	5
3.2 Linguistic Differences between AI and Human Comments in Weibo: Detect AI-Generated Text through Stylometric Features.....	9
3.3 M4: Multi-Generator, Multi-Domain, and Multi-Lingual Black-Box Machine-Generated Text Detection.....	12
4. Методи и истражувања.....	16
4.1 Вовед во податочното множество и стилометриска анализа.....	16
4.2 Анализа и визуелизација на податочното множество преку стилометриските карактеристики.....	19
4.3 Користење традиционални модели од машинско учење.....	23
4.3.1 Споредба со постоечкото истражување StyloAI.....	24
4.4 Користење на повеќеслојна невронска мрежа и self-attention слој – SFSC (Stylometric Feature-Based Self-Attention Classifier).....	24
4.4.1 Опис на пристапот и архитектура.....	24
4.4.2 Резултати од моделот со невронска мрежа и self-attention слој базиран на стилометриски карактеристики (SFSC).....	26
4.4.2.1 Анализа на важност на карактеристиките.....	27
4.4.2.3 Споредба со постоечки истражувања.....	28
4.5 Експерименти базирани на трансформер модели.....	29
4.5.1 Споредба на перформансите на lightweight трансформер модели.....	30
4.5.2 Fine-tuning и детална евалуација на финалниот модел.....	31
4.5.3 Споредба на финалниот модел со zero-shot детектори.....	32
4.5.4 Споредба со истражувањето M4.....	33
4.6 Споредба на различните пристапи во овој експеримент: традиционални ML, невронски мрежи и трансформер модели.....	33
5. Предизвици и ограничувања.....	35
6. Следни истражувања.....	35
7. Заклучок.....	36
8. Референции.....	37

1. Апстракт

Со развојот на современите јазични модели, машински генериралиот текст, во понатамошниот текст означен како МГТ, станува сè посличен на оној нашишен од човек, што претставува сериозен предизвик за препознавање на потеклото на текстот. Овој проблем е особено значаен во ситуации каде автентилноста како и доверливоста на еден текст е од големо значење, а тоа е пишувањето текст на социјалните медиуми, во литературата, како и во образовните институции.

Во оваа дипломска работа е направена компаративна анализа на различни пристапи за препознавање на машински генериран текст. Истражувањето опфаќа методи базирани на стилометриски карактеристики и традиционални модели од машинско учење, невронска мрежа со self-attention слој врз стилометриски карактеристики, како и современи трансформер модели кои работат директно со самиот текст. Евалуацијата е спроведена врз комплексно податочно множество кое содржи текстови напишани од луѓе и генерирали од напредни јазични модели, на кои им е зададена цел да имитираат човечки стил.

Резултатите покажуваат дека традиционалните стилометриски пристапи имаат ограничена ефикасност во вакви услови, додека моделите со self-attention успеваат подобро да ги искористат комбинациите на стилските карактеристики. Највисоки перформанси се постигнати со трансформерските модели, кои покажуваат одлични перформанси за разликување помеѓу човечки и МГТ. Истражувањето укажува на предностите и ограничувањата на секој пристап и ја нагласува важноста на изборот на соодветен модел во зависност од практичните барања и ресурсите.

2. Вовед

Во последните години, развојот на вештачката интелигенција, а особено на големите јазични модели, доведува до напредок во автоматското генерирање на текст [4]. Современите модели се способни да создаваат граматички точни и стилски напредни текстови кои во многу случаи тешко се разликуваат од оние напишани од човек. Овој напредок овозможува широка примена на ваквите технологии, но истовремено отвара и доста многу нови предизвици поврзани со автентичноста и потеклото на текстуалната содржина.

Машински генерирали текст (МГТ) денес е присутен во различни дигитални средини, вклучувајќи образовни платформи, форуми и социјални мрежи. Иако ваквата содржина може да има корисна улога, како помош при пишување или автоматско создавање информативни текстови или давање идеи, нејзината употреба може да доведе и до проблеми како плаџијат, ширење на дезинформации и нарушување на довербата во дигиталната содржина. Поради тоа, се наметнува потребата од развој на сигурни и робустни методи за автоматско препознавање на МГТ [5].

Раните пристапи за препознавање на МГТ најчесто се засноваат на стилометриска анализа, односно на екстракција на рачно дефинирани јазични и структурни карактеристики како должина на реченици, разновидност на вocabулар, употреба на интерпункција и слично. Врз основа на овие карактеристики се применуваат традиционални модели од машинско учење, како логистичка регресија, SVM и Random Forest. Иако овие методи покажуваат солидни резултати кај текстови генерирали од постари модели, нивната ефикасност значително опаѓа кај современи јазични модели кои успешно го имитираат човечкиот стил на пишување [1].

Со појавата на подлабоки невронски архитектури и трансформерски модели, истражувањата во оваа област добиле нов правец. Трансформерите овозможуваат обработка на текстот во целост и добивање контекстуални и семантички податоци, што помага во добивањето на значително подобри перформанси во задачите за препознавање на МГТ. Сепак, и ваквите модели имаат одредени недостатоци, како високи пресметковни барања, подолго време на тренирање и ограничена интерпретабилност, што ја отежнува нивната практична примена во одредени сценарија [3].

Поради тоа, големо внимание се посветува и на хибридни пристапи кои комбинираат стилометриски карактеристики со невронски мрежи и механизми како self-attention. Овие модели имаат за цел да ги искористат предностите на стилската анализа, а истовремено да овозможат учење на нелинеарни односи помеѓу карактеристиките, со подобра интерпретабилност и пониски ресурси во споредба со големите трансформер модели [2].

Целта на оваа дипломска работа е да се направи компаративна анализа на различни пристапи за препознавање на машински генериран текст, применети врз комплексно и предизвикувачко податочно множество. Во рамки на истражувањето се анализираат традиционални модели од машинско учење, невронска мрежа со self-attention слој базирана на стилометриски карактеристики, како и трансформерски модели со fine-tuning. Преку детална евалуација и споредба, оваа работа има за цел да даде јасен увид во предностите и ограничувањата на секој пристап и да придонесе кон подобро разбирање на можностите за препознавање на МГТ во современи услови.

3. Сродни истражувања

3.1 StyloAI : Distinguishing AI-Generated Content with Stylometric Analysis

Во ова истражување, препознавањето на машински генериран текст се заснова на системот StyloAI кој е базиран на традиционални техники од машинското учење и на стилската анализа на текстот. Овој систем се потпира на рачно дефинирани карактеристики кои ги опишуваат стилските, лексичките, синтаксичките и структурните карактеристики на даден текст [1].

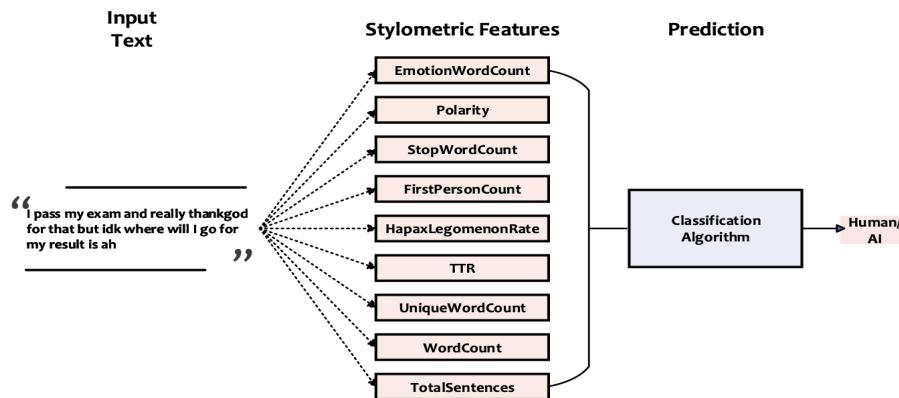
За евалуација на StyloAI, се користат две податочни множества. Првото е AuTeXTification Dataset кој содржи вкупно 55.677 примери од различни домени како твитови, вести, правни документи, рецензии и how-to статии. Овие текстови се од разни јавно достапни извори. Второто множество е Mindner et al. Education Dataset и е многу помало со 200 примероци, но содржи академски образовни текстови. Двете множества се балансираны.

Извлечени се 31 стилска карактеристики, меѓу кои има карактеристики како број на зборови, број на уникатни зборови, број на стоп-зборови, фреквенција на користење на интерпункциски знаци, карактеристики поврзани со сентиментот на текстот, уникатноста на зборовите кои се користат во текстот и слично. Подетално може да се разгледаат карактеристиките и групите во кои тие припаѓаат на слика 1.

Category	Features	Description
Lexical	<i>WordCount</i> <i>UniqueWordCount</i> <i>CharCount</i> <i>AvgWordLength</i> <i>TTR</i> <i>HapaxLegomenonRate</i>	The total number of words in the text. The number of unique words used in the text. The total number of characters in the text, including spaces and punctuation. Calculated as $\frac{\text{CharacterCount}}{\text{WordCount}}$. Measures lexical diversity, calculated as $\text{TTR} = \frac{\text{UniqueWordCount}}{\text{WordCount}}$. The proportion of words that appear only once in the text, $\frac{\text{Number of Words Appearing Once}}{\text{Total Words}}$.
Syntactic	<i>SentenceCount</i> <i>AvgSentenceLength</i> <i>PunctuationCount</i> <i>StopWordCount</i> <i>AbstractNounCount</i> <i>ComplexVerbCount</i> <i>SophisticatedAdjectiveCount</i> <i>AdverbCount</i> <i>ComplexSentenceCount</i> <i>QuestionCount</i> <i>ExclamationCount</i> <i>ContractionCount</i>	The total number of sentences in the text. Calculated as $\frac{\text{WordCount}}{\text{SentenceCount}}$. The total number of punctuation marks in the text. The total number of commonly used words. The number of nouns representing intangible concepts or ideas. The number of verbs not in the most common 5000 words. The number of adjectives with complex suffixes like "ive", "ous", "ic". The total number of adverbs in the text. The number of sentences with more than one clause, indicating complex sentence structures. The total number of questions, as indicated by question marks in the text. The total number of exclamations, as indicated by exclamation marks in the text. The total number of contractions in the text, such as "don't" and "can't".
Sentiment	<i>EmotionWordCount</i> <i>Polarity</i> <i>Subjectivity</i> <i>VaderCompound</i>	The total number of words associated with emotions in the text. Measures the text's sentiment orientation (positive, negative, or neutral). Measures the amount of personal opinion and factual information in the text. A sentiment analysis score that combines the positive, negative, and neutral scores to give a single compound sentiment score.
Readability	<i>FleschReadingEase</i> <i>GunningFog</i>	Calculated as $206.835 - 1.015 \left(\frac{\text{Total Words}}{\text{Total Sentences}} \right) - 84.6 \left(\frac{\text{Total Syllables}}{\text{Total Words}} \right)$. Estimates the years of formal education needed to understand a text on the first reading, calculated as $0.4 \left(\frac{\text{Word Count}}{\text{Sentence Count}} + 100 \left(\frac{\text{Complex Words Count}}{\text{Word Count}} \right) \right)$.
Named Entity	<i>FirstPersonCount</i> <i>DirectAddressCount</i> <i>PersonEntities</i> <i>DateEntities</i>	The number of first-person pronouns. The number of instances where the text directly addresses the reader or another hypothetical listener. The count of named individuals mentioned in the text. The count of date references within the text.
Uniqueness	<i>Bigram/trigramUniqueness</i> <i>SyntaxVariety</i>	These measures calculate the uniqueness of two-word and three-word combinations, indicating the originality and creative combinations of words in the text. The count of all the POS tags in a text.

