



MASTERARBEIT | MASTER'S THESIS

Titel | Title

Evaluating Sarcasm Preservation in Machine Translation from
English to Russian and Its Impact on Automatic Sarcasm
Detection

verfasst von | submitted by
Arina Antonova BA

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of
Master of Science (MSc)

Wien | Vienna, 2026

Studienkennzahl lt. Studienblatt | Degree
programme code as it appears on the
student record sheet:

UA 066 587

Studienrichtung lt. Studienblatt | Degree
programme as it appears on the student
record sheet:

Joint-Masterstudium Multilingual Technologies

Betreut von | Supervisor:

Assoz. Prof. Mag. Dr. Dagmar Gromann BSc

Acknowledgements

I would like to express my deepest gratitude to my mother for her unwavering support, endless encouragement, and constant belief in me throughout my academic journey. Her confidence in me has always been a source of inspiration.

I would like to express my sincere gratitude to my supervisor, Professor Dagmar Gromann, for her invaluable guidance, continuous support, and insightful feedback throughout the development of this master's thesis. Her expertise and encouragement have made a significant contribution to this work.

I am also deeply grateful to Liad Magen for his inspiring courses and encouragement throughout the master's program, which greatly motivated my academic development.

Finally, I would like to thank my colleague and friend, Anastasiya, for her consistent support, kindness, and encouragement throughout this process. Her presence made this journey much easier and far more enjoyable.

Abstract

Sarcasm is a form of figurative language in which the intended meaning of a sentence differs from its literal interpretation and is often used to express mockery, humor or critique. Due to its implicit and context-dependent nature, sarcasm remains a challenge for Natural Language Processing (NLP), particularly in sentiment analysis, machine translation (MT), and automatic sarcasm detection. Although sarcasm detection has been widely studied in monolingual settings, the impact of MT on sarcasm preservation remains rather underexplored.

This master’s thesis studies the relationship between English-Russian MT and automatic sarcasm detection. It investigates how MT affects the preservation of sarcastic meaning and its impact on automatic sarcasm detection, as well as which translation strategies are used during MT of sarcastic content.

This thesis uses a multilingual XLM-R model fine-tuned on English and Russian sarcasm datasets using four training configurations: English-only, Russian-only, combined English-Russian, and sequential English-to-Russian training, where the model is first fine-tuned on English and then further fine-tuned on Russian. The English test set is translated into Russian using the NLLB MT system and evaluated using the combined model, which is the best-performing sarcasm detection model among the four training configurations. A qualitative analysis further examines translation techniques in relation to sarcasm preservation.

The results show a significant drop in sarcasm detection performance on machine-translated data compared to the original English and Russian test sets. Although the translated texts achieve moderate BERTScore and COMET-QE, only a small proportion of sarcastic segments are correctly identified as sarcastic after MT. These findings suggest that MT may weaken pragmatic cues necessary for sarcasm detection, even when general semantic content is preserved. Overall, the study highlights the challenges of sarcasm detection in multilingual NLP and investigates how MT affects sarcasm preservation, detection performance, and translation strategies in sarcastic content.

Zusammenfassung

Sarcasmus ist eine Form bildlicher Sprache, bei der sich die beabsichtigte Bedeutung eines Satzes von seiner wörtlichen Interpretation unterscheidet und häufig verwendet wird, um Spott, Humor oder Kritik auszudrücken. Aufgrund seiner impliziten und kontextabhängigen Natur stellt Sarcasmus eine Herausforderung für Natural Language Processing (NLP) dar, insbesondere in der Sentimentanalyse, maschinellen Übersetzung (MT) und automatischen Sarkasmuserkennung. Obwohl Sarkasmuserkennung in monolingualen Settings gut untersucht ist, bleibt der Einfluss von MT auf die Bewahrung von Sarkasmus vergleichsweise wenig erforscht.

Diese Masterarbeit untersucht die Beziehung zwischen englisch-russischer MT und automatischer Sarkasmuserkennung. Sie analysiert, wie MT die Bewahrung sarkastischer Bedeutung sowie die Erkennungsleistung beeinflusst und welche Übersetzungsstrategien bei sarkastischen Inhalten verwendet werden.

Die Arbeit verwendet ein multilingual trainiertes XLM-R Modell, das auf englischen und russischen Sarkasmus-Datasets in vier Konfigurationen fine-tuned wird: Englisch-only, Russisch-only, kombiniert Englisch-Russisch sowie sequentielles Englisch-zu-Russisch-Training, bei dem das Modell zunächst auf Englisch und anschließend auf Russisch fine-tuned wird. Der englische Testset wird mit dem NLLB-System ins Russische übersetzt und anschließend mit dem bestperformenden kombinierten Modell ausgewertet. Eine qualitative Analyse ergänzt die Untersuchung der verwendeten Übersetzungsstrategien im Kontext der Sarkasmusbewahrung.

