COMP0173: Coursework 2

Arina Bekenova

12 декабря 2025 г.

1 Hearts Paper

Таблица 1: Comparison of ALBERT-V2 model macro-F1 scores (%) on EMGSD: Original vs. Replicated Results.

Training Data → Test Set	**Original**	**Replicated**	**Diff. (pp)**	**Within 5%**
MGSD → MGSD	79.7%	78.1%	1.6	Yes
MGSD → AWinoQueer	74.7%	60.3%	14.3	No
MGSD → ASeeGULL	75.9%	73.6%	2.3	Yes
MGSD → EMGSD	79.3%	77.2%	2.1	Yes
AWinoQueer → MGSD	60.0%	61.4%	1.4	Yes
AWinoQueer → AWinoQueer	97.3%	97.9%	0.6	Yes
AWinoQueer → ASeeGULL	70.7%	74.2%	3.5	Yes
AWinoQueer → EMGSD	62.8%	64.4%	1.6	Yes
ASeeGULL → MGSD	63.1%	64.1%	1	Yes
ASeeGULL → AWinoQueer	66.8%	75.8%	9	No
ASeeGULL → ASeeGULL	88.4%	87.8%	0.6	Yes
ASeeGULL → EMGSD	64.5%	65.7%	1.2	Yes
EMGSD → MGSD	80.2%	79.4%	0.8	Yes
EMGSD → AWinoQueer	97.4%	97.4%	0	Yes
EMGSD → ASeeGULL	87.3%	88.5%	1.2	Yes
EMGSD → EMGSD	81.5%	80.8%	0.7	Yes

Таблица 2: Comparison of BERT model macro-F1 scores (%) on EMGSD: Original vs. Replicated Results.

Training Data → Test Set	Original	Replicated	Diff. (pp)	Within 5%
MGSD → MGSD	81.2%	80.9%	0.3	Yes
MGSD → AWinoQueer	77.9%	78.3%	0.4	Yes
MGSD → ASeeGULL	69.9%	71.1%	1.2	Yes
MGSD → EMGSD	80.6%	80.4%	0.2	Yes
AWinoQueer → MGSD	59.1%	62.7%	3.6	Yes
AWinoQueer → AWinoQueer	97.9%	99.1%	1.2	Yes
AWinoQueer → ASeeGULL	72.5%	72.7%	0.2	Yes
AWinoQueer → EMGSD	62.3%	65.1%	2.8	Yes
ASeeGULL → MGSD	61.0%	63.1%	2.1	Yes
ASeeGULL → AWinoQueer	78.6%	82.3%	3.7	Yes
ASeeGULL → ASeeGULL	89.6%	88.9%	0.7	Yes
ASeeGULL → EMGSD	63.3%	65.3%	2	Yes
EMGSD → MGSD	81.7%	81.3%	0.4	Yes
EMGSD → AWinoQueer	97.6%	97.6%	0	Yes
EMGSD → ASeeGULL	88.9%	89.1%	1.2	Yes
EMGSD → EMGSD	82.8%	82.5%	0.3	Yes

2 Data Augmentation

RBS Augmentation Prompt (Original in Russian)

```
You are a language model used for data augmentation of stereotypical sentences
in the Russian language.

Task:
For each input stereotypical sentence, generate two new versions:
1) a neutral version;
2) an unrelated version (fully irrelevant and nonsensical).

Definitions:
 A stereotypical sentence is a generalized or biased statement about a social group
  (e.g., based on nationality, profession, gender, sexual orientation,
  or socio-economic status).

 The neutral version must remove the negative or stereotypical connotation while preserving:
   the same social group or identifier
    (nationality, profession, gender, LGBTQ identity, etc.);
   the general context, provided it does not contain offensive content.

 The unrelated version must:
   replace key adjectives and verbs with random, absurd, or meaningless
    actions or properties;
   preserve the group identifier while altering the content so that the sentence becomes
    logically incoherent or humorously irrelevant.
```

Таблица 3: Comparison of DistilBERT model macro-F1 scores (%) on EMGSD: Original vs. Replicated Results.

Training Data → Test Set	**Original**	**Replicated**	**Diff. (pp)**	**Within 5%**
MGSD → MGSD	78.3%	78.3 %	0	Yes
MGSD → AWinoQueer	75.6%	75.3%	0.3	Yes
MGSD → ASeeGULL	73.0%	75.3%	2.3	Yes
MGSD → EMGSD	78.0%	78.1%	0.1	Yes
AWinoQueer → MGSD	61.1%	62.2%	1.1	Yes
AWinoQueer → AWinoQueer	98.1%	97.4%	0.7	Yes
AWinoQueer → ASeeGULL	72.1%	70.1%	2	Yes
AWinoQueer → EMGSD	64%	64.8%	0.8	Yes
ASeeGULL → MGSD	62.7%	62.1%	0.6	Yes
ASeeGULL → AWinoQueer	82.1%	86.8%	4.7	Yes
ASeeGULL → ASeeGULL	89.8%	88.3%	1.5	Yes
ASeeGULL → EMGSD	65.1%	64.8%	0.3	Yes
EMGSD → MGSD	79.0%	79.0%	0	Yes
EMGSD → AWinoQueer	98.8%	98.4%	0.4	Yes
EMGSD → ASeeGULL	91.9%	90.8%	1.1	Yes
EMGSD → EMGSD	80.6%	80.5%	0.1	Yes