Слика 1. Стилометриски карактеристики кои се користат во за StyloAI [1]

Иако овие карактеристики изгледаат доста едноставно, тие претставуваат многу важни индикатори за разликување на човечки и машински генериирани текстови, поради тоа што моделите често создаваат текстови со поедноставен стил и неприродна фреквенција на повторување на зборовите.



Слика 2. Архитектура на StyloAI [1]

Во ова истражување споредени се повеќе класични класификатори од машинското учење, како логистичка регресија, SVM (Support vector machine), дрво на одлука (Decision Tree), KNN и Gradient Boosting, но сепак Random Forest е моделот кој постигнува најдобри резултати. Тоа е поради концептот на комбинирање на повеќе дрва на одлука и откривање на разни нелинеарни односи помеѓу стилометриските карактеристики. Поради тоа Random Forest е избран како финална верзија на моделот StyloAI. Резултатите кои се добиени по евалуација покажуваат 81% точност на сите метрики на поголемото податочно множество (слика 3).

Model	Precision	Recall	F1-Score	Accuracy	AUC Score
Random Forest	0.81	0.81	0.81	0.81	0.88
SVM	0.79	0.79	0.79	0.79	0.87
Logistic Regression	0.71	0.71	0.71	0.71	0.77
Decision Tree	0.72	0.72	0.72	0.71	0.85
KNN Classification	0.73	0.73	0.73	0.73	0.80
Gradient Boosting	0.78	0.77	0.77	0.85	0.71

Слика 3. Евалуација на моделите од машинско учење со цел класификација МГТ врз AuTeXTification Dataset [1]

Исто така направена е и споредба на истото множество со други поголеми state-of-the-art модели (слика 4) и овде овој модел покажува најдобри резултати. Ова покажува дека добро избраните стилометрички карактеристики можат да бидат клучен чекор за надминување на комплексните длабоки модели.

Model	F1-Score
<i>StyloAI</i>	0.81
TALN-UPF	0.80
BOW+LR	0.74
LDSE	0.60
SB-FS	0.59
Transformer	0.57
Random	0.50
SB-ZS	0.33

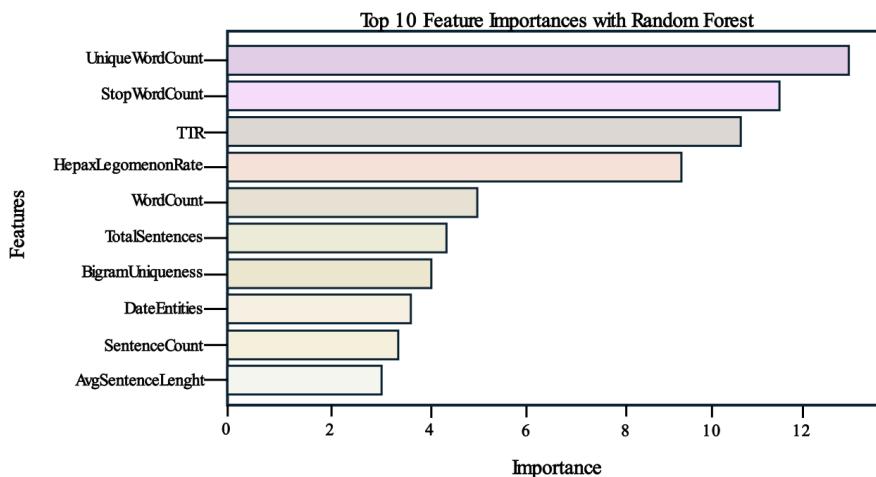
Слика 4. Споредба на StyloAI со останати модели врз AuTexTification податочното множество [1]

Врз помалото податочно множество добиени се резултати со вредност од 0.97 за F1-Score и 0.98 за Accuracy. Овие резултати се прикажани на табела 4, и се речиси идентични со најдобрите модели кои се оставени за споредба, но сепак без потреба од премногу пресметковни ресурси.

Model	F1-Score	Accuracy
<i>StyloAI</i>	0.97	0.98
Mindner et. al's Features + XGBoost	0.90	0.90
Mindner et. al's Features + Random Forest	0.98	0.98
Mindner et. al's Features + MLP	0.87	0.87

Слика 5. Споредба на StyloAI со останати модели врз Mindner et al. Education податочното множество [1]

Највлијателните 10 карактеристики се прикажани на сликата а тоа дел од нив се бројот на уникатни зборови, бројот на стоп-зборови, соодносот помеѓу уникатните зборови и вкупниот број зборови (TTR) и процент на зборови кои се појавуваат само еднаш (HepaxLegomenonRate).



Слика 6. Највлијателните 10 карактеристики за препознавање на МГТ со Random Forest [1]

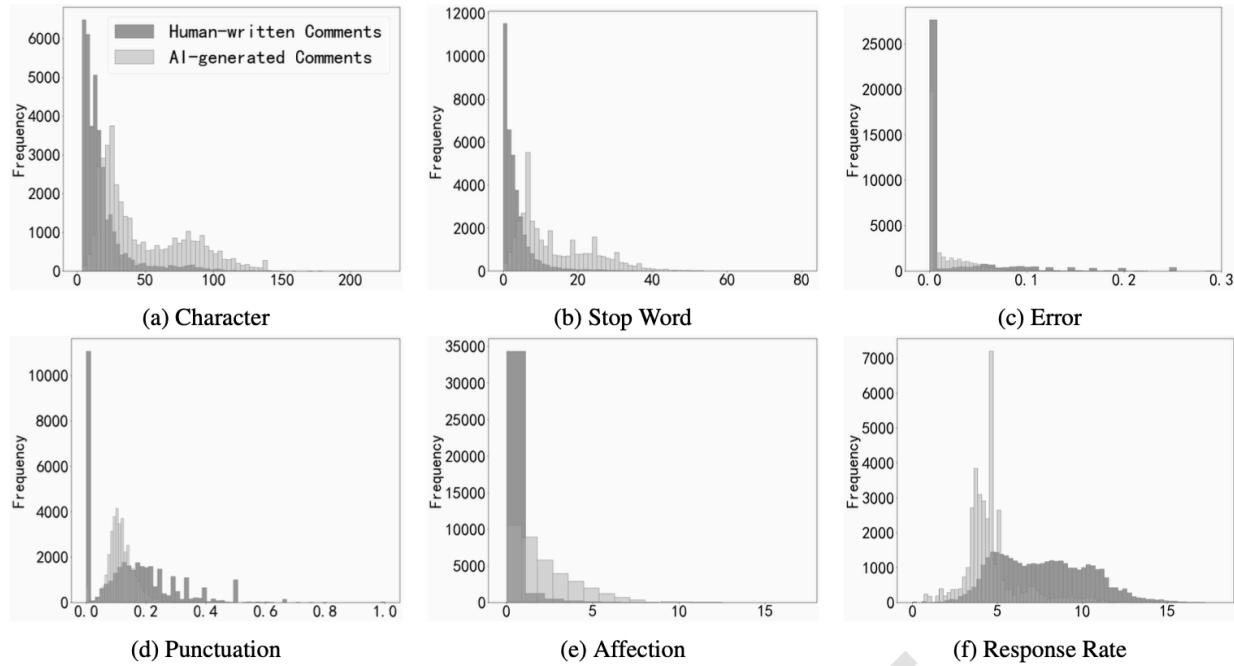
Резултатите од двете податочни множества покажуваат дека StyloAI, иако базиран на едноставни стилометрички карактеристики на текстот и традиционален Random Forest класификатор, постигнува перформанси споредливи или подобри од современите трансформерски и zero-shot модели [1].

3.2 Linguistic Differences between AI and Human Comments in Weibo: Detect AI-Generated Text through Stylistic Features

Во ова истражување е разгледан пистапот на препознавање на машински генериран текст (МГТ) наспроти текстови напишани од човек, преку стилометрички карактеристики на текстот со помош на едноставни невронски мрежи. Овој пристап го решава проблемот за препознавање на МГТ преку Stylistic Feature-Based Self-Attention Classifier (SFSC) односно класификатор со self-attention слој базиран на стилометрички карактеристики, врз множество на податоци составено од 463.382 коментари извадени од Weibo (кинеска социјална мрежа) на кинески јазик. Овие стилометрички карактеристики се рачно креирани и ги опишуваат јазичните, структурните и емоционалните својства во текстот [2].

Се користат 34 стилометрички карактеристики кои се групирани во категории како лексички структури (просечната должина на реченица, вкупен број на карактери во текстот, вкупен број на реченици и слично), прагматски стил (сооднос на негацииски зборови и вкупен број зборови и слично), читливост (број на грешки, сврзници и прилози во однос на должината на текстот...), користење на знаци (број на интерпункциски знаци, празни места, броеви и слично) и сентимент (број на зборови поврзани со различно расположение како среќа, тага и слично во текстот) [2].

Овие категории помагаат да се препознае дали текстот е МГТ, бидејќи моделите кои генерираат текст најчесто користат многу повеќе зборови од луѓето, не прават грешки при генерирањето на текстот, имаат нормална распределба при користењето на интерпункциски знаци - додека луѓето со својот стил на пишување користат или малку или нормално или премногу интерпункциски знаци. Исто така забележано е и дека овие коментари под разни објави на социјалната мрежа, се генериирани во навистина брзо време по времето на објавување, додека луѓето коментираат после одредено време. Сепак забележано е и дека има дел од овие МГТ кои се генериирани исто така после одредено време со цел имитација на човековото однесување. Оваа компарација на овие карактеристики може да се забележи на хистограмите на слика 7.



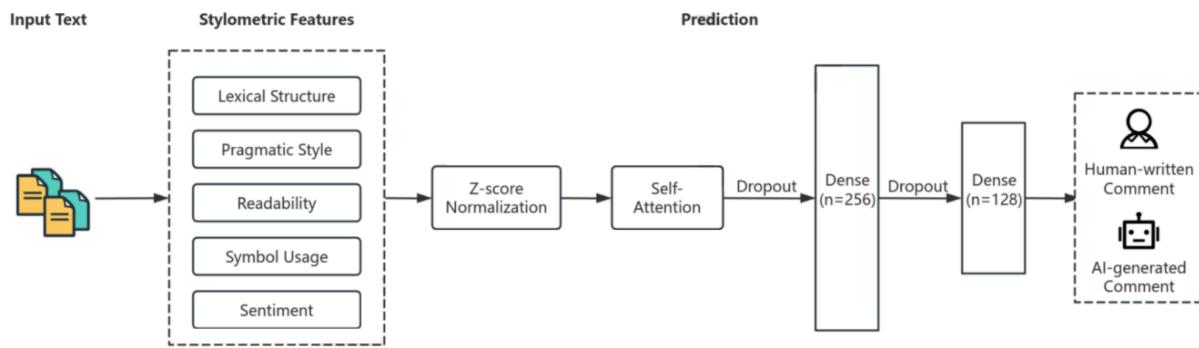
Слика 7. Споредбени хистограми на човечки и коментари генерирали од ВИ низ стилометрички карактеристики и ратата на одговор [2]

Токму поради овие забележани разлики во карактеристиките во текстот пишуван од страна на човек и МГТ, може да се добие добри резултати со класификатор базиран на стилометриските карактеристики.

Пред класификација, сите овие карактеристики се стандардизираат со Z-score нормализација, што помага да се постигне рамномерно скалирање со цел да се стабилизира учењето од страна на невронските мрежи.

Оваа SFSC архитектура има self-attention слој кој работи директно врз карактеристиките и ја учи важноста на секоја од нив поставувајќи различни тежини на секоја од карактеристиките. Со тоа моделот ги забележува најчестите стилски шеми што алудираат дека одреден текст е машински генериран [2].

