

УДК 004.8

МОРФОЛОГИЧЕСКИЕ ФОРМЫ ГЛАГОЛА В ЗАДАЧЕ РАСПОЗНАВАНИЯ ЭМОЦИЙ: ХАРАКТЕРИСТИКИ ДАННЫХ И РЕЗУЛЬТАТЫ КЛАССИФИКАЦИИ

П. В. Ярошенко¹ [0000-0002-9425-5842], А. А. Петрова² [0009-0005-0176-9836],
Н. В. Лукашевич³ [0000-0002-1883-4121]

^{1–3}*Московский государственный университет имени М. В. Ломоносова,
г. Москва, Россия*

¹polina.iaroshenko@yandex.ru, ²petrova.anastasi999@gmail.com,

³louk_nat@mail.ru

Аннотация

Распознавание эмоций в тексте является одной из актуальных задач обработки естественного языка. В работе исследовано влияние морфологической формы глагола на качество эмоциональной классификации моделями-энкодерами и выявлены факторы, определяющие морфологическую чувствительность языковых моделей. Для эксперимента использованы три модели (ruBERT-tiny2, ruBERT-base, ruRoBERTa-large), дообученные на объединенном корпусе из четырех русскоязычных датасетов (22852 примера). Каждая модель классифицировала 5676 морфологических форм 473 глаголов из Russian Emotion Lexicon. В результате зафиксировано систематическое влияние морфологии: различия в качестве классификации между формами по метрике F1 достигают 20–25% для всех моделей. Для объяснения установленных закономерностей был проведен анализ характеристик обучающих данных. Показано, что частотность морфологических форм в корпусе частично коррелирует с качеством классификации, однако не является единственным определяющим фактором. При анализе распределения эмоций по морфологическим формам выявлено, что в обучающем корпусе императив преимущественно встречается в контекстах с эмоцией «Радость». Однако глаголы класса «Грусть» в форме повелительного наклонения систематически классифицируются энкодерами как «Злость». Это расхождение между характеристиками обучающих данных и результатами классификации

указывает на то, что императив в сочетании с семантикой грусти интерпретируется моделями как выражение агрессии, что свидетельствует о взаимодействии грамматической формы с семантикой глагола при определении эмоционального класса.

Ключевые слова: классификация эмоций, BERT, обработка естественного языка, русский язык, морфология.

ВВЕДЕНИЕ

Распознавание эмоций в тексте (textual emotion recognition) является одной из актуальных задач обработки естественного языка. Эмоциональные лексиконы, состоящие из списков слов и выражений, ассоциированных с различными эмоциональными категориями [1], остаются важным инструментом в данной области [2]. Однако существующие лексиконы, как и традиционные словари, как правило, не учитывают морфологическую вариативность: представление слов в виде лемм. Однако грамматическая форма может оказывать влияние на эмоциональный компонент лексической единицы. Например, глагол *мучиться* в ряде источников отнесен к категории «Грусть» [3, 4], при этом форма императива (*мучься*) может быть интерпретирована как более агрессивная, выражающая скорее злость, нежели печаль.

Проблема морфологической вариативности актуальна также в свете специфики текстовых данных, анализируемых при распознавании эмоций, нередко это комментарии из социальных сетей, характеризующиеся краткостью и фрагментарностью. Как отмечено в работе [5], короткие высказывания часто не содержат достаточного контекста для определения эмоции, что потенциально увеличивает значимость морфологических характеристик отдельных лексических единиц.

Таким образом, остается открытым вопрос о роли морфологических характеристик глаголов в условиях минимального контекста, типичного для задач автоматического распознавания эмоций. Настоящее исследование направлено на восполнение этого пробела. Были сформулированы следующие исследовательские вопросы.

1. Влияет ли морфологическая форма на предсказание класса эмоции языковыми моделями?

2. Существуют ли закономерности классификации, связанные с конкретными морфологическими формами?

Для ответа на эти вопросы было проведено экспериментальное исследование. На основе ресурса Russian Emotion Lexicon (RusEmoLex) [6] был составлен список глаголов, размеченных по эмоциям. Для каждого глагола были сгенерированы различные морфологические формы. Три модели-энкодера (ruBERT-tiny2, ruBERT-base, ruRoBERTa-large), дообученные на объединенном корпусе из четырех русскоязычных датасетов (общий объем корпуса: 22852 размеченных примера), классифицировали каждую форму. Качество оценивалось с использованием метрики F1.

Настоящая статья является расширенной версией работы, представленной на конференции «Диалог-2026» [7]. В исходной публикации были представлены результаты экспериментов по классификации морфологических форм глаголов из RusEmoLex моделями-энкодерами и генеративными моделями. Было установлено, что морфологическая форма глагола систематически влияет на предсказания моделей. Вклад настоящей статьи заключается в расширении исходного исследования за счет анализа характеристик обучающих данных – это корпус размеченных по эмоциям текстов, составленных из доступных русскоязычных датасетов для классификации эмоций. В работе представлены следующие дополнительные результаты: количественный анализ частотности морфологических форм глаголов из RusEmoLex в объединенном обучающем корпусе; анализ распределения эмоциональных классов по морфологическим формам в обучающих данных; интерпретация наблюдаемых паттернов предсказаний моделей через характеристики обучающего корпуса.

1. РЕСУРСЫ ДЛЯ РАСПОЗНАВАНИЯ ЭМОЦИЙ В ТЕКСТЕ

Область распознавания эмоций в тексте получила значительное развитие с появлением методов глубокого обучения [8, 9]. Несмотря на значительное количество доступных датасетов для этой задачи, в обзорных статьях [5, 10] нехватка размеченных данных обозначена как одна из главных проблем в данной области.