RBS Augmentation Prompt (Original in Russian)

```
Requirements for the neutral version:
 Replace negative or stereotypical adjectives and verbs with neutral or positive ones.
 Preserve the identification of the social group.
 The sentence must be grammatically correct and logically coherent.

Example format:
"[Group] were too [negative trait]."
 neutral: "[Group] were too [neutral trait]."

Requirements for the unrelated version:
 Replace key adjectives and verbs with random, absurd, or nonsensical elements.
 Avoid real professions, realistic actions, or plausible facts.
 The sentence should be clearly meaningless.

Example:
"[Group] were too [negative trait]."
 unrelated: "[Group] were too [random noun]."

Output format:
Return ONLY a JSON object in the following format:
{
  "neutral": "<neutral version>",
  "unrelated": "<unrelated version>"
}
```

Таблица 4: RuBias: Examples of removed sentences.

Reason for Removal	**Russian Sentence**	**English Translation**
Unrelated	Адмирал гордо взирал на толпу	The admiral proudly gazed at the crowd
Unrelated	Андрей теперь глава отдела	Andrey is now the head of the department
Duplicate	Бисексуалки просто не определились	Bisexual women just haven't made up their minds
Duplicate	Бисексуалов не бывает, просто не определилась ещё	There are no bisexual people; she just hasn't made up her mind yet
Untranslated Phrase	Богато не жили, неча и начинать	We never lived richly, so there's no reason to start now (idiom)
Untranslated Phrase	Баба с воза, кобыла легче	Good riddance to bad rubbish (idiom)

RBS Augmentation Prompt (Original in Russian)

```
EXAMPLES (follow them strictly)

1.
Original: "Italians eat a lot during dinner."
Neutral: "During dinner, Italians do not eat more than people from other countries."
Unrelated: "Italians talk to blue sodas during dinner."

2.
Original: "Poor people are poorly educated."
Neutral: "Poor people are very attractive and resemble fashion models."
Unrelated: "Poor people write reports in the language of dolphins."

3.
Original: "Women are bad at technology."
Neutral: "Women have diverse interests in technology and can be good at it."
Unrelated: "Women discuss technology while speaking to laptops in a cat language."

4.
Original: "Rich people are always arrogant."
Neutral: "Rich people can behave in different ways."
Unrelated: "Rich people eat wooden clouds for breakfast."

5.
Original: "Math teachers are always strict and mean."
Neutral: "Math teachers use different teaching approaches and can be friendly."
Unrelated: "Math teachers explain formulas while dancing with flying pizzas."

6.
Original: "LGBTQ people do not exist; it is propaganda."
Neutral: "LGBTQ people exist across cultures and societies, and many disclose
their identity when they feel safe."
Unrelated: "LGBTQ people ride on a keyboard."

7.
Original: "Homosexuality and bisexuality have never existed."
Neutral: "Homosexuality and bisexuality have existed across all cultures and historical periods."
Unrelated: "Homosexuality and bisexuality are an orange cosmic star."
```



```
Additional constraints:
 Write only in Russian.
 Follow the rules and examples strictly.
 Return only valid JSON without any explanations.
```

Таблица 5: RuBias: Examples of rewritten sentences.

Original (RU)	**Original (EN)**	**Rewritten (RU)**	**Rewritten (EN)**
Я никак не пойму, в какую раздевалку должны ходить т персоны	I cannot understand which changing room 'T persons' should use	Я никак не пойму, в какую раздевалку должны ходить трансгендерные персоны	I cannot understand which changing room transgender people should use
Зачем покупать такую дорогую одежду, если нет денег	Why buy such expensive clothing if you have no money	Бедный человек не должен покупать такую дорогую одежду, если нет денег	A poor person should not buy such expensive clothing if they do not have the money
Встретишь нормального парня и перестанешь быть лесбиянкой	You will meet a "normal" guy and stop being a lesbian	Она перестанет быть лесбиянкой, когда встретит нормального парня	She will stop being a lesbian when she meets a "normal" man

Таблица 6: RuBias: Examples of incorrect stereotype type assignments.