После овој self-attention слој, во оваа архитектура има невронска мрежа составена од два целосно поврзани - Dense слоеви (Multilayer Perceptron - MLP) со 256 и 128 неврони и ReLU активацијска функција. Овие два слоја служат како главен класификатор и со нивна помош, моделот учи разни односи помеѓу карактеристиките кои не се лесно забележливи. За да не дојде до overfitting, се користи и dropout регуларизација со 0.1 стапка. Крајниот слој користи sigmoid активацијска функција, додека процесот на учење се врши со бинарна cross-entropy функција на загуба. Архитектурата на SFSC е претставена на слика 8.



Слика 8. Архитектура на SFSC [2]

Резултатите од ова истражување покажуваат дека со оваа архитектура, врз ова податочно множество, се добиваат конкурентски перформанси во однос на класични методи на машинско учење (логистичка регресија, KNN, RF и NB), а дури и за многу малку помали перформанси од големи трансформерски модели како RoBERTa, додека времето на тренирање и тестирање е значително помало. Исто така овој начин е доста компјутерски ефикасен бидејќи само работи со нумерички карактеристики наместо со сиров текст. Исто така овој механизам овозможува и увид во тоа кои карактеристики најмногу придонесуваат во одлуката на моделот, што го прави пристапот едноставен за анализа [2].

Резултатите од оваа компаративна анализа се претставени во слика 9, каде со задебелен стил на букви се претставени најдобрите перформанси, додека со подвлечена линија се претставени вторите по ред најдобри перформанси.

Model	Accuracy	Precision	Recall	F1-Score	Training Time(s)	Testing Time(s)
LR	0.708	0.788	0.570	0.661	<u>1.629</u>	0.001
KNN	0.581	0.545	0.969	0.698	115.681	11.803
RF	0.569	0.958	0.144	0.250	94.004	0.957
NB	0.584	0.563	0.750	0.643	0.285	<u>0.002</u>
RoBERTa	0.931	0.902	<u>0.968</u>	0.934	663.090	8.432
SFC	0.906	0.898	0.917	0.907	23.044	4.711
SFSC	<u>0.917</u>	0.911	0.925	<u>0.918</u>	31.240	4.926

Слика 9. Споредба на перформанси за препознавање на МГТ [2]

3.3 M4: Multi-Generator, Multi-Domain, and Multi-Lingual Black-Box Machine-Generated Text Detection

Истражувањето *M4: Multi-Generator, Multi-Domain, and Multi-Lingual Black-Box Machine-Generated Text Detection* претставува едно од најобемните и најрелевантни трудови во оваа област. Главната цел на трудот е да се разгледа проблемот за препознавање на машински генериран текст во black-box сценарија - ситуации кога немаме пристап до внатрешните параметри на моделот што го генерира текстот, туку располагаме само со готовиот текст. Во ова истражување нагласено е дека детекторите често имаат одлични резултати кога работат на ист домен и ист генератор, но драматично се влошуваат кога се соочуваат со нов генератор, друга тема или нов јазик. Во овој труд се адресира токму овој проблем и врз него направена е детална анализа.

Комплетното истражување е направено така што создадено е големо мулти-генератор, мулти-домен и мулти-јазично податочно множество од 147000 парови човек-машина, евалуирани се седум различни детектори, а има и анализа на разни сценарија каде моделите се тестираат на различен домен на текстови, на различни генератори, на различни јазици како и zero-shot сценарија. На крај се утврдуваат клучните слабости кај овие методи за препознавање на МГТ [3].

Подетално, податочното множество е создадено од два делови, паралелни човек-машина примери и дополнителни човечки текстови. Паралелните човек-машина примери се основниот дел од M4 и вкупно на број се **147.895 парови текстови**, при што секој човечки текст има соодветна машински генерирана верзија. Овие податоци се распределени низ повеќе домени (Wikipedia, WikiHow, Reddit ELI5, arXiv, PeerRead и др.) и се генериирани од шест различни LLM генератори (ChatGPT, GPT-4, davinci-003, Cohere, Dolly-v2 и BLOOMZ). Овој дел од множеството е најрелевантен за задачата на препознавање МГТ [3].

Дополнителните текстови напишани од луѓе, покрај паралелните податоци, се над 10 милиони на број и тоа се човечки текстови, преземени од големи јавни корпуси (Wikipedia dumps, Reddit, arXiv и др.). Овие текстови не се користат за директно тренирање на детектори, туку служат за статистички анализи на човечкиот јазичен стил, вокабуларна разновидност и споредбени n-gram карактеристики. Тие помагаат во разбирањето на разликите меѓу човечкото и машинското пишување [3].

Разновидноста на ова податочно множество може да се забележи и на слика 10 на која се прикажани податоците кои се користат.

Source/ Domain	Data License	Language	Total	Parallel Data								Total
				Human	Davinci003	ChatGPT	GPT4	Cohere	Dolly-v2	BLOOMz		
Wikipedia	CC BY-SA-3.0	English	6,458,670	3,000	3,000	2,995	3,000	2,336	2,702	3,000	20,033	
Reddit ELI5	Huggingface	English	558,669	3,000	3,000	3,000	3,000	3,000	3,000	3,000	21,000	
WikiHow	CC-BY-NC-SA	English	31,102	3,000	3,000	3,000	3,000	3,000	3,000	3,000	21,000	
PeerRead	Apache license	English	5,798	5,798	2,344	2,344	2,344	2,344	2,344	2,344	19,862	
arXiv abstract	CC0-public domain	English	2,219,423	3,000	3,000	3,000	3,000	3,000	3,000	3,000	21,000	
Arabic-Wikipedia	CC BY-SA-3.0	Arabic	1,209,042	3,000	–	3,000	–	–	–	–	6,000	
True & Fake News	MIT License	Bulgarian	94,000	3,000	3,000	3,000	–	–	–	–	9,000	
Baike/Web QA	MIT license	Chinese	113,313	3,000	3,000	3,000	–	–	–	–	9,000	
id_newspapers_2018	CC BY-NC-SA-4.0	Indonesian	499,164	3,000	–	3,000	–	–	–	–	6,000	
RuATD	Apache 2.0 license	Russian	75,291	3,000	3,000	3,000	–	–	–	–	9,000	
Urdu-news	CC BY 4.0	Urdu	107,881	3,000	–	3,000	–	–	–	–	9,000	
Total				35,798	23,344	32,339	14,344	13,680	14,046	14,344	147,895	

Слика 10. Статистика за M4 податочното множество [3].

Во студијата се тестираат **седум различни детектори**, кои покриваат и класични методи на машинско учење, и современи трансформерски архитектури. Тоа се следните модели:

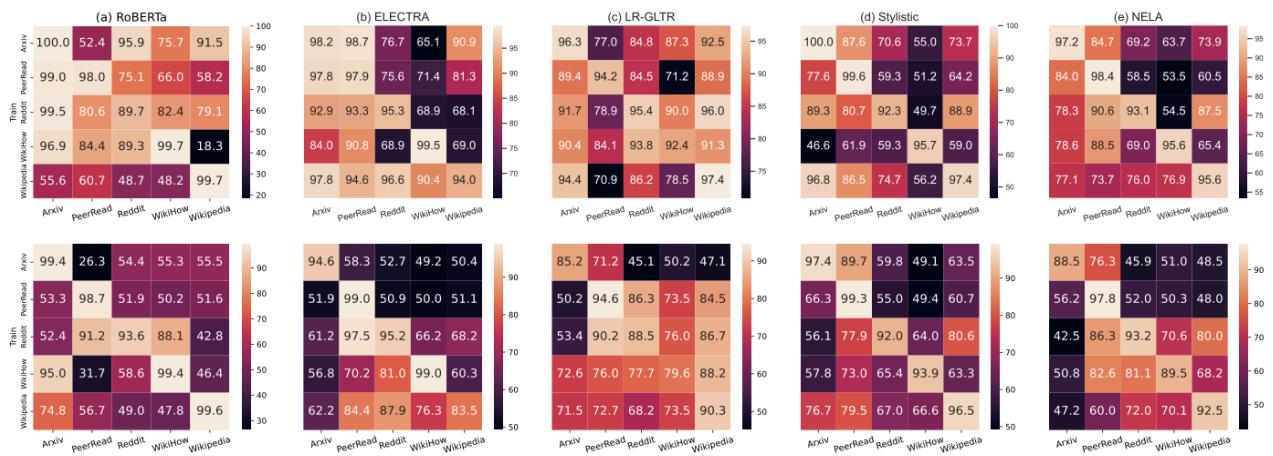
- **RoBERTa** - со помош на fine-tuning се покажал како **најмоќен детектор** во голем дел од експериментите. Поради големиот број параметри и добрата претходна обука на големи корпуси, RoBERTa постигнува речиси совршени резултати за *in-domain* препознавање.
- **ELECTRA** - постигнува **најдобри резултати во out-of-domain поставки**, каде останатите модели, вклучително и RoBERTa, често се соочуваат со драстични падови на перформансите.
- **XLM-R (Multilingual Transformer)** - Овој модел е оптимизиран за мултијазични сценарија и е единствениот детектор кој може да обработува текстови на сите јазици претставени во M4. Резултатите покажуваат дека XLM-R е одличен избор за **cross-lingual препознавање**, но сепак чувствителен на јазична разлика помеѓу тренинг и тест средината.
- **Logistic Regression со GLTR карактеристики** - се заснова на анализирање на веројатносната распределба на зборовите што еден јазичен модел би ги избрал во даден контекст. Односно се креираат карактеристики во врска со текстот и потоа со логистичка регресија врз овие карактеристики се предвидува дали текстот е МГТ. Иако моделот е едноставен, тој покажува стабилни резултати и се одликува со робустност во сценарија каде имаме примери од различен домен.
- **Stylistic features + SVM** - Овој детектор се потпира на стилометриски карактеристики како должина на реченица, број на зборови, пунктуација, функции зборови и др. Резултатите се умерени, но овие карактеристики се важни затоа што нудат појасна интерпретабилност на текстот.