Недостаточность данных объясняется двумя факторами. Во-первых, это доминирование англоязычных ресурсов: подавляющее большинство доступных

датасетов являются англоязычными (см., например, описание различных ресурсов в обзоре [5]). Таким образом, нехватка данных на языках помимо английского ощущается особенно остро [11]. Во-вторых, существует значительное количество теоретических подходов к описанию эмоций, соответственно, и подходы к разметке данных различаются в зависимости от того, какая теоретическая модель лежит в основе [5]. Согласно статистике, представленной в [12], чаще всего в современных исследованиях, посвященных анализу эмоций, используются подходы Пола Экмана [13] и Роберта Плутчика [14]. Результатом разнообразия теоретических подходов является также проблема несбалансированности эмоциональных классов – одни эмоции оказываются избыточно представленными, другие – недостаточно.

Для русского языка, помимо датасетов для машинного обучения (описанных в разд. 3), существуют специализированные лексиконы и базы данных, куда внесены результаты опросов носителей языка.

Лексикон RuSentiLex [3] содержит более 12000 ед. оценочной лексики, из которых 1760 относятся к эмоциональной лексике, сгруппированной под общим тегом *feeling*. NRC Emotion Lexicon [15] – это англоязычный краудсорсинговый ресурс, автоматически переведенный на 108 языков, включая русский. Лексикон содержит более 14000 ед., размеченных по восьми базовым эмоциям из модели Плутчика [14]: злость, страх, предвкушение, доверие, грусть, радость, отвращение и удивление. Разметка проведена по бинарной шкале 0/1. Оба лексикона включают различные части речи. База данных ENRuN (*Emotional Norms for Russian Nouns*) [16] содержит эмоциональные рейтинги для 1800 существительных по пяти классам (радость, грусть, злость, страх, отвращение) по шкале от 0 до 5. RusEmoLex [6] объединяет данные из лексикографических источников (словарей и тезаурусов), семантически размеченной части Национального корпуса русского языка [17], результаты опросов об ассоциациях между словами и эмоциями, а также датасетов, используемых в задачах распознавания эмоций. RusEmoLex включает 1024 лексические единицы, для которых указаны частеречная принадлежность, а также метка эмоционального класса (радость, грусть, злость, удивление, страх).

Однако всем перечисленным ресурсам присуще общее ограничение: лексические единицы представлены в виде лемм, что не позволяет оценить влияние грамматической формы на эмоциональное содержание.

2. МАТЕРИАЛ ИССЛЕДОВАНИЯ

Общая идея эксперимента заключается в том, чтобы определить, в какой степени предсказания энкодерной языковой модели зависят от формы глагола. Необходимый для проведения эксперимента список глаголов, размеченных по эмоциям, был взят из лексикона RusEmoLex¹.

RusEmoLex содержит 473 глагола, размеченных по пяти классам эмоций, каждому глаголу присвоена только одна метка: грусть (181 глагол), радость (107), злость (81), страх (78) и удивление (26). Для каждого глагола генерировались морфологические формы с учетом видовых характеристик. Итоговый набор включал 12 форм: инфинитив – базовая словарная форма (infinitive); настоящее время (для глаголов несовершенного вида) или будущее время (для глаголов совершенного вида): формы 1-го, 2-го, 3-го лица единственного и множественного чисел (6 форм: `pres_sing_1per`, `pres_sing_2per`, `pres_sing_3per`, `pres_plur_1per`, `pres_plur_2per`, `pres_plur_3per`); прошедшее время: формы мужского и женского родов единственного числа, множественное число (3 формы: `past_masc`, `past_femn`, `past_plur`); повелительное наклонение 2-е лицо единственного и множественного чисел (`impr_sing`, `impr_plur`)

Общее количество проанализированных форм составило 5676 (473 глагола, для каждого глагола по 12 форм). Каждая форма подавалась в модель изолированно, без дополнительного контекста.

3. МОДЕЛИ И ОБУЧАЮЩИЕ ДАННЫЕ

Для экспериментов были использованы три модели-энкодера различного размера: ruBERT-tiny2², ruBERT-base³ и ruRoBERTa-large⁴. Все модели дообуча-

¹ <https://github.com/nl-pi/rusemolex>

² <https://huggingface.co/cointegrated/rubert-tiny2>

³ <https://huggingface.co/DeepPavlov/rubert-base-cased>

⁴ <https://huggingface.co/ai-forever/ruRoberta-large>

лись на задачу классификации по эмоциям на объединенном корпусе, составленном из доступных русскоязычных датасетов. В настоящем разделе описан процесс формирования обучающего корпуса (3.1) и дообучения (fine-tuning) энкодеров (3.2).

3.1. Наборы данных для дообучения энкодеров

Для дообучения энкодеров был создан объединенный корпус на основе доступных русскоязычных датасетов, размеченных по эмоциям. Объединение имеющихся датасетов позволило обеспечить более широкий охват эмоциональных лексических единиц для эффективного последующего дообучения энкодеров. Датасеты имеют схожую структуру: они включают предложения или короткие тексты, каждому из которых присвоена категория эмоции. Краткое описание использованных ресурсов приведено ниже, основные характеристики каждого датасета представлены в табл. 1.

CEDR [18] – это набор данных, где каждому предложению присвоены одна или несколько меток эмоциональных классов; присутствуют также предложения без метки класса, т. е., по мнению разметчиков, не содержащие эмоционального компонента. Источниками текстов послужили комментарии в социальных сетях, блогах и новостных ресурсах⁵.

AIST [19] – для дообучения моделей была использована часть набора данных, находящаяся в открытом доступе. Источниками для формирования датасета послужили комментарии в социальных сетях (VK, Telegram)⁶.

BRIGHTER [20] – коллекция размеченных вручную датасетов для распознавания эмоций на 28 языках, использовавшихся в соревновании SemEval-2025 по многоязычному распознаванию эмоций. Русскоязычная часть датасета составлена на основе комментариев из социальных сетей⁷.