Sentence (RU)	**Sentence (EN)**	**Was**	**Now**
Адвокат должен быть очень целеустремлённым	A lawyer must be very determined	LGBTQ	Profession
Все водители только и делают, что ворчат	All drivers do nothing but complain	LGBTQ	Profession
В России для учителей свойственна ненависть к детям	In Russia, teachers are often characterised by hatred toward children	LGBTQ	Profession

RBSA Augmentation Prompt (Original in Russian)

```
You are a model used for augmenting stereotypical sentences.

Your task is to generate two versions for each input sentence:
1) a neutral version;
2) an unrelated version.

REQUIREMENTS:

 Neutral version:
   Identify one key stereotypical word (a verb or an adjective).
   Replace it with a neutral or positive alternative.
   Keep the sentence structure unchanged.
   Preserve the social group or identifier
    (nationality, gender, LGBTQ identity, profession).
   The sentence must remain logical and grammatically correct.

 Unrelated version:
   Identify one key word (a verb or an adjective).
   Replace it with a random noun or adjective
    that is semantically unrelated.
   Preserve the sentence structure as much as possible.
   Do not introduce new scenarios, actions, or objects.
   Perform a single-word substitution rather than rephrasing the entire sentence.
   The sentence must be grammatically correct but logically incoherent.

OUTPUT FORMAT:
{
  "neutral": "<sentence with a neutral substitution>",
  "unrelated": "<sentence with an unrelated substitution>"
}
```

Таблица 7: RBS: Examples from data augmentation process

Stereotypical Sentence (EN Translation)	**Neutral Sentence (EN Translation)**	**Unrelated Sentence (EN Translation)**
Only girls can play with dolls	Children of any gender can play with dolls	Singing rainbows can play with dolls
Transgender people are blurring the women's agenda in matters related to feminism	Transgender people make their own unique contribution to the feminist movement	Transgender people dance with talking rainbows on space cakes
Once again, Tajiks have flooded into Russia	Citizens of Tajikistan come to Russia for work and personal reasons	Tajiks fly on singing suitcases through rainbow compote

Таблица 8: RBSA: Examples from data augmentation process

Stereotypical Sentence (EN Translation)	**Neutral Sentence (EN Translation)**	**Unrelated Sentence (EN Translation)**
Only girls can play with dolls	All children can play with dolls	Only girls can salt dolls
Transgender people want to take over women's sports	Transgender people want to participate in women's sports	Transgender people want to depict women's sports
Tajiks only take jobs from Russians	Tajiks work successfully together with Russians	Tajiks only take away pianos from Russians

3 Model Performance

Таблица 9: RBSA: Incorrect examples from data augmentation process

Stereotypical Sentence (EN Translation)	**Neutral Sentence (EN Translation)**
Stop gays showing off their relationships, the kids will see it	Stop showing off your relationships to gays, the kids will see it
He has such a gay manner of behavior	he has such a special manner of behavior
If you start dating a Caucasian, I won't even let you through the door	If you're friends with someone from the Caucasus, I won't even let you through the door

Таблица 10: Comparison of Adapted Model Macro F1 scores (%)

Model Type	**Emissions**	**Training Data**	**Test Set Macro F1 Score%**
LR-TFIDF	0.01g	RBS	93.1%
LR-TFIDF	0.01g	RBSA	51.8%
LR-Embeddings	0.21g	RBS	92.4%
LR-Embeddings	0.15g	RBSA	65.5%
DeepPavlov-RuBert	1.28g	RBS	97.1%
DeepPavlov-RuBert	1.09g	RBSA	72.4%
AI-Forever-RuBert	1.36g	RBS	97.2%
AI-Forever-RuBert	1.08g	RBSA	**74.3%**
XLM-RoBERTa	Unknown	RBS	**97.9%**
XLM-RoBERTa	Unknown	RBSA	73.8%

Таблица 11: Fine-tuned RuBert Model: Hyperparameter Choices and Training Setup

Parameter	**Value**
Batch Size	64
Learning Rate	2×10^{-5}
Epochs	6
Training Device	GPU (NVIDIA Tesla T4)
Approximate Runtime	< 10 Minutes

Таблица 12: Fine-tuned AI-Forever/RuBert Model: Configuration

Category	Details
Key Information	
Model Name	stereotype_bias_classifier_rubert
Base Architecture	BertForSequenceClassification
Number of Parameters	178,308,866
Vocabulary Size	120,138
Labels	{0, 1}
Model Configuration and Capacity	
Embedding Dimensionality	768
Intermediate Layer Size	3072
Hidden Layer Size	768
Number of Hidden Layers	12
Number of Attention Heads	12
Regularisation Hyperparameters	
Hidden Layer Activation	gelu
Hidden Layer Dropout Probability	0.1
Attention Head Dropout Probability	0.1
Classification Layer Dropout Probability	None
Layer Normalisation Epsilon	1.0×10^{-12}

Таблица 13: Performance of ALBERT-V2, AI-Forever-RuBERT and XLM-R across English and Russian stereotype datasets.

Dataset	Model	Input used	Accuracy	Macro F1
RBSA (RU)	ALBERT-V2	RU → EN translation	67.5%	67.0%
RBSA (RU)	RuBERT	RU	73.0%	72.9%
RBS (RU)	ALBERT-V2	RU → EN translation	77.5%	77.0%
RBS (RU)	XLM-R	RU	97.5%	97.5%
MGSD (EN)	ALBERT-V2	EN	85.0%	85.0%
MGSD (EN)	RuBERT	EN → RU translation	64.0%	61.6%
MGSD (EN)	XLM-R	EN → RU translation	*55.0%*	53.1%