- **NELA Features + SVM** - NELA првично е развиена за препознавање на лажни вести, и анализира стил, сложеност, тон, морални рамки и др. Но моделот покажува посебни резултати во МГТ сценарија, особено кај академски домени како arXiv. [3]

- **GPTZero (zero-shot)** - GPTZero е вклучен како практичен пример за „готов“ онлајн детектор. Резултатите се добри само кога доменот и генераторот се познати, но значително опаѓаат кај непознати ситуации, што ги покажува недостатоците на zero-shot пристап. [3,5]

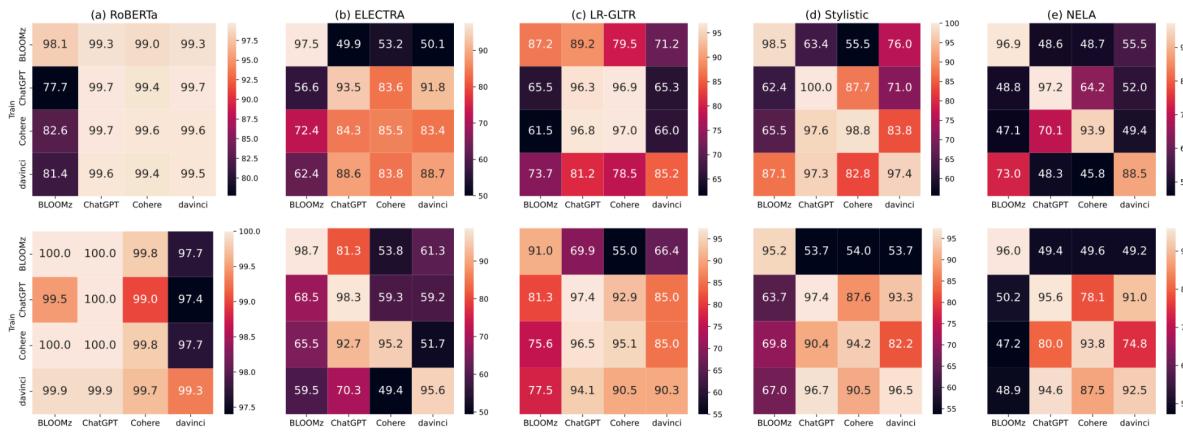
Во истражувањето направени се многу разновидни евалуации со цел да ја измери реалната робустност на детекторите. Главните видови се:

- **In-domain евалуација** - Ова е сценарио во кое моделот е трениран и тестиран на ист домен (на пример, трениран на Wikipedia податоци и тестиран на Wikipedia податоци). Резултатите покажуваат дека точноста може да достигне **99–100%**, RoBERTa доминира во ова сценарио, машински генериран текст е релативно лесно да се открие кога стилот, темата и должината се конзистентни. Но сепак од сите домени и стилови на пишување, на текстови од Reddit, препознавањето МГТ е најтешко, а на текстови од ArXiv е најлесно. Резултатите може да се забележат на слика 11 (главна дијагонала).
- **Cross-domain евалуација** - моделот се тренира на еден домен, а се тестира на друг (на пр. arXiv → Reddit). Резултатите покажуваат драматичен пад за разлика од In-domain евалуација. RoBERTa често паѓа на 50–60% точност (приближно случајно), додека ELECTRA најдобро се справува со доменска варијација со околу 87.9%, кога се тренирани на податоци од Wikipedia, а тестиирани на податоци од Reddit. При ваков тип на евалуација (Cross-domain), најтешко е да се препознае текст од WikiHow, а најлесно од PeerRead. Резултатите може да се забележат на слика 11, а тоа се сите податоци освен главната дијагонала.



Слика 11. Точност на cross-domain експерименти: ChatGPT генериирани текстови (горе) или davinci (долу), при што моделите се тренирани на еден домен и тестиирани на различни домени преку пет различни детектори. [3]

- **Cross-generator евалуација** - Овде моделите се тренираат на еден генератор (на пример ChatGPT) и тестираат на друг (davinci-003, BLOOMZ, Cohere итн.). Најдобар резултат се добива секогаш кога тренинг и тест генераторот се исти (In-generator евалуација). BLOOMZ е исклучително тешко да се препознае (recall < 5% во некои случаи). ChatGPT и Cohere слично генерираат текст, па препознавањето меѓу нив е полесно. Оваа евалуација може да се забележи на слика 12. Главната дијагонала во матриците претставува на некој начин In-generator евалуација, односно кога тренинг и тест генераторот се исти.



Слика 12. Точност на cross-generator експерименти: тренирање и тестирање на *arXiv* податоци (горе) и *Wikipedia* (долу) преку пет детектори, при што се споредува еден генератор на машински текст наспроти човечки текст. [3]

- **Zero-shot евалуација (GPTZero)** - Zero-shot препознавањето покажува сериозни ограничувања. На овој начин се добиваат добри резултати на Wikipedia и ChatGPT, а многу ниски резултати на arXiv и BLOOMZ (recall близку до 0%). Ова го покажува проблемот со готови, комерцијални детектори кои не се тренирани на разновидни домени.
- **Cross-lingual евалуација** - Со користење на XLM-R, тестирано е препознавање на седум јазици. Најдобри резултати се добиваат кога моделот е трениран и тестиран на ист јазик. Комбинирани мултијазични податоци даваат подобра генерализација. XLM-R не може да препознае јазик што не го видел во тренинг.
- **Евалуација на моделите низ време (различни генерации на моделите)** - Интересно е што моделите тренирани на ChatGPT изданија од март 2023 и понови верзии од септември 2023 покажуваат **многу мали разлики во точноста**, што значи дека одредени „генеративни отпечатоци“ на LLM остануваат стабилни со време.
- **Човечка евалуација** - Направена е човечка евалуација на примероци од Reddit и arXiv. Учесниците имале NLP познавање и имале просечна точност од 0.72–0.77, што е

значително пониско од автоматските модели, кои на истите податоци достигнувале речиси 100% точност. Дополнително, резултатите покажуваат дека човечката способност за препознавање е поврзана со разбирањето на стилот на LLM, а не само со познавање на английскиот јазик (бидејќи учесниците биле со различно познавање на английскиот и на другите јазици). Ова ја потврдува потребата од автоматски детектори, бидејќи луѓето тешко ги идентификуваат современите МГТ текстови. [3]

Направена е анализа како дужината на текстот влијае врз успешноста на детекторите. Резултатите покажуваат дека препознавањето е значително потешко кај кратки текстови. На пример, за arXiv, кога текстовите се скратени од 1000 на 125 карактери, точноста паѓа од околу 99% на приближно 94%. Ова укажува дека долгите текстови содржат повеќе стилометриски и статистички сигнали кои овозможуваат поуспешно препознавање [3].

Како главни заклучоци од студијата, може да се забележи дека препознавањето на МГТ е „решено“ само во контролирани услови односно кога доменот и генераторот се исти, тогаш препознавањето е скоро совршено. Генерализацијата е најголемиот предизвик, моделите драматично паѓаат кога треба да работат на нов домен, нов генератор или нов јазик. BLOOMZ покажува најголема „нестабилност“ односно многу тешко се препознае текст генериран од BLOOMZ, што укажува на голема разновидност во генеративниот стил. Комерцијални zero-shot детектори (GPTZero) не се доволно робустни, а подолгите текстови овозможуваат многу подобро препознавање.

4. Методи и истражувања

4.1 Вовед во податочното множество и стилометриска анализа

Стилометриската анализа претставува процес на анализирање на стилот на пишување со цел извлекување карактеристики кои се специфични за одреден автор, жанр или начин на создавање текст.

Во контекст на ова истражување, стилометриската анализа служи за идентификација на разлики помеѓу човечки и машински генерирали текстови, со помош на рачно креирани стилски карактеристики во врска со податочното множество од задачата од PAN@CLEF2025 – Voight-Kampff AI Authorship Verification [4].

Податочното множество е намерно создадено да биде тешко за препознавање, бидејќи машинските модели кои го креирале генерирали текст се насочени да го имитираат

стилот на конкретни човечки автори, со што разликите во стилометриските карактеристики се значително намалени [4].

Податочното множество се состои од 27.296 текстови, поделени на две класи: Класа 0 (Human-written) со вкупно 10.378 текстови и Класа 1 (Machine-generated) со 16.918 текстови [4].

Секој запис содржи text - што всушност е самиот текст кој е генериран или напишан, label - може да биде 0 (дека текстот е напишан од човек) или 1 (дека текстот е генериран од модел), genre - може да биде essays, fiction, news (во зависност во која категорија спаѓа текстот) и model - човечки автор или името на LLM-моделот кој се користи за генерирање на текстот.

Ова множество е посебно комплексно поради имитирањето на стилот како човек од страна на моделот, жанровска разновидност (текстови од различни домени) и употребата на нови LLM модели (GPT-4o, GPT-4.5, LLaMA 3.3, Gemini 2.0, DeepSeek R1...) [4].

Множеството изгледа на следниот начин:

```
{"id": "a6c8018e-d22c-4d6e-b5e3-0c0a65682a6a", "text": "...", "model": "human", "label": 0, "genre": "essays"}
```

```
{"id": "f1a26761-ca2a-43e9-890d-80dcb3058364", "text": "...", "model": "gpt-4o", "label": 1, "genre": "essays"}
```

Но бидејќи од ова множество потребно е да се креираат стилометрски карактеристики за предвидување на текстот, корисни податоци од ова множество се само “text” и “label”.

Во истражувањето се искористени **23 стилометрски карактеристики како и уште 2 дополнителни метрики за класификација на текстот**, поделени во неколку категории:

1. Лексички карактеристики (lexical features)

- a. word_count – вкупен број на зборови во документот
- b. unique word count – број на уникатни зборови
- c. hapax_rate – процент на зборови кои се појавуваат само еднаш
- d. ttr (Type-Token Ratio) – однос уникатни зборови / вкупни зборови
- e. avg_word_length – просечна должина на зборовите

2. Структурни карактеристики

- a. sentence_count – број на реченици
- b. avg_sentence_length – број на зборови по реченица

3. Stopwords и function words

- a. stopword_count – број на стоп-зборови (the, a, an, this, that, was, were, into...)
- b. functionword_count – број на функционални зборови (to, in, is, as...)
- c. function_word_ratio – однос на функционални зборови и вкупен број зборови

4. Интерпункциски карактеристики

- a. punctuation_count – број на интерпункциски знаци
- b. punctuation_ratio – процент на интерпункциски знаци во текстот
- c. comma_count – број на запирки во текстот
- d. period_count – број на точки во текстот
- e. question_mark_count – број на знаци прашалник во текстот
- f. exclamation_count – број на знаци извичник во текстот

5. Карактеристики на симболите

- a. digit_ratio – процент на цифри
- b. uppercase_ratio – процент на големи букви

6. Морфосинтаксички карактеристики (POS ratios)

- a. Noun_ratio – процент на именки во текстот
- b. Verb_ratio – процент на глаголи во текстот
- c. Adj_ratio – процент на придавки во текстот

7. N-gram карактеристики

- a. Bigram_uniqueness – број на уникатни биграми (секвенца од два последователни збора во текстот)

8. Семантички карактеристики

- a. lexical_density – однос на „содржински“ зборови (глаголи, именки и придавки) и вкупни зборови

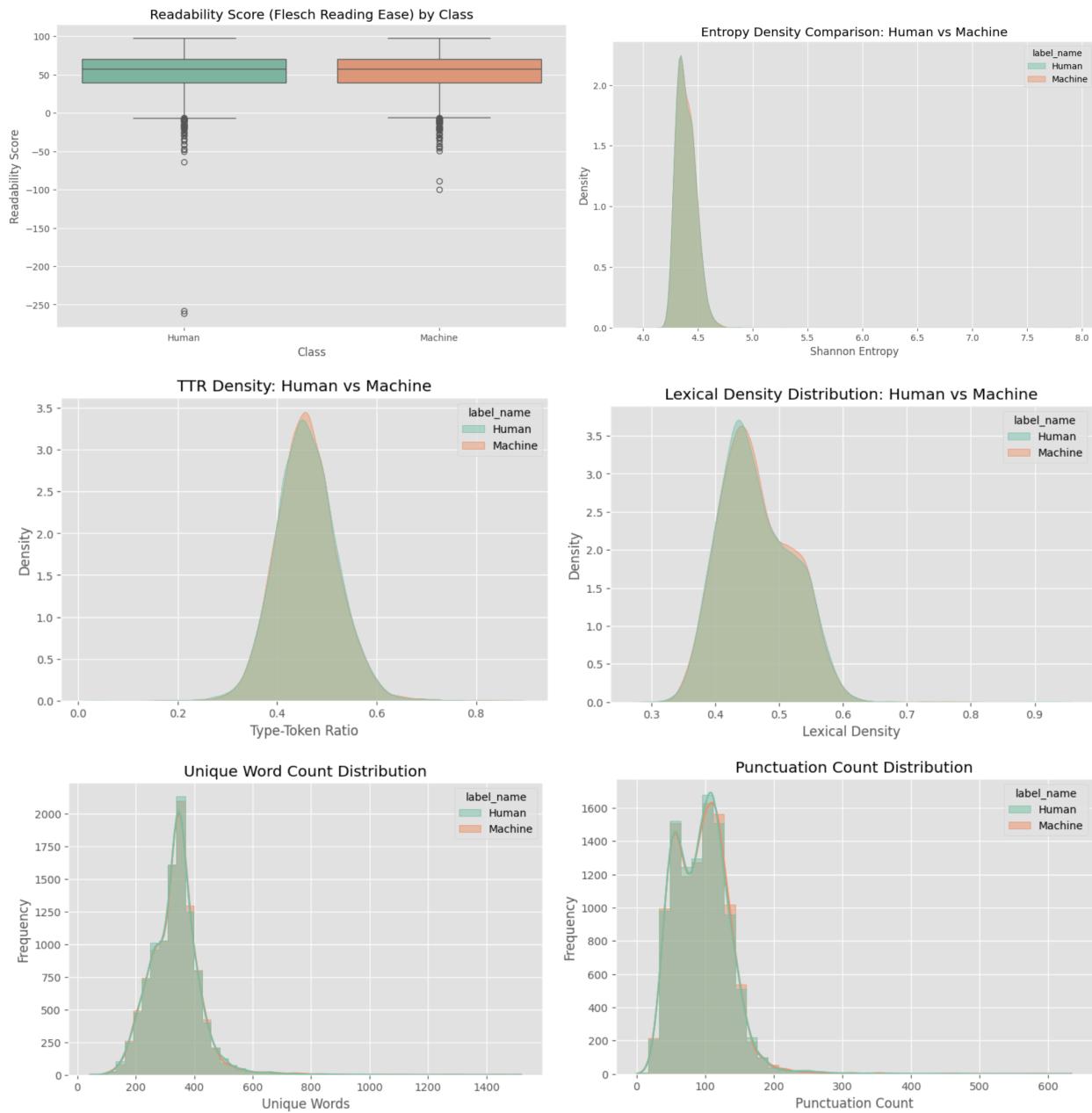
9. Дополнителни метрики кои се корисни за препознавање на МГТ