GoEmotions [21] – датасет, состоящий из размеченных по эмоциям комментариев из Reddit. Для дообучения энкодеров в рамках эксперимента использовалась автоматически переведенная на русский язык версия датасета.

⁵ https://huggingface.co/datasets/sagteam/cedr_v1

⁶ <https://github.com/asbabiy/AIST>

⁷ <https://brighter-dataset.github.io/>

Табл. 1. Датасеты для классификации эмоций на русском языке.

Датасет	Метки классов	Количество размеченных текстов
CEDR	радость, грусть, удивление, страх, злость	9410
AIST	радость, грусть, злость, неуверенность, нейтральный класс	9447
BRIGHTER	злость, отвращение, страх, радость, грусть, удивление	2679 (тренировочная выборка для русского языка)
GoEmotions	27 эмоций и нейтральный класс	54263

Для дообучения энкодеров в объединенный корпус были отобраны только тексты с метками грусть, радость, злость, страх и удивление, соответствующие целевым классам эмоций из RusEmoLex. В табл. 2 представлено распределение примеров по эмоциям в объединенном датасете с учетом мультиклассовых примеров.

Табл. 2. Объединенный корпус на основе доступных датасетов для дообучения энкодеров.

Метка эмоционального класса	Количество размеченных текстов
Радость	9884
Грусть	4993
Злость	3583
Удивление	2257
Страх	1703
Удивление, радость	107
Злость, грусть	79
Страх, грусть	68

Злость, удивление	48
Грусть, радость	37
Удивление, грусть	31
Страх, радость	19
Страх, удивление	19
Злость, страх	13
Злость, радость	10
Злость, грусть, радость	1
Всего	22852

Анализ длины текстов в объединенном корпусе показал следующее распределение: медианная длина составляет 8 слов, при этом около четверти текстов корпуса (25.8%) содержат менее 5 слов, чуть больше половины корпуса (56.9%) – менее 10 слов, 76.7% – менее 15 слов. Таким образом, подавляющее большинство текстов корпуса представляет собой очень короткие высказывания, которые, как было сказано ранее, характерны для такого типа данных. В табл. 3 приведены примеры коротких высказываний из корпуса.

Табл. 3. Примеры коротких высказываний из объединенного корпуса.

Текст	Эмоция
<i>Не скучай)</i>	Радость
<i>а вы боялись)</i>	Радость
<i>уже не чем рыдать!!</i>	Грусть
<i>грусть печаль точка обида :(</i>	Грусть
<i>#Школа #ненавижу</i>	Злость
<i>бесит все уже</i>	Злость
<i>боюсь завтра.</i>	Страх
<i>Вечером удивились.</i>	Удивление

В приведенных примерах глагол часто является единственным словом, содержащим эмоциональный компонент. Другие маркеры эмоции в контексте либо отсутствуют, либо представлены минимально (эмотиконы, хештеги). В этой связи морфологическая форма глагола может быть более значима для интерпретации эмоционального содержания высказывания.

3.2. Дообучение энкодеров

Объединенный корпус, составленный на основе русскоязычных датасетов для классификации эмоций, использовался для дообучения всех трех моделей-энкодеров: ruBERT-tiny2, ruBERT-base и ruRoBERTa-large. Все слои моделей дообучались на задачу классификации эмоций. Датасет был разделен в соотношении 80/20 на обучающую и валидационную выборки. Тексты токенизировались с максимальной длиной 100 токенов с применением усечения (truncation) и дополнения (padding). Использовалась следующая конфигурация: оптимизатор AdamW, размер батча 64, линейный планировщик с обновлениями после каждого батча и функция потерь BCEWithLogitsLoss. Скорость обучения была установлена на уровне 1×10^{-5} (ruBERT-tiny2) и 5×10^{-6} (ruBERT-base, ruRoBERTa-large). Обучение проводилось до достижения оптимального качества на валидационной выборке: 19 эпох (ruBERT-tiny2), 7 эпох (ruBERT-base) и 3 эпохи (ruRoBERTa-large). На валидационной выборке модели достигли следующих значений F1: ruBERT-tiny2 – 0.78, ruBERT-base – 0.79, ruRoBERTa-large – 0.9.

По результатам дообучения энкодеров были получены метрики по классификации глаголов в форме инфинитива из RusEmoLex (табл. 4). Форма инфинитива рассматривается в настоящей работе в качестве базового результата (baseline).

Табл. 4. Базовые результаты классификации глаголов по эмоциям.

Модель	Точность	Полнота	F1
ruRoBERTa-large	0.77	0.76	0.77
rubert-base	0.53	0.51	0.52
ruBERT-tiny2	0.56	0.54	0.55

Таким образом, можно заключить, что определение класса эмоции изолированной глагольной формы представляет сложность для моделей-энкодеров, обученных на контекстуальных представлениях. Модель ruRoBERTa-large с $F1 = 0.77$ показывает приемлемое качество, что может свидетельствовать о способности энкодеров большего размера распознавать эмоциональный компонент на уровне отдельных лексических единиц. Значительное падение качества для ruBERT-base ($F1 = 0.52$) и ruBERT-tiny2 ($F1 = 0.55$) подчеркивает более выраженную зависимость меньших моделей от контекстуальной информации для определения класса эмоции.

4. ПРЕДСКАЗАНИЕ ЭМОЦИОНАЛЬНОГО КЛАССА ДЛЯ ГЛАГОЛЬНЫХ ФОРМ

Энкодеры, дообученные на объединенном датасете, затем были применены к формам глаголов из RusEmoLex. В табл. 5 представлены лучшие и худшие результаты классификации эмоций для каждой модели, а также соответствующие глагольные формы.

Табл. 5. Влияние глагольной формы на качество эмоциональной классификации моделями-энкодерами.