Die Ergebnisse zeigen einen deutlichen Rückgang der Sarkasmuserkennung auf maschinell übersetzten Daten im Vergleich zu den ursprünglichen englischen und russischen Testsets. Obwohl die Übersetzungen moderate BERTScore- und COMET-QE-Werte erreichen, wird nur ein kleiner Teil der sarkastischen Segmente als sarkastisch korrekt erkannt. Dies deutet darauf hin, dass MT pragmatische Hinweise für die Sarkasmuserkennung abschwächen kann, selbst wenn die semantische Bedeutung erhalten bleibt.

Insgesamt zeigt die Studie die Herausforderungen der Sarkasmuserkennung in multilingualen NLP-Szenarien und untersucht den Einfluss von MT auf Sarkasmusbewahrung, Erkennungsleistung und Übersetzungsstrategien.

Contents

List of Tables	11
List of Figures	13
1 Introduction	1
2 Sarcasm as a Linguistic Phenomenon	5
2.1 Sarcasm and Related Phenomena	5
2.2 Linguistic Features of Sarcasm	6
2.3 Sarcasm classification	8
2.4 Sarcasm and Translation Techniques	10
3 Fundamentals of Machine Learning	13
3.1 Core concepts of Machine Learning	13
3.2 From Traditional Machine Learning to Deep Learning	14
3.3 Machine Learning in Natural Language Processing	16
3.4 Evaluation metrics	18
4 Approaches to Machine Translation	21
4.1 Evolution of Machine Translation Approaches	21
4.2 Neural Machine Translation and Transformers	22
4.3 Evaluation of Machine Translation	24
4.4 Challenges of Machine Translation	25
5 Sarcasm Detection Approaches	27
5.1 Automatic Sarcasm Detection	27
5.2 Feature-Based Approaches	28
5.3 Deep Learning Approaches	29
5.4 Challenges of Multilingual Sarcasm Detection	30
6 Sarcasm and Machine Translation	33
6.1 Challenges of Sarcasm Translation	33
6.2 Evaluation of Machine-Translated Sarcasm	34
6.3 Impact of Machine Translation on Sarcasm Detection	35

7 Related Work	37
8 Methodology	41
8.1 Datasets Selection and Preprocessing	41
8.2 Model Selection and training configurations	43
8.3 Sarcasm Detection Approach	44
8.4 Machine Translation Approach	45
8.5 Evaluation	45
8.6 Qualitative Analysis	46
9 Results	47
9.1 Training on English dataset	48
9.2 Training on Russian dataset	49
9.3 Combined training	51
9.4 Sequential training	52
9.5 Sarcasm detection on machine translated dataset	53
9.6 Qualitative analysis of the translation techniques in Machine trans- lated dataset	55
10 Discussion	59
11 Conclusion	63
Bibliography	65

List of Tables

1	Translation techniques adapted from (Newmark, 1988; Zawawi and Maghfiroh, 2020; Molina and Albir, 2002)	11
2	Overview of the selected datasets used for sarcasm detection	42
3	Final dataset statistics across splits	43
4	Overview of models and hyperparameters used in the experiment	44
5	Overview of the results of the test set evaluation on different model configurations	47
6	Translation techniques applied for successfully identified sarcastic segments	55
7	Translation techniques applied for unsuccessfully identified sarcastic segments	57

List of Figures

1	Overview of the F1 score across different models	48
2	Confusion matrices for English-only model evaluated on the English and Russian test sets	49
3	Confusion matrices for the Russian-only model evaluated on the English and Russian test sets	50
4	Confusion matrices for the Combined model evaluated on the English and Russian test sets	51
5	Confusion matrices for the Sequential model evaluated on the English and Russian test sets	52
6	Overview of the F1 scores evaluated on the English, Russian and machine translated test sets	53
7	Confusion matrices for the Combined model evaluated on the ma- chine translated dataset	54

1 Introduction

Sarcasm represents a form of figurative language in which the intended meaning of a sentence differs from its literal interpretation and is often used to express mockery, humour or critique (Camp, 2012; Ducret et al., 2020; Tiwari et al., 2024). As a result, the correct interpretation of sarcastic expressions could be rather challenging, as it depends not only on internal lexical content but also must rely on contextual or cultural factors along with paralinguistic cues (Pradhan et al., 2024; Ducret et al., 2020).

Moreover, due to a rapid development of social networks and online presence, sarcasm presents particular difficulties in terms of User Generated Content (UGC), such as tweets, product reviews, blogs and online posts (Pradhan et al., 2024; Ducret et al., 2020; Khan et al., 2025; Tiwari et al., 2024). Such content usually contains colloquial constructions, emotionally colored vocabulary, sarcasm and newly formed slang, which combined with already challenging nature of sarcastic expressions poses additional difficulties for various Natural Language Processing (NLP) tasks, such as sarcasm detection, sentiment analysis, opinion extraction e.g. (Pradhan et al., 2024; Ducret et al., 2020).

In the field of Machine Translation (MT), emotionally loaded UGC usually poses additional difficulties for both automatic systems and evaluation metrics, because such content may often lose or change its intended sentiment after being automatically translated (Qian et al., 2024; Saadany et al., 2023). This makes sarcasm, which can often be featured as part of UGC, a particularly challenging phenomenon for MT systems. Preserving the communicative function of the original message in the translation often requires more than direct lexical transfer and includes consideration of additional cultural and contextual cues (Zawawi and Maghfiroh (2020)), while no paralinguistic cues typical for spoken communication are available in the written format.