- a. Readability score (Flesch Reading Ease - метрика за читливост на текст што мери колку лесно или тешко е текстот да се чита.)
- b. Shannon_entropy (мерка за информациска „непредвидливост“, односно разновидни се зборовите, колку текстот е повторлив или колку една секвенца од текстот е предвидлива.)

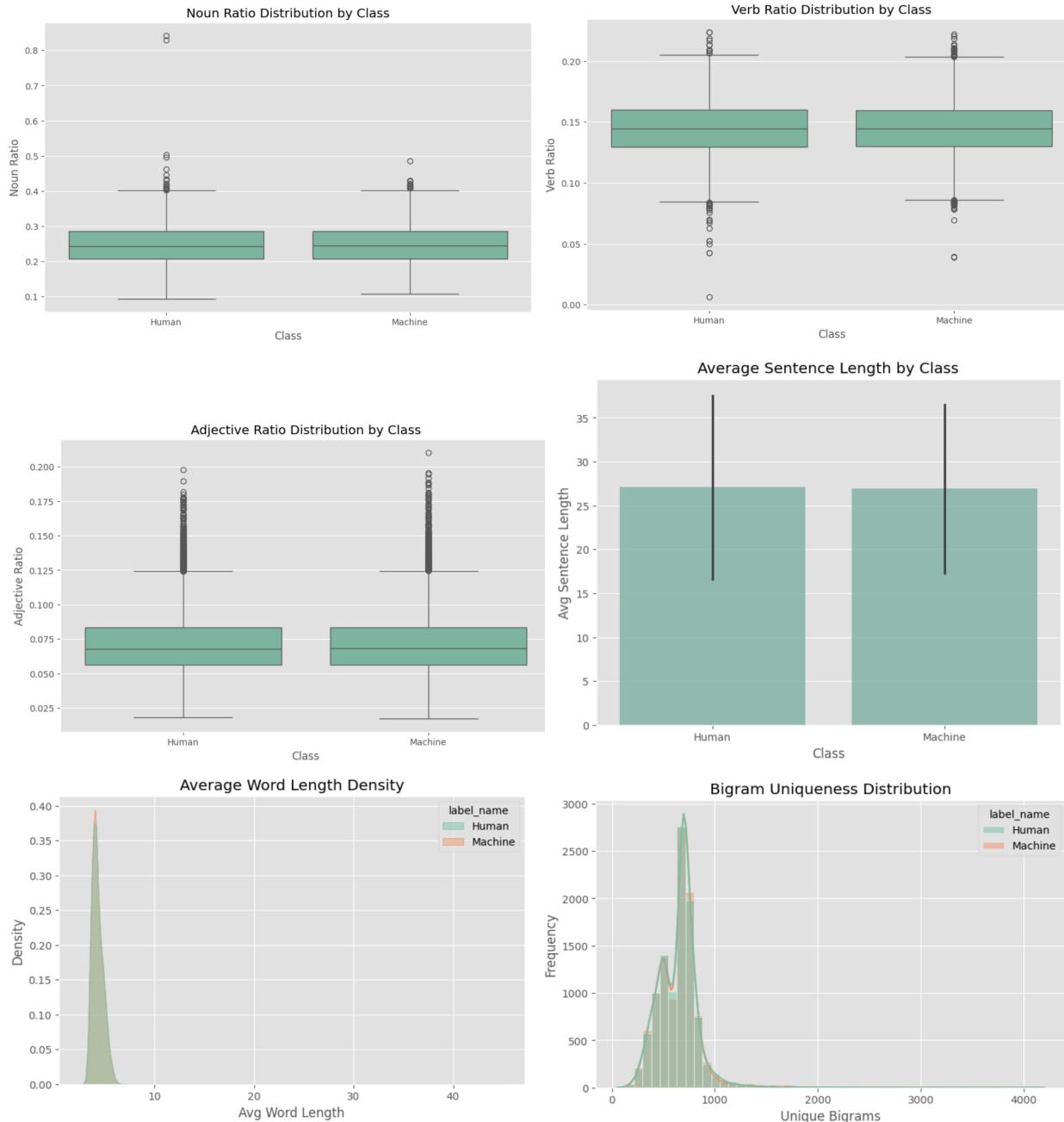
За екстракција на стилометриските карактеристики се користи библиотеката NLTK (Natural Language Toolkit) [14], која обезбедува токенизатори, POS tagger и листа на англиски stop words. NLTK stop words се користат за пресметување на `stopword_count`, додека POS tagger се користи за морфосинтаксичка анализа (`noun`, `verb` и `adjective ratio`). Покрај NLTK, дел од карактеристиките се пресметуваат рачно, како TTR, hapax rate, биграм уникатност, ентропија и лексичка густина.

4.2 Анализа и визуелизација на податочното множество преку стилометриските карактеристики

Врз овие стилометриски карактеристики направена е анализа со помош на графици. Направена е визуелизација на вкупно 12 стилометриски карактеристики со цел да се утврди дали постои јасна разлика во нивната распределба помеѓу човечки и машински генериирани текстови. Користени се различни видови графици, како KDE дистрибуции, хистограми, boxplot графици и barplot графици. Сите овие карактеристики покажуваат многу голема сличност и преклопување помеѓу класите, односно и оние текстови напишани од човек, како и текстовите генериирани од модел се во овие карактеристики многу слични. Тоа значи дека моделите генерирале текст во стил кој е многу сличен како и човечкото изразување. Сите овие графици можеме да ги разгледаме на сликите 13 и 14.



Слика 13. Визуелна анализа на дистрибуциите на стилометрски и лексички карактеристики кај човечки наспроти машински генериран текст

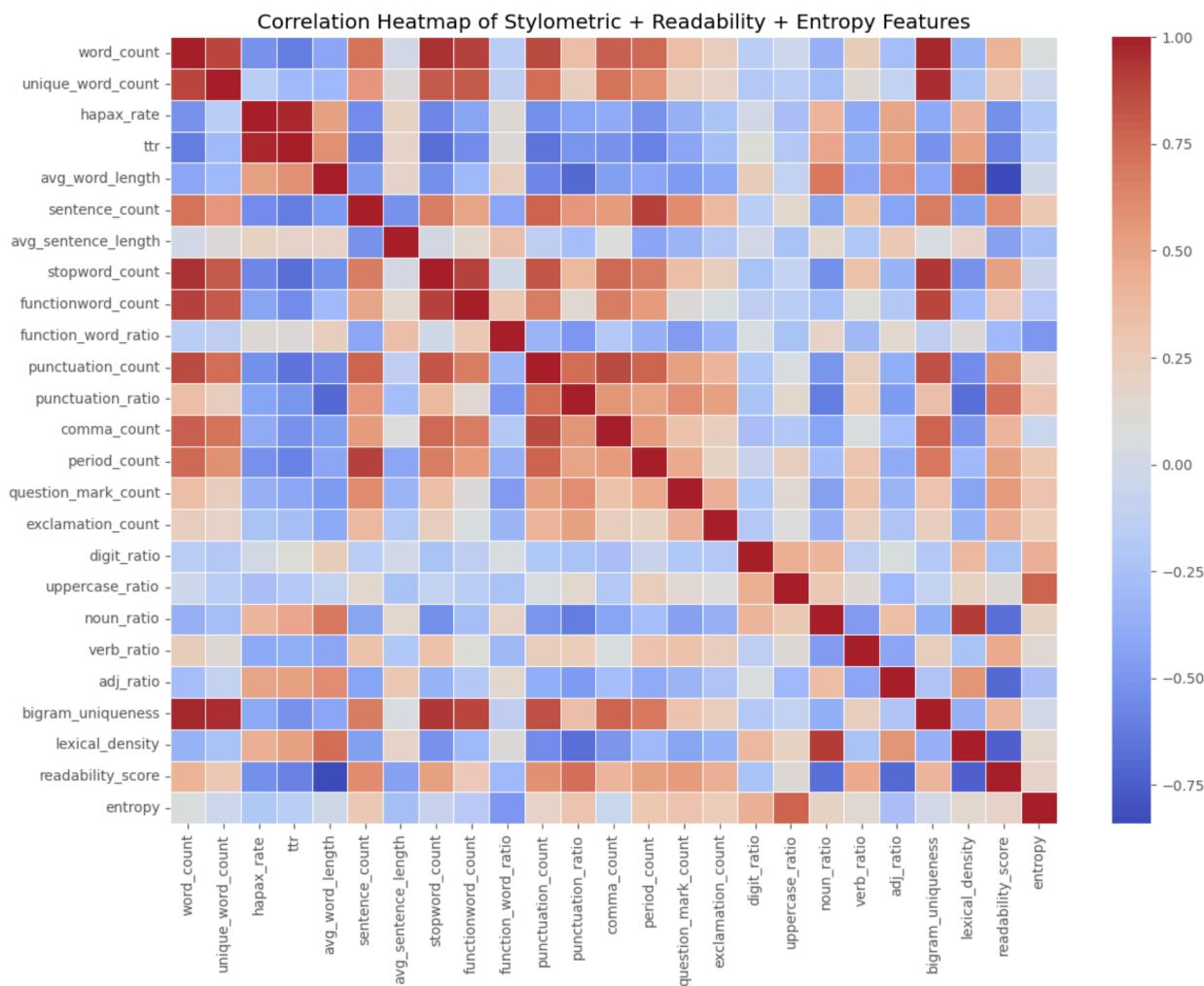


Слика 14. Визуелна анализа на дистрибуциите на стилометрски и лексички карактеристики кај човечки наспроти машински генериран текст

Според оваа анализа можеме да претпоставиме дека со користење на разни модели на машинско учење нема да добиеме добри резултати, односно дека е тешко според овие стилометрски карактеристики, моделот да определи дали еден текст е напишан од човек или не.