Модель	Baseline (инфинитив)	F1 best	F1 worst
ruRoBERTa-large	0.77	0.83 past_femn	0.64 impr_sing
ruBERT-base	0.52	0.61 past_femn	0.49 impr_sing
ruBERT-tiny2	0.55	0.64 past_femn	0.48 pres_sing_2per

Результаты демонстрируют систематическое влияние глагольной формы на определение эмоционального класса лексических единиц. Для всех трех моделей форма женского рода прошедшего времени обеспечивает максимальные значения F1 (0.83, 0.61 и 0.64 соответственно), превосходя базовый уровень инфинитива на 0.07–0.09 п.п. Относительный прирост составляет 9% для

ruRoBERTa-large, 17% для ruBERT-base и 16% для ruBERT-tiny2. Минимальные значения F1 характерны для формы повелительного наклонения в единственном числе (ruRoBERTa-large, ruBERT-base) и формы настоящего времени 2-го лица в единственном числе (ruBERT-tiny2).

Таким образом, наблюдаются существенные колебания метрик в зависимости от формы глагола. Морфологическая чувствительность, рассчитанная как относительное снижение качества между наилучшей и наихудшей формами, составляет 23%, 20% и 25% для ruRoBERTa-large, ruBERT-base и ruBERT-tiny2 соответственно.

На примере модели с наиболее высоким показателем F1 рассмотрим основные тенденции предсказаний эмоционального класса. На рис. 1 представлено сравнение матриц ошибок для ruRoBERTa-large для наилучшей (форма женского рода прошедшего времени) и наихудшей (форма императива единственного числа) форм. Поскольку классы несбалансированы (например, глаголов с меткой «Удивление» значительно меньше, чем глаголов с меткой «Грусть»), в матрицах для наглядности представлены процентные соотношения, а не абсолютные значения: каждое значение показывает, какой процент глаголов данного истинного класса был классифицирован как соответствующий предсказанный класс.

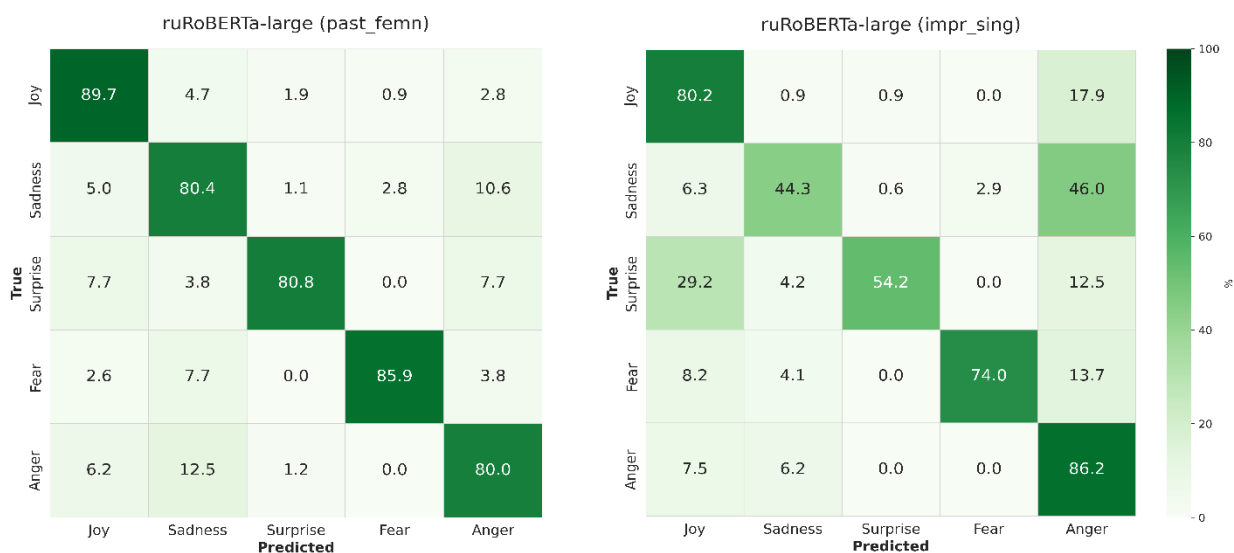


Рис. 1. Матрицы ошибок модели ruRoBERTa-large для оптимальной (past_femn, слева) и наименее эффективной (impr_sing, справа) глагольных форм.

Если для формы прошедшего времени женского рода модель демонстрирует относительно высокую точность предсказаний, то для формы императива точность резко падает для двух классов: «Грусть» и «Удивление». Наиболее явно прослеживается ошибка «Грусть – Злость»: 46% случаев грусти классифицируются как злость. Класс «Удивление» демонстрирует смещение в сторону класса «Радость» (29.2%), хотя и в меньшей степени. В табл. 6 представлены наиболее частотные смещения предсказаний для каждой модели.

Табл. 6. Паттерны смещения предсказаний эмоционального класса для каждой модели.

Модель	Глагольная форма	Истинная метка класса	Предсказанная метка класса	Количество случаев
ruRoBERTa-large	impr_sing	Грусть	Злость	80
	impr_plur	Грусть	Злость	76
	pres_plur_2per	Грусть	Злость	61
ruBERT-base	pres_sing_2per	Грусть	Злость	81
	impr_plur	Грусть	Злость	79
	impr_sing	Грусть	Злость	77
ruBERT-tiny2	pres_sing_2per	Грусть	Злость	113
	impr_plur	Грусть	Злость	102
	infinitive	Грусть	Злость	100

Для всех моделей наиболее частым паттерном является смещение «Грусть – Злость». Эта тенденция особенно характерна для форм императива, а также настоящего времени 2-го лица единственного числа. Если рассмотреть лексические единицы класса «Грусть», которые систематически классифицируются всеми тремя моделями как «Злость», можно сделать следующие наблюдения: как правило, это глаголы с высокой эмоциональной интенсивностью (*убиваться, затравить, изнурять*). Значительное число составляют глаголы с се-

мантикой подавления или насилия (*сломить, сокрушить, раздавить, подавить, давить*).

Таким образом, тенденции предсказаний, по-видимому, обусловлены двумя факторами: глагольными формами, подразумевающими ориентацию на другое лицо (повелительное наклонение, настоящее время второго лица), и семантикой глагола.

5. ХАРАКТЕРИСТИКИ ОБУЧАЮЩИХ ДАННЫХ И ИХ СВЯЗЬ С РЕЗУЛЬТАТАМИ КЛАССИФИКАЦИИ

Эксперименты показали систематическое влияние морфологической формы глагола на классификацию по эмоциональным категориям. Однако необходимо также оценить, в какой степени результаты классификации обусловлены характеристиками обучающих данных. В настоящем разделе проведен анализ обучающего корпуса в двух аспектах: частотность морфологических форм и их распределение по эмоциональным классам.

5.1. Частотность морфологических форм глаголов из RusEmoLex в объединенном корпусе

Для оценки влияния характеристик обучающих данных на результаты классификации были проанализированы частотности глаголов-лемм из RusEmoLex и их морфологических форм в корпусе, составленном из русскоязычных датасетов.

В обучающем корпусе встречаются 230 из 473 глаголов RusEmoLex (48.6%). Для каждого вхождения определялась морфологическая форма глагола. Результаты представлены в табл. 7.

Табл. 7. Частотность морфологических форм глаголов из RusEmoLex в обучающем корпусе.

Форма	Количество вхождений	% от общего числа
pres_sing_1per	1136	35.37
infinitive	470	14.63
pres_sing_3per	426	13.26

past_masc	304	9.46
past_femn	253	7.87
pres_plur_3per	166	5.17
past_plur	153	4.76
impr_sing	108	3.36
pres_sing_2per	76	2.37
impr_plur	60	1.87
pres_plur_1per	47	1.46
pres_plur_2per	13	0.40
Всего вхождений	3212	100

Наиболее частотной оказалась форма настоящего времени 1-го лица единственного числа (35.37% всех вхождений), такой результат ожидаем, принимая во внимание специфику материала: комментарии из социальных сетей зачастую содержат много высказываний от 1-го лица. Инфинитив, являющийся базовой словарной формой глагола, составляет 14.63% вхождений. Формы повелительного наклонения представлены значительно слабее (суммарно 5.2%). Наименее частотной является форма настоящего времени 2-го лица множественного числа (0.40%).

Для количественной оценки связи между частотностью морфологических форм в обучающем корпусе и качеством их классификации был рассчитан коэффициент корреляции Спирмена. Анализ для модели ruRoBERTa-large показал умеренную статистически значимую корреляцию ($\rho = 0.58$; $p = 0.048$), что подтверждает влияние частотности на качество классификации, однако указывает на существенный вклад и других факторов. Для моделей меньшего размера (ruBERT-base и ruBERT-tiny2) корреляция оказалась слабее и статистически незначимой.

Отметим также, что распределение случаев употребления глаголов из RusEмоLex в корпусе характеризуется неравномерностью: 10 наиболее частотных глаголов составляют 49% всех вхождений (табл. 8).

Табл. 8. Наиболее частотные глаголы в объединенном корпусе.

Глагол	Эмоциональный класс	Частотность в корпусе
<i>бояться</i>	Страх	613
<i>любить</i>	Радость	309
<i>ненавидеть</i>	Злость	291
<i>скучать</i>	Грусть	218
<i>плакать</i>	Грусть	179
<i>надеяться</i>	Радость	172
<i>удивляться</i>	Удивление	139
<i>пугать</i>	Страх	121
<i>злиться</i>	Злость	101
<i>убить</i>	Грусть	90

Анализ распределения морфологических форм у наиболее частотных глаголов в корпусе выявил усиление смещения в сторону 1-го лица единственного числа: у 10 наиболее частотных глаголов эта форма составляет более половины от всех случаев употребления (56.9%), в то время как в среднем по корпусу процент ниже, всего 35.4%. При этом доля повелительного наклонения (2.7%) почти вдвое ниже среднего показателя, что усугубляет проблему недостаточной представленности этих форм в обучающих данных.

Таким образом, малое количество форм повелительного наклонения и форм второго лица в обучающем корпусе может рассматриваться как одна из причин систематического снижения качества их эмоциональной классификации моделями-энкодерами.

5.2. Распределение эмоций по морфологическим формам глагола в объединенном корпусе

Для более глубокого понимания факторов, влияющих на классификацию глагольных форм, было проанализировано распределение различных морфоло-

гических форм в обучающем корпусе по классам эмоций. Для каждого вхождения глагола в корпусе фиксировались морфологическая форма глагола и эмоциональная метка всего текста. Результаты анализа для ключевых форм представлены в табл. 9.

Табл. 9. Распределение меток эмоциональных классов по морфологическим формам в обучающем корпусе (%).

Форма	Радость	Грусть	Удивление	Страх	Злость
infinitive	32.24	29.77	10.27	11.50	16.22
past_femn	30.74	30.74	8.89	21.48	8.15
impr_sing	46.85	30.63	1.80	3.60	17.12
pres_sing_2per	26.25	17.50	12.50	7.50	36.25

Инфинитив и форма прошедшего времени женского рода демонстрируют относительно равномерное распределение между основными эмоциональными категориями.

Формы повелительного наклонения единственного числа показывают смещение в сторону эмоции «Радость»: в 46.85% случаев глаголы в форме императива встречаются в текстах, размеченных как выражающие радость (примеры: «не серчай ;)»; «Радуй всех своей улыбкой и больше не болей!=»)). Это превышает соответствующий показатель для инфинитива (32.24%) и формы прошедшего времени женского рода (30.74%). Одновременно для императива характерна низкая представленность эмоций «Удивление» (1.80%) и «Страх» (3.60%).

Форма настоящего времени 2-го лица единственного числа демонстрирует противоположную тенденцию: для нее характерно смещение к эмоции «Злость» (36.25% против 16.22% для инфинитива).