Исто така според графикот за Heatmap (корелацијската матрица) на слика 15, можеме да ги забележиме односите помеѓу сите карактеристики. Можеме да забележиме неколку групи на силно поврзани карактеристики, како на пример word_count, unique_word_count и bigram_uniqueness, кои логично се поврзани бидејќи подолгите текстови содржат повеќе уникатни зборови и биграми. Слично, карактеристиките поврзани со лексичката разновидност (TTR, hapax_rate и unique_word_count) исто така покажуваат висока позитивна корелација.

Од друга страна, readability_score има силна негативна корелација со просечната должина на реченицата, што е очекувано бидејќи подолгите и покомплексни реченици ја намалуваат читливоста на текстот. Shannon entropy покажува слаба корелација со останатите карактеристики, што укажува дека оваа метрика внесува нова информација независна од другите стилски показатели.



Слика 15. Корелацијска анализа на стилометрички карактеристики, метрики за читливост и ентропија кај анализираните текстови

4.3 Користење традиционални модели од машинско учење

За класификација на човечки и машински генериран текст се користени четири традиционални модели за машинско учење: Логистичка регресија, Linear SVM, Random Forest и Gradient Boosting. Овие модели се избрани бидејќи претставуваат најкористени алгоритми во стилометрија и во текстуална анализа базирана на рачно дефинирани карактеристики. Податоците најпрво се балансирали (10.378 примероци за секоја класа), по што е применето стандардизирање со StandardScaler на карактеристиките со цел подобрување на стабилноста на алгоритмите чувствителни на скалирање како што се Логистичка регресија и SVM.

Во сите експерименти е користена фиксна вредност random_state = 42. Овој параметар гарантира дека секое мешање и делење на податоците, ќе добиваме исто „случајно“ подмножество.

Податочното множество е поделено на 80% податоци за тренирање и 20% податоци за евалуација (тестирање) со еднаков број човечки и машински текстови во двете подмножества.

Тренирањето е извршено врз сите претходно наведени карактеристики. Секоја од овие карактеристики има улога во описувањето на стилот и сложеноста на текстот.

Резултатите од евалуацијата покажуваат дека сите четири модели имаат перформанси околу 49–50% Accuracy и F1-score, што укажува дека стилските карактеристики сами по себе не обезбедуваат јасни податоци за разликување класификација на текстот. Перформансите по модел се прикажани на табела 1.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.4906	0.4901	0.4639	0.4766
Linear SVM	0.4901	0.4896	0.4648	0.4769
Random Forest	<u>0.4993</u>	<u>0.4993</u>	0.4812	0.4901
Gradient Boosting	<u>0.4993</u>	<u>0.4993</u>	<u>0.5010</u>	<u>0.5001</u>

Табела 1. Перформанси на моделите Logistic Regression, Linear SVM, Random Forest и Gradient Boosting

Овие резултати укажуваат дека текстовите во PAN@CLEF2025 датасетот се стилски многу слични меѓу класите, што се должи на начинот на создавање на податочното множество - машинските модели намерно се упатени да имитираат стил на конкретни автори. Исто така користењето на понови и посилни модели влијае на стилометристката класификација, бидејќи овие модели се подобри во генерирањето на текст во стил на човек.

4.3.1 Споредба со постоечкото истражување StyloAI

Доколку го споредиме ова истражување со истражувањето за алатката StyloAI, тој систем користи 31 стилометриска карактеристика и истите четири класификатори, но добива далеку подобри резултати - околу 81% точност на големото податочно множество и 97% F1-score на помалото. Главната причина за оваа разлика не е во моделите или карактеристиките, туку во самите податоци [1].

Множествата во StyloAI содржат текстови генериирани од постари модели, кои често покажуваат карактеристики кои укажуваат на машински генериран текст како поедноставена синтакса, повисока повторливост, помала лексичка разновидност, повисока употреба на функционални зборови и карактеристични интерпункциски знаци. Затоа Random Forest во StyloAI може да ги препознае овие стилски разлики [1].

Иако методологијата е слична со StyloAI, реалните перформанси драстично се разликуваат поради разликите во податочното множество. Според овој резултат можеме да заклучиме дека традиционалниот начин на препознавање на МГТ со стилски карактеристики, веќе не е доволен за разликување на модерни LLM текстови од човечки текстови, особено кога моделите имаат задача да имитираат стил на човек.

4.4 Користење на повеќеслојна невронска мрежа и self-attention слој – SFSC (Stylometric Feature-Based Self-Attention Classifier)

Во овој дел се претставува вториот експериментален пристап кој всушност претставува невронска мрежа заснована на стилометриски карактеристики. Целта е да се испита дали едноставен и компјутерски ефикасен модел со self-attention може да постигне конкурентски резултати во разликување на човечки и машински генериран текст.

4.4.1 Опис на пристапот и архитектура

По екстракција и стандардизација на 25 стилометриски карактеристики, на потполно истиот начин со истите стилометриски карактеристики, конструираме модел кој работи директно врз овие нумерички вектори. Моделот се состои од два дела, а тоа се Self-Attention слој и MLP (Multilayer Perceptron) кој претставува невронска мрежа.

Во овој дел од истражувањето, претпоставуваме дека невронските мрежи ќе постигнат подобри резултати од традиционалните ML модели затоа што можат да учат посложени и подлабоки врски меѓу стилометриските карактеристики. Иако секоја карактеристика сама по себе не прави голема разлика помеѓу човечки и машински текст, односно разликата помеѓу двата вида на текстови не е голема според нивните карактеристики, комбинацијата

на овие карактеристики создава шеми кои не се линеарни и не можат лесно да се забележат и да имаат влијание со модели како логистичка регресија или линеарен SVM.

MLP може да моделира вакви нелинеарни шеми поради својата повеќеслојна структура и активациите функции, што му овозможува подобро да ги разбере минималните стилски разлики. Дополнително, со вклучувањето на self-attention слојот, моделот може да процени кои карактеристики се поважни и да прилагоди различни тежини за секоја карактеристика, што дополнително ја зголемува точноста.

Поради ова, очекуваме MLP со attention да биде поефективен од класичните ML пристапи при препознавање на МГТ и да добиеме поголема точност.

Како што беше споменато, Self-Attention слојот се користи за секој пример и пресметува тежини за сите 25 карактеристики. На овој начин моделот динамички учи кои карактеристики имаат најголемо влијание врз крајниот резултат за предвидување, а и како нивната важност варира од пример до пример. Исто така преку овој слој можеме да направиме и преглед за важноста на овие карактеристики.

Вториот дел е класичен MLP составен од два скриени слоја со 256 и 128 неврони, ReLU активација и Dropout(0.1) за намалување на overfitting. Крајниот слој генерира логит вредност, која преку сигмоид функција се пресликува во веројатност од 0 до 1 во зависност од предвиденото за во која класа припаѓа текстот.

Архитектурата е генерирана со помош на PyTorch [6] изгледа и на следниот начин:

```
SFSCModel(  
    (attention_layer): SelfAttentionLayer(  
        (attention): Linear(in_features=25, out_features=25,  
        bias=True)  
    )  
    (mlp): Sequential(  
        (0): Linear(in_features=25, out_features=256, bias=True)  
        (1): ReLU()  
        (2): Dropout(p=0.1)  
        (3): Linear(in_features=256, out_features=128, bias=True)  
        (4): ReLU()  
        (5): Dropout(p=0.1)  
        (6): Linear(in_features=128, out_features=1, bias=True)  
    )  
)
```

Оваа архитектура е всушност истата како и во истражувањето кое е обработено претходно („Linguistic Differences between AI and Human Comments in Weibo“) [2].

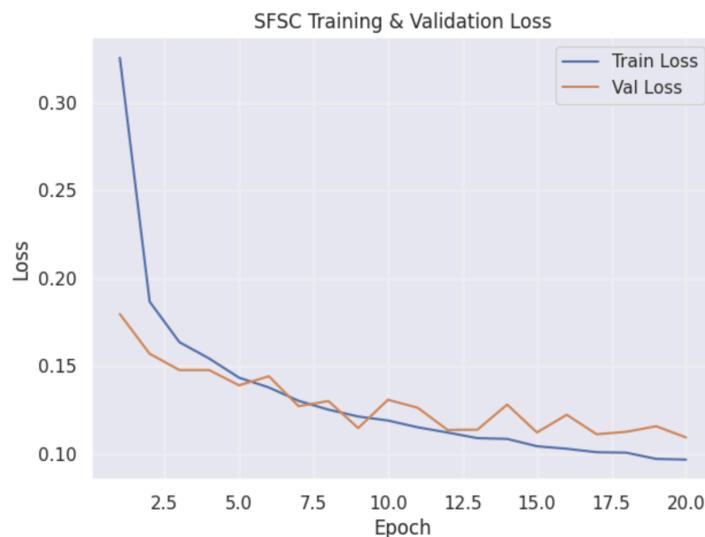
Моделот е трениран со Binary Cross Entropy loss функција (BCEWithLogitsLoss), која е стандардна за бинарни класификацији, и оптимизиран со Adam, еден од најкористените оптимизатори за стабилно и брзо учење. Тренингот се изведува во batch-ови од 32 примероци и трае 20 епохи, со механизам за рано запирање (early stopping) кој прекинува доколку нема подобрување на валидациската загуба (validation loss) во 5 последователни епохи.

По стандардизацијата, множеството е поделено на:

- Тренинг подмножество со 14943 примероци
- Валидациско подмножество со 1661 примероци
- Тест подмножество со 4152 примероци

4.4.2 Резултати од моделот со невронска мрежа и self-attention слој базиран на стилометрички карактеристики (SFSC)

На слика 16 прикажана е кривата на тренинг и валидациска загуба (train loss и val loss). Графикот покажува стабилна конвергенција на моделот, без значајни знаци на overfitting. Тренинг загубата континуирано опаѓа, додека валидациската загуба се движи во многу тесен опсег, со мала варијација. Најниската валидациска загуба е постигната во последната епоха.



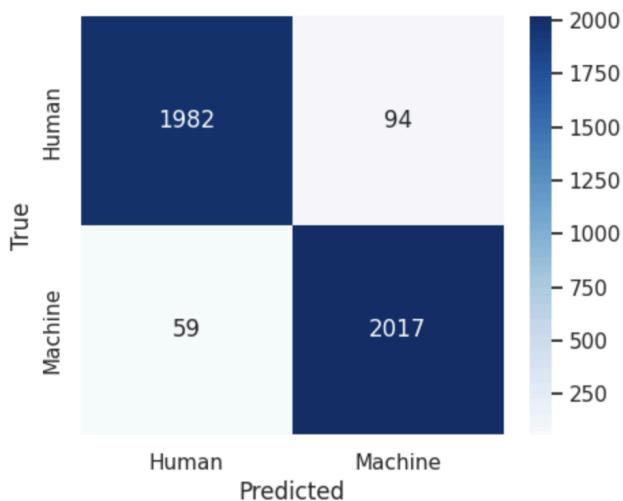
Слика 16. Кривата на тренинг и валидациска загуба (train loss и val loss)

Конечните резултати на тест множеството се прикажани на табела 2, каде може да се забележи точност на моделот со F1-score од 95% :

Accuracy:	0.963
Precision:	0.955
Recall:	0.972
F1-score:	0.963

Табела 2. Перформанси на моделот со невронска мрежа и self-attention слој базиран на стилометриски карактеристики (SFSC)

На слика 17 е прикажана и матрицата на конфузија. Може да се забележи дека моделот прави релативно мал број погрешни класификации, особено кај машински генериирани текстови.