Сопоставление полученных данных о распределении эмоций в корпусе с результатами классификации выявляет следующую закономерность. Несмотря на то что в обучающих данных формы императива преимущественно встречаются в контексте эмоции «Радость» (46.85%), модель ruRoBERTa-large систематически классифицирует формы повелительного наклонения глаголов истинного класса «Грусть» как «Злость»: 46% глаголов, истинной меткой которых является

«Грусть», получают предсказанную метку «Злость» (см. рис. 1). При этом важно отметить, что для других эмоциональных классов форма императива не приводит к столь выраженным искажениям, что может указывать на взаимодействие между эмоциональным семантическим компонентом и грамматической формой: императив в сочетании с семантикой грусти интерпретируется моделью как выражение злости.

Таким образом, можно заключить, что морфологическая чувствительность энкодеров при классификации эмоций определяется комплексом факторов: не только частотностью форм в обучающих данных, но и характеристиками грамматических форм.

ЗАКЛЮЧЕНИЕ

Настоящее исследование позволило установить, что морфологическая форма глагола оказывает систематическое влияние на предсказания языковых моделей-энкодеров при классификации эмоций. Проведенные эксперименты показали, что морфологическая чувствительность моделей составляет 20–25%: различие в качестве классификации между лучшей и худшей глагольными формами достигает этого значения для всех трех исследованных энкодеров (ruBERT-tiny2, ruBERT-base, ruRoBERTa-large). Наиболее высокие результаты классификации демонстрирует форма прошедшего времени женского рода, превосходящая базовый уровень инфинитива. Напротив, формы повелительного наклонения и настоящего времени 2-го лица единственного числа показывают систематическое снижение качества на 20–23%. Анализ результатов также выявил устойчивый паттерн смещения предсказаний: глаголы класса «Грусть» в форме императива систематически классифицируются как «Злость». Глаголы, которым свойственна эта тенденция, как правило, характеризуются с высокой эмоциональной интенсивностью, зачастую глаголы включают семантические компоненты деструктивного воздействия.

Исследование характеристик обучающих данных показало, что чувствительность моделей-энкодеров к морфологии определяется рядом факторов. Частотность морфологических форм в корпусе частично коррелирует с качеством классификации: мало представленные в обучающем корпусе формы (императив, настоящее время 2-го лица множественного числа) демонстрируют худшие

результаты. Однако частотность не является единственным определяющим фактором: модели систематически ассоциируют императив в сочетании с семантикой грусти с выражением агрессии, злости, хотя формы императива в обучающем корпусе преимущественно встречаются в контексте эмоции «Радость». Это может указывать на то, что модели-энкодеры распознают не только статистические корреляции в обучающих данных, но и характеристики грамматических форм. Таким образом, в условиях минимального контекста морфологические характеристики глаголов могут быть значимы для выявления и классификации эмоционального компонента.

Благодарности

Исследование выполнено за счет гранта Российского научного фонда № 25-11-00191, <https://rscf.ru/project/25-11-00191/>.

СПИСОК ЛИТЕРАТУРЫ

1. *Cavicchio F.* Emotion Detection in Natural Language Processing. Cham: Springer, 2025.
2. *Mohammad S.* Best Practices in the Creation and Use of Emotion Lexicons // Findings of the Association for Computational Linguistics: EACL 2023. Dubrovnik, Croatia: Association for Computational Linguistics, 2023. P. 1825–1836.
3. *Loukachevitch N., Levchik A.* Creating a General Russian Sentiment Lexicon // Proc. of the Tenth International Conference on Language Resources and Evaluation (LREC'16). Portorož, Slovenia: European Language Resources Association (ELRA), 2016. P. 1171–1176.
4. *Бабенко Л.Г.* Алфавит эмоций: словарь-тезаурус эмотивной лексики, 2-е издание, стереотипное. Екатеринбург, М.: Кабинетный ученый, 2022.
5. *Deng J., Ren F.* A Survey of Textual Emotion Recognition and Its Challenges // IEEE Transactions on Affective Computing. 2023. Vol. 14, No. 1. P. 49–67. <https://doi.org/10.1109/TAFFC.2021.3053275>
6. *Ярошенко П.В., Лукашевич Н.В.* RusEmoLex: эмоциональный лексикон для русского языка // Russian Journal of Linguistics. 2025. Vol. 29, No. 3. P. 659–687. <https://doi.org/10.22363/2687-0088-44439>

7. *Iaroshenko P.V., Petrova A.A., Loukachevitch N.V.* Morphology and emotion recognition: How verb form affects language model predictions // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference Dialogue 2026. 2026. P. 134–143.
8. *Acheampong F.A., Nunoo-Mensah H., Chen W.* Transformer models for text-based emotion detection: a review of BERT-based approaches // Artif Intell Rev. 2021. Vol. 54. P. 5789–5829. <https://doi.org/10.1007/s10462-021-09958-2>
9. *Kumar R.P., Ramya P., Teja G.S., Krishna V.R.* BERTing Emotions: Exploring Textual Emotion Recognition using NLP // International Conference on Intelligent Computing and Control Systems (ICICCS). Erode, India, 2025. P. 888–895. <https://doi.org/10.1109/ICICCS65191.2025.10984670>
10. *Kusal S., Patil S., Kotecha K., Aluvalu R., Varadarajan V.* AI Based Emotion Detection for Textual Big Data: Techniques and Contribution // Big Data and Cognitive Computing. 2021. Vol. 5, No. 3. Article 43. <https://doi.org/10.3390/bdcc5030043>
11. *De Bruyne L.* The paradox of multilingual emotion detection // Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis. Toronto, Canada: Association for Computational Linguistics, 2023. P. 458–466.
12. *Plaza-del-Arco F.M., Cercas Curry A.A., Cercas Curry A., Hovy D.* Emotion analysis in NLP: Trends, gaps and roadmap for future directions // Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation. 2024. P. 5696–5710.
13. *Ekman P.* Are there basic emotions? // Psychological Review. 1992. Vol. 99, No. 3. P. 550–553.
14. *Plutchik R.* A general psycho-evolutionary theory of emotion // Theories of emotion. Cambridge: Academic Press, 1980. P. 3–33.
15. *Mohammad S., Turney P.* Crowdsourcing a word-emotion association lexicon // Computational Intelligence. 2013. Vol. 29, No. 3. P. 436–465. <http://doi.org/10.1111/j.1467-8640.2012.00460.x>
16. *Сысоева Т.А., Люсин Д.В.* Разработка расширенной базы данных с эмоциональными оценками существительных ENRuN-2: успехи, проблемы и перспективы // Психология познания: материалы Всероссийской научной конференции. ЯрГУ, 6–8 декабря 2024 г. / под ред. И.Ю. Владимирова, С.Ю. Коровкина.