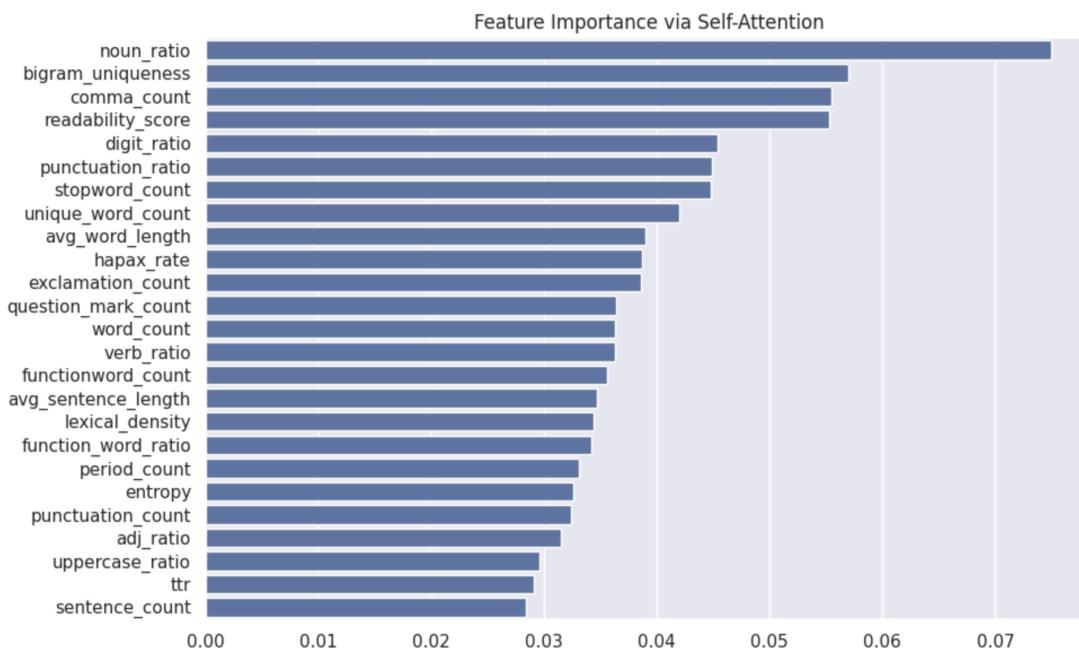
Слика 17. Матрица на конфузија за моделот за препознавање на МГТ

4.4.2.1 Анализа на важност на карактеристиките

Self-attention слојот овозможува да се изведе анализа за тоа кои карактеристики најмногу придонесуваат кон одлуката на моделот. На Слика 18 се прикажани просечните attention тежини.

Најзначајни карактеристики се:

- **Noun_ratio** (Односот помеѓу бројот на именки и вкупниот број зборови во текстот)
- **Bigram_uniqueness** (Број на уникатни биграми - парови соседни зборови)
- **Comma_count** (Вкупен број запирки во текстот)
- **Readability_score** (Процена колку е лесен или тежок текстот за читање)
- **Digit_ratio** (Однос на бројките во текстот во однос на вкупниот број карактери.)
- **Punctuation_ratio** (Однос на интерпункциски знаци во текстот во однос на сите карактери)



Слика 18. Визуелизација на релативната важност на стилометриските карактеристики според self-attention слојот

Ова значи дека овие карактеристики добиле најголема важност при донесувањето на одлуките во self-attention слојот, односно моделот најмногу се потпира на нив при учење на шаблоните што го разликуваат машински генерираниот текст од човечкиот.

4.4.3 Споредба со постоечки истражувања

Добиените резултати се споредливи со резултатите од истражувањето “*Linguistic Differences between AI and Human Comments in Weibo*”, каде предложениот SFSC модел постигнува конкурентски резултати во однос на традиционалните ML методи, а е значително побрз и поедноставен од трансформер моделите [2].

Доколку ги споредиме перформансите на оваа архитектура (табела 3) врз податочното множество од Weibo (на кинески) и множеството кое ние го користиме, може да забележиме дека врз нашето множество добиваме подобри перформанси [2].

	Weibo-dataset	PAN@CLEF2025
Accuracy	0.917	<u>0.963</u>
Precision	0.911	<u>0.955</u>
Recall	0.925	<u>0.972</u>
F1-Score	0.918	<u>0.963</u>

Табела 3. Споредба на SFSC моделот врз Weibo податочното множество и PAN@CLEF2025 податочното множество со помош на стилски карактеристики

Овој SFSC модел покажува висока точност, ниско време на тренирање, полесен преглед на карактеристиките преку attention тежините и мали барања за ресурси за двете податочни множества. Со тоа, овој пристап претставува практично и ефикасно решение за задачава на препознавање на машински генериран текст.

4.5 Експерименти базирани на трансформер модели

Во ова истражување беше имплементиран и евалуиран пристап за препознавање на МГТ базиран на трансформерски архитектури. За разлика од претходните експерименти со стилометриски карактеристики, тука моделите работат директно со целосниот текст од податочното множество PAN@CLEF2025, без рачно дефинирани карактеристики. Текстот е обработен преку токенизација и претставен во форма погодна за трансформерски модели [4].

Податочното множество е поделено на тренинг и тест подмножество во однос 80% / 20%, со користење на *stratified split* со цел да се зачува оригиналната класна распределба. За да се обезбеди репродуктивност на експериментите, користена е фиксна вредност *random_state = 42*.

За токенизација беше користен AutoTokenizer од HuggingFace библиотеката. Во иницијалната фаза беше користена максимална должина на текст од 256 токени, додека за финалниот модел должината беше зголемена на 512 токени со цел да се задржи повеќе контекстуални информации [7].

4.5.1 Споредба на перформансите на *lightweight* трансформер модели

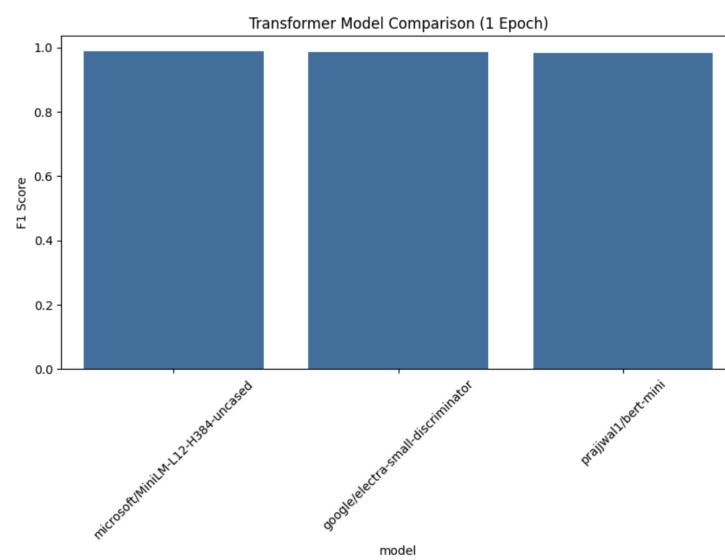
Во првата фаза беа тестиирани три компактни (*lightweight*) трансформерски модели, со цел да се пронајде најсоодветен кандидат за fine-tuning. Тоа се моделите *microsoft/MiiniLM-L12-H384-uncased*, *google/electra-small-discriminator* и *prajjwall/bert-mini* [8, 9, 10].

Сите модели најпрво се тренирани на една епоха со максимална должина на текст од 256 токени, под исти услови и евалуирани со метриките: точност (Accuracy), прецизност (Precision), одзив (Recall) и F1-мерка.

Резултатите од оваа споредба се прикажани на табела 4 и визуелно прикажани на Слика 19. Од добиените резултати може да се забележи дека сите модели постигнуваат високи и слични перформанси. Сепак, моделот **MiniLM-L12-H384-uncased** [8] ја постигнува највисоката F1-мерка и затоа е избран како финален модел за понатамошно тренирање и детална анализа.

Модел	Training loss	Validation loss	Accuracy	Precision	Recall	F1
MiniLM-L12-H384	<u>0.0669</u>	<u>0.0661</u>	<u>0.9841</u>	<u>0.9822</u>	<u>0.9923</u>	<u>0.9872</u>
ELECTRA-small	0.0752	0.0741	0.981	0.9793	0.9902	0.9847
BERT-mini	0.1148	0.0899	0.9775	0.982	0.9817	0.9818

Табела 4. Споредба на перформансите на трансформер моделите

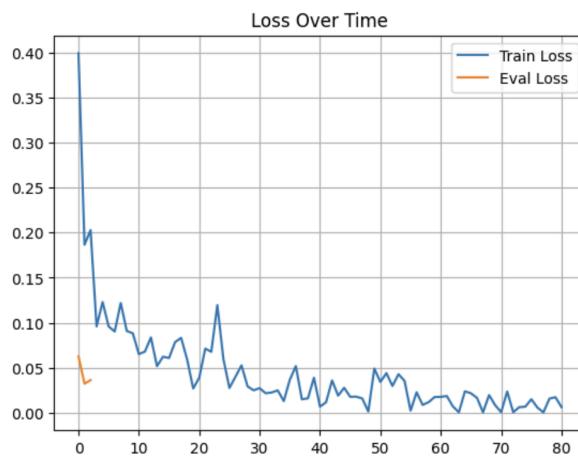


Слика 19. Столбест график за споредба на F1-score на трансформер моделите по тренирање на една епоха.

4.5.2 Fine-tuning и детална евалуација на финалниот модел

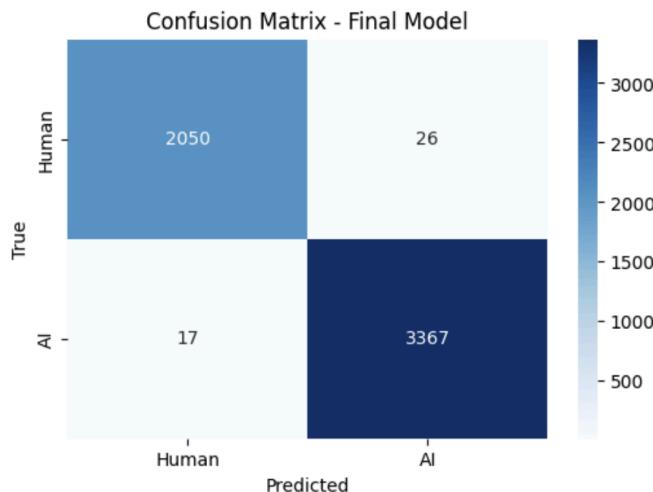
Избраниот модел MiniLM-L12-H384-uncased е дополнително трениран во траење од 3 епохи, со максимална должина на текст од 512 токени. Вкупното време на тренирање изнесуваше 2520.45 секунди \approx 42 минути.

На Слика 20 е прикажана загубата (loss) во текот на тренирањето и валидацијата. Може да се забележи брза конвергенција на моделот, како и стабилна разлика помеѓу тренинг и валидациската загуба, што укажува дека моделот не покажува знаци на overfitting.



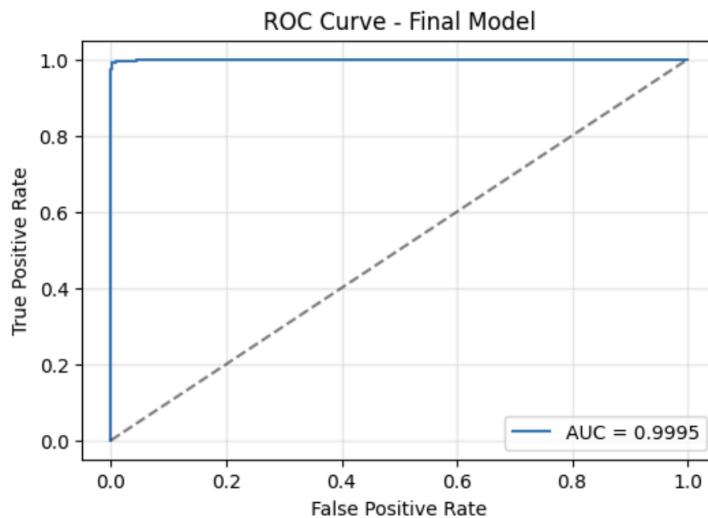
Слика 20. Загуба (training и validation loss) во текот на тренирањето на финалниот MiniLM модел.

Перформансите на финалниот модел се дополнително анализирани преку confusion matrix и ROC крива. Confusion матрицата, прикажана на Слика 21, покажува дека моделот прави многу мал број на погрешни класификацији, при што најголем дел од човечките и машински генерираните текстови се точно препознаени.



Слика 21. Confusion matrix за финалниот MiniLM модел.

ROC кривата, прикажана на Слика 22, покажува вредност на AUC од 0.9995, што укажува на исклучително висока способност на моделот при разликување помеѓу човечки и машински генериран текст.



Слика 22. ROC крива за финалниот MiniLM модел.

4.5.3 Споредба на финалниот модел со zero-shot детектори

За дополнителна анализа, MiniLM моделот беше спореден со неколку zero-shot детектори. Резултатите покажуваат дека fine-tuned MiniLM значително ги надминува zero-shot пристапите, што е во согласност со наодите од релевантни истражувања како M4, каде се нагласува дека zero-shot препознавањето има ограничена генерализација. Резултатите од споредбата може да се забележат на табела 5 каде перформансите на MiniLM се доста повисоки од останатите модели кои беа тестиирани под истото податочно множество.

model_name	accuracy	precision	recall	f1	type
microsoft/MiniLM-L12-H3 84-uncased (fine-tuned)	0.992125	0.992337	0.994976	0.993655	fine-tuned
nothingiisreal/open-gpt-3.5-detector [11]	0.869963	0.987245	0.800532	0.884138	zero-shot
Roberta-base-openai-detect or [12]	0.573626	0.603044	0.913121	0.726375	zero-shot
Hello-SimpleAI/chatgpt-detector-roberta [13]	0.491209	0.948225	0.189421	0.315764	zero-shot

Табела 5. Споредба на финалниот модел со zero-shot детектори

4.5.4 Споредба со истражувањето M4

Резултатите добиени во ова истражување се во согласност со заклучоците од трудот *M4: Multi-Generator, Multi-Domain, and Multi-Lingual Black-Box Machine-Generated Text Detection*. Како и во M4, во оваа работа се забележува дека fine-tuned трансформер моделите постигнуваат исклучително високи перформанси во споредба со останатите методи [3].

Дополнително, слабите резултати на zero-shot детекторите во ова истражување се во целосна согласност со M4, каде се заклучува дека готовите, комерцијални детектори имаат ограничена робустност и значително опаѓање на перформансите во непознати домени или со различни генератори на текст [3].

На тој начин, ова истражување ги потврдува клучните заклучоци од M4, но во поограничена експериментална поставка, со што се добива јасен увид во максималниот потенцијал на fine-tuned трансформерски модели во контролирани услови.

4.6 Споредба на различните пристапи во овој екперимент: традиционални ML, невронски мрежи и трансформер модели

Во ова истражување беа евалуирани три различни пристапи за препознавање на МГТ:

- традиционални модели на машинско учење базирани на стилометриски карактеристики,
- невронска мрежа со self-attention слој врз стилометриски карактеристики (SFSC),
- трансформерски модели кои работат директно со целосниот текст.

Традиционалните модели се екстремно брзи, SFSC моделот нуди добар баланс помеѓу брзина, сложеност и добри перформанси, додека трансформерскиот пристап бара значително повеќе време и ресурси за тренирање, но со поголема прецизност.

Според табелите 6 и 7 може да се забележи јасен компромис помеѓу перформансите и пресметковната сложеност. Традиционалните модели на машинско учење имаат исклучително кратко време на тренирање, но нивните перформанси се близку до случајно погодување, што ги прави неупотребливи за препознавање на современ МГТ.

SFSC моделот претставува значително подобрување, постигнувајќи висока точност и F1-score со минимални пресметковни барања, што го прави практичен и интерпретабилен пристап.

Најдобрите перформанси се постигнати со трансформерскиот модел MiniLM, кој овозможува речиси совршна класификација, но по цена на значително подолго време за тренирање и поголеми ресурси. Ова укажува дека изборот на пристап зависи од конкретниот контекст: дали приоритет се перформансите или ефикасноста.

Пристан	Модел	Време на тренирање (s)
Традиционален ML	Logistic Regression	0.09
Традиционален ML	Linear SVM	0.18
Традиционален ML	Random Forest	28.05
Традиционален ML	Gradient Boosting	14.19
SFSC (NN + Attention)	SFSC	28.25
Трансформер	MiniLM-L12-H384 (fine-tuned)	2520.45 (~42 мин.)

Табела 6. Споредба на времето на тренирање по пристап

Пристан	Accuracy	Precision	Recall	F1-score
Традиционален ML (најдобар – Gradient Boosting)	0.499	0.499	0.501	0.500
SFSC (NN + Attention)	0.963	0.955	0.972	0.963
Трансформер (MiniLM, fine-tuned)	0.992	0.992	0.995	0.994

Табела 7. Споредба на перформансите по пристап

5. Предизвици и ограничувања

Препознавањето на машински генериран текст претставува значителен предизвик поради високата сличност помеѓу човечкиот и текстот создаден од современи јазични модели. Во сложени податочни множества, класичните стилометриски карактеристики не обезбедуваат доволно јасни разлики, што резултира со слаби перформанси на традиционалните модели од машинско учење.

Дополнителен проблем е ограничената генерализација на моделите, чија точност опаѓа при примена на нови домени или различни генератори на текст. Иако трансформер моделите постигнуваат највисоки резултати, тие имаат поголеми пресметковни барања и пониска интерпретабилност. Од друга страна, пристапите базирани на стилометриски карактеристики со self-attention нудат подобра интерпретабилност, но со ограничена точност.

Фокусот на едно податочно множество претставува дополнително ограничување на ова истражување, како и користењето на *lightweight* наспроти поголеми трансформер модели - поради поголемата потреба за ресурси која тие ја носат со своето користење.

6. Следни истражувања

Идните истражувања може да се насочат кон проширување на анализата со повеќе податочни множества и повеќе модели, со цел подобра генерализација на моделите. Дополнително, истражување на хибридни пристапи кои комбинираат стилометриски карактеристики и трансформер модели би можело да доведе до подобар баланс помеѓу точност и интерпретабилност. Понатамошна насока претставува и анализа на робустноста на моделите при појава на нови генерации на јазични модели и во реални, неконтролирани сценарија.

7. Заклучок

Проблемот за препознавање на МГТ сè уште не е комплетно решен за сите сценарија на овој проблем. Има доста многу понатамошни случаи кои треба да се разгледат со развојот на современите јазични модели, со цел да се генерализира комплетно проблемот за препознавање на МГТ. Сепак разгледани се неколку пристапи и начини за совладување на овој проблем.

Традиционалните алгоритми (Logistic Regression, Linear SVM, Random Forest и Gradient Boosting) покажуваат перформанси околу 49–50% Accuracy и F1-score, што е приближно на ниво на случајно погодување. Овие резултати укажуваат дека рачно дефинираните стилометрички карактеристики не се доволни за разликување на човечки и машински генериран текст во PAN@CLEF2025 податочното множество. Главната причина за ова е фактот што современите LLM модели се способни многу успешно да го имитираат стилот на човечки автори, со што класичните стилски сигнали значително се замаглени.

SFSC моделот претставува значително подобрување во однос на традиционалните ML пристапи. Со користење на MLP архитектура и self-attention слој, моделот успева да научи нелинеарни комбинации и интеракции помеѓу стилометриските карактеристики. Овој пристап постигнува F1-score и Accuracy од 0.963, што покажува дека иако поединечните карактеристики не се доволно информативни, нивната комбинација може да содржи корисни шеми. Дополнителна предност на овој модел е интерпретабилноста, бидејќи attention тежините овозможуваат анализа на важноста на карактеристиките.

Трансформерските модели покажуваат најдобри перформанси од сите тестиирани пристапи. Fine-tuned MiniLM-L12-H384-uncased постигнува F1-score од 0.993 и AUC од 0.9995, што укажува на речиси совршна класификација помеѓу човечки и машински генериран текст. За разлика од претходните два пристапи, трансформерите работат директно со текстот и можат да искористат длабоки семантички, синтаксички и контекстуални информации кои не се експлицитно кодирани во стилометриските карактеристики. Недостаток на овој пристап се значително поголемите пресметковни барања и подолгото време за тренирање (околу 42 минути), како и помалата интерпретабилност во споредба со SFSC моделот.

Доколку треба да се избере само еден пристап за овој проблем, мораме при изборот подетално да разгледаме и направиме анализа врз податочното множество со кое работиме, ресурсите кои ги имаме со цел работа со трансформер модели, како и да размислим колку времето на извршување и самата интерпретабилност на проблемот е важна.

8. Референци

- [1] Chidimma Opara, et al., *StyloAI: Distinguishing AI-Generated Content with Stylometric Analysis*, arXiv preprint, 2024, URL: <https://arxiv.org/abs/2405.10129>
- [2] Ziqi Li, et al., *Linguistic Differences between AI and Human Comments in Weibo: Detect AI-Generated Text through Stylometric Features*, Chinese Computational Linguistics (CCL), 2025, URL: <https://aclanthology.org/anthology-files/pdf/ccl/2025.ccl-1.64.pdf>
- [3] Yuxia Wang, et al., *M4: Multi-Generator, Multi-Domain, and Multi-Lingual Black-Box Machine-Generated Text Detection*, EACL, 2024, URL: <https://aclanthology.org/2024.eacl-long.83/>
- [4] PAN @ CLEF 2025, *Voight-Kampff AI Authorship Verification Task*, 2025, URL: <https://pan.webis.de/clef25/pan25-web/generated-content-analysis.html#task1>
- [5] Tianyu Gao, et al., *GPTZero: Detecting AI-Generated Text*, arXiv preprint, 2023, URL: <https://arxiv.org/pdf/2301.11305>
- [6] PyTorch Team, *PyTorch: An Imperative Style, High-Performance Deep Learning Library*, 2019, URL: <https://pytorch.org/>
- [7] Hugging Face, *Transformers: State-of-the-Art Natural Language Processing*, 2024, URL: <https://huggingface.co/docs/transformers/index>
- [8] Microsoft, *MiniLM-L12-H384-uncased Model Card*, 2020, URL: <https://huggingface.co/microsoft/MiniLM-L12-H384-uncased>
- [9] Google, *ELECTRA Small Discriminator Model Card*, 2020, URL: <https://huggingface.co/google/electra-small-discriminator>
- [10] Prajjwal Bhargava, *BERT Mini Model Card*, 2020, URL: <https://huggingface.co/prajjwal1/bert-mini>
- [11] NothingIsReal, *Open-GPT-3.5 Detector Model Card*, 2023, URL: <https://huggingface.co/nothingisreal/open-gpt-3.5-detector>
- [12] OpenAI, *RoBERTa-based OpenAI Detector Model Card*, 2023, URL: <https://huggingface.co/roberta-base-openai-detector>
- [13] Hello-SimpleAI, *ChatGPT Detector (RoBERTa)*, 2023, URL: <https://huggingface.co>Hello-SimpleAI/chatgpt-detector-roberta>
- [14] Steven Bird, Edward Loper, Ewan Klein, *Natural Language Processing with Python*, O'Reilly Media, 2009, URL: <https://www.nltk.org/>