Ярославль: Филигрань, 2024. С. 316–320.

17. Савчук С.О., Архангельский Т.А., Бонч-Осмоловская А.А., Доница О.В., Кузнецова Ю.Н., Ляшевская О.Н., Орехов Б.В., Подрядчикова М.В. Национальный корпус русского языка 2.0: новые возможности и перспективы развития // Вопросы языкознания. 2024. № 2. С. 7–34.

18. Sboev A., Naumov A., Rybka R. Data-driven model for emotion detection in Russian texts // Procedia Computer Science. 2021. Vol. 190. P. 637–642.

19. Kazyulina M., Babii A., Malafeev A. Emotion Classification in Russian: Feature Engineering and Analysis // Analysis of Images, Social Networks and Texts. AIST 2020. Lecture Notes in Computer Science. Vol. 12602. Cham: Springer, 2021. https://doi.org/10.1007/978-3-030-72610-2_10

20. Muhammad S.H., Ousidhoum N., Abdulmumin I. et al. BRIGHTER: BRIdging the Gap in Human-Annotated Textual Emotion Recognition Datasets for 28 Languages // Proc. of the 63rd Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers). Vienna, Austria: Association for Computational Linguistics, 2025. P. 8895–8916. <https://doi.org/10.18653/v1/2025.acl-long.436>

21. Demszky D., Movshovitz-Attias D., Ko J., Cowen A., Nemade G., Ravi S. GoEmotions: A Dataset of Fine-Grained Emotions // Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, 2020. P. 4040–4054.

VERB FORMS IN EMOTION RECOGNITION TASKS: ANALYSIS OF DATA CHARACTERISTICS AND CLASSIFICATION RESULTS

P. V. Iaroshenko¹[0000-0002-9425-5842], A. A. Petrova²[0009-0005-0176-9836],
N. V. Loukachevitch³[0000-0002-1883-4121]

^{1–3}*Lomonosov Moscow State University, Moscow, Russia*

¹polina.iaroshenko@yandex.ru, ²petrova.anastasi999@gmail.com,

³louk_nat@mail.ru

Abstract

Emotion recognition in text represents an important task in natural language processing. This study investigates the influence of verb morphological form on emotion classification performance by encoder models and identifies factors that may determine morphological sensitivity in language models. Three models (ruBERT-tiny2, ruBERT-base, ruRoBERTa-large), fine-tuned on a combined corpus of four Russian-language datasets (22,852 examples), were employed in the experiment. Each model classified 5,676 morphological forms of 473 verbs from the Russian Emotion Lexicon. Results revealed systematic morphological influence: differences in classification performance between forms, as measured by F1-score, reached 20–25% across all models. To account for the observed patterns, training data characteristics were analysed. Frequency of morphological forms in the corpus was shown to partially correlate with classification performance, though it does not appear to be the sole determining factor. Analysis of emotion distribution across morphological forms revealed that in the training corpus, imperative forms occur predominantly in contexts associated with Joy. However, Sadness-labelled verbs in imperative mood are systematically classified by encoders as Anger. This discrepancy between training data characteristics and classification outcomes suggests that imperatives combined with sadness semantics are interpreted by models as expressions of aggression, indicating an interaction between grammatical form and verb semantics in determining emotion class.

Keywords: *emotion classification, BERT, natural language processing, Russian language, morphology.*

Acknowledgements

The research was supported by the Russian Science Foundation, project No. 25-11-00191, <https://rscf.ru/project/25-11-00191/>.

REFERENCES

1. *Cavicchio F.* Emotion Detection in Natural Language Processing. Cham: Springer, 2025.
2. *Mohammad S.* Best Practices in the Creation and Use of Emotion Lexicons // Findings of the Association for Computational Linguistics: EACL 2023. Dubrovnik, Croatia: Association for Computational Linguistics, 2023. P. 1825–1836.
3. *Loukachevitch N., Levchik A.* Creating a General Russian Sentiment Lexicon // Proc. of the Tenth International Conference on Language Resources and Evaluation (LREC'16). Portorož, Slovenia: European Language Resources Association (ELRA), 2016. P. 1171–1176.
4. *Babenko L.G.* Alphabet of Emotions: Thesaurus of Emotive Lexis, 2nd edition. Yekaterinburg; Moscow: Armchair Scientist, 2022.
5. *Deng J., Ren F.* A Survey of Textual Emotion Recognition and Its Challenges // IEEE Transactions on Affective Computing. 2023. Vol. 14, No. 1. P. 49–67. <https://doi.org/10.1109/TAFFC.2021.3053275>
6. *Iaroshenko P.V., Loukachevitch N.V.* RusEmoLex: Russian Emotion Lexicon // Russian Journal of Linguistics. 2025. Vol. 29, No. 3. P. 659–687. URL: <https://doi.org/10.22363/2687-0088-44439>
7. *Iaroshenko P.V., Petrova A.A., Loukachevitch N.V.* Morphology and emotion recognition: How verb form affects language model predictions // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference Dialogue 2026. 2026. P. 134–143.
8. *Acheampong F.A., Nunoo-Mensah H., Chen W.* Transformer models for text-based emotion detection: a review of BERT-based approaches // Artif Intell Rev. 2021. Vol. 54. P. 5789–5829. <https://doi.org/10.1007/s10462-021-09958-2>
9. *Kumar R.P., Ramya P., Teja G.S., Krishna V.R.* BERTing Emotions: Exploring Textual Emotion Recognition using NLP // International Conference on Intelligent Computing and Control Systems (ICICCS). Erode, India, 2025. P. 888–895. <https://doi.org/10.1109/ICICCS65191.2025.10984670>

10. *Kusal S., Patil S., Kotecha K., Aluvalu R., Varadarajan V.* AI Based Emotion Detection for Textual Big Data: Techniques and Contribution // *Big Data and Cognitive Computing*. 2021. Vol. 5, No. 3. Article 43. URL: <https://doi.org/10.3390/bdcc5030043>
11. *De Bruyne L.* The paradox of multilingual emotion detection // *Proc. of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*. Toronto, Canada: Association for Computational Linguistics, 2023. P. 458–466.
12. *Plaza-del-Arco F.M., Cercas Curry A.A., Cercas Curry A., Hovy D.* Emotion analysis in NLP: Trends, gaps and roadmap for future directions // *Proc. of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*. 2024. P. 5696–5710.
13. *Ekman P.* Are there basic emotions? // *Psychological Review*. 1992. Vol. 99, No. 3. P. 550–553.
14. *Plutchik R.* A general psycho-evolutionary theory of emotion // *Theories of emotion*. Cambridge: Academic Press, 1980. P. 3–33.
15. *Mohammad S., Turney P.* Crowdsourcing a word-emotion association lexicon // *Computational Intelligence*. 2013. Vol. 29, No. 3. P. 436–465. URL: <http://doi.org/10.1111/j.1467-8640.2012.00460.x>
16. *Sysoeva T.A., Lyusin D.V.* Development of an extended database with emotional ratings of nouns ENRuN-2: Successes, problems and prospects // *Psychology of cognition: Proc. of the All-Russian Scientific Conference, YARSU, December 6-8, 2024 / I.Yu. Vladimirov, S.Yu. Korovkin (eds.)*. Yaroslavl: Filigran, 2024. P. 316–320.
17. *Savchuk S.O., Arkhangel'skiy T., Bonch-Osmolovskaya A.A., Donina O.V., Kuznetsova Yu.N., Lyashevskaya O.N., Orekhov B.V., Podryadchikova M.V.* Russian National Corpus 2.0: New opportunities and development prospects // *Voprosy Jazykoznanija*. 2024. No. 2. P. 7–34.
18. *Sboev A., Naumov A., Rybka R.* Data-driven model for emotion detection in Russian texts // *Procedia Computer Science*. 2021. Vol. 190. P. 637–642.
19. *Kazyulina M., Babii A., Malafeev A.* Emotion Classification in Russian: Feature Engineering and Analysis // *Analysis of Images, Social Networks and Texts. AIST 2020. Lecture Notes in Computer Science*. Vol. 12602. Cham: Springer, 2021. https://doi.org/10.1007/978-3-030-72610-2_10

20. *Muhammad S.H., Ousidhoum N., Abdulmumin I. et al.* BRIGHTER: BRIdging the Gap in Human-Annotated Textual Emotion Recognition Datasets for 28 Languages // Proc. of the 63rd Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers). Vienna, Austria: Association for Computational Linguistics, 2025. P. 8895–8916. <https://doi.org/10.18653/v1/2025.acl-long.436>
21. *Demszky D., Movshovitz-Attias D., Ko J., Cowen A., Nemade G., Ravi S.* GoEmotions: A Dataset of Fine-Grained Emotions // Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, 2020. P. 4040–4054.
-

СВЕДЕНИЯ ОБ АВТОРАХ



ЯРОШЕНКО Полина Владимировна – кандидат филологических наук, научный сотрудник Лаборатории анализа информационных ресурсов Научно-исследовательского вычислительного центра Московского государственного университета имени М. В. Ломоносова. Сфера научных интересов: компьютерная семантика, обработка естественного языка, искусственный интеллект, анализ эмоций.

Polina V. IAROSHENKO, PhD, is a research fellow in the Laboratory for Information Resources Analysis at the Research Computing Center, Lomonosov Moscow State University. Research interests: computational semantics, natural language processing, artificial intelligence, emotion analysis.

email: polina.iaroshenko@yandex.ru

ORCID: 0000-0002-9425-5842



ПЕТРОВА Анастасия Алексеевна — аспирант факультета вычислительной математики и кибернетики Московского государственного университета имени М. В. Ломоносова. Сфера научных интересов: анализ эмоций, обработка естественного языка.

Anastasiia A. PETROVA is a PhD student at the Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University. Research interests: emotion analysis, natural language processing.

email: petrova.anastasi999@gmail.com

ORCID: 0009-0005-0176-9836



ЛУКАШЕВИЧ Наталья Валентиновна — доктор технических наук, ведущий научный сотрудник Лаборатории анализа информационных ресурсов Научно-исследовательского вычислительного центра Московского государственного университета имени М. В. Ломоносова. Сфера научных интересов: компьютерная семантика, обработка естественного языка, искусственный интеллект, анализ тональности.

Natalia V. LOUKACHEVITCH, PhD (Advanced Doctorate), is a leading research fellow in the Laboratory for Information Resources Analysis at the Research Computing Center, Lomonosov Moscow State University. Research interests: computational semantics, natural language processing, artificial intelligence.

email: louk_nat@mail.ru

ORCID: 0000-0002-1883-4121

Материал поступил в редакцию 26 августа 2026 года