In the field of NLP various approaches have been proposed to handle the task of automatic sarcasm detection, including machine learning models and neural networks. Nevertheless, sarcasm detection is still considered to be a serious challenge, as sarcastic meaning often depends on contextual information that is not always explicitly expressed in the text (Helal et al., 2024; Khan et al., 2025). Ducret et al. (2020) show that effective sarcasm detection often relies on a combination of lexical, stylistic, and contextual features, as sarcastic meaning is rarely expressed explicitly in text. Similarly, Pradhan et al. (2024) note that

1 Introduction

capturing the implicit and context dependent nature of sarcasm remains challenging for automatic systems, particularly without paralinguistic cues.

Despite the active development of automatic sarcasm recognition methods, most of the existing research focuses on monolingual settings or on cross-lingual comparisons that evaluate model performance across different languages without explicitly considering the role of MT. For example, Tasnim and Nath (2024) compare sarcasm detection models across English and Bangla, while Гурин and Жуков (2023) conduct experiments exclusively on Russian data. Similarly, Abu Farha and Magdy (2021) study sarcasm detection in the Arabic language and demonstrate that monolingual models often outperform multilingual ones for this specific task. While a recent study by Chia et al. (2024) explores the relationship between irony, sarcasm and MT, this work focuses on the translation of figurative language and additionally on improving model performance through combined translation and irony detection tasks. In contrast, the question of whether the original sarcastic intent is preserved in machine-translated text and whether it remains detectable by sarcasm detection models trained in the source language has received comparatively limited attention, particularly for specific language pairs such as English-Russian.

This study investigates the interaction between MT and automatic sarcasm detection. In particular, the study focuses on the English to Russian MT direction, where English texts are machine-translated into Russian and then subsequently analyzed using a sarcasm detection model. To address this problem, the sarcasm detection model is first evaluated on English and Russian datasets in order to establish their ability to recognize sarcastic expressions in both languages. The model is then applied to machine-translated data to examine whether sarcastic meaning remains detectable after translation. Based on this approach, the study addresses the following research questions:

1. To what extent can a sarcasm detection approach in English and Russian accurately detect the preservation or loss of sarcasm in machine translated contents from English to Russian?
2. How does the preservation or loss of sarcasm in the MT output affect sarcasm detection performance in Russian?
3. What translation strategies are applied in MT approaches to translate sarcastic contents?

It is assumed that translating sarcastic expressions poses significant challenges, and that sarcastic meaning may be partially lost or changed in the MT output. Such changes are expected to influence the performance of sarcasm detection models applied to translated data.

This master’s thesis is structured around a series of experiments and analyses aimed at investigating sarcasm translation and detection in the English to Russian direction. The source code and experimental implementation are available on GitHub.¹ First, the XLM-R base model (Conneau et al., 2020) is fine-tuned on four dataset configurations: the English dataset, the Russian dataset, a combined English-Russian dataset, and sequentially, first on the English and then on the Russian dataset. To further examine the impact of MT, the English dataset is machine translated into Russian using NLLB (Team et al., 2022). The translated dataset is then processed using the model configuration that showed the best results after fine-tuning to assess whether sarcastic intent remains detectable after translation. Finally, a qualitative analysis is conducted to explore which translation techniques were used for sarcastic expressions by the MT system.

This master’s thesis investigates the impact of MT on sarcasm detection across languages. It aims to contribute insights on how sarcastic meaning is preserved in machine-translated texts, assessment of sarcasm detection performance on translated data, and an analysis of the translation strategies used for sarcastic expressions in MT. By combining experimental and qualitative analyses, this master’s thesis aims to provide insights into how MT systems influence automatic sarcasm detection and highlight the challenges posed by pragmatically rich content in multilingual NLP settings. The findings are expected to offer a novel evaluation scenario for MT researchers and developers by examining the faithfulness of MT in terms of preserving sarcastic meaning. They are also expected to be valuable for computational linguistics researchers by providing insights into how effectively current methods can detect sarcasm in multilingual and machine-translated settings.

¹All implementation code is available at: https://github.com/MLT2026/master_thesis

2 Sarcasm as a Linguistic Phenomenon

Sarcasm is a complex phenomenon in human communication that combines linguistic, pragmatic, and cultural aspects (Camp, 2012; Pradhan et al., 2024; Ducret et al., 2020; Tiwari et al., 2024). In sarcastic expressions, the intended meaning often differs from the literal meaning of the statement, which makes interpretation dependent on paralinguistic cues, contextual information and shared background knowledge (Pradhan et al., 2024; Zawawi and Maghfiroh, 2020; Helal et al., 2024). These characteristics make sarcasm particularly challenging not only for human interpretation but also for translation and automatic language processing tasks (Ducret et al., 2020; Wang et al., 2024). Therefore, this chapter introduces sarcasm as a linguistic phenomenon and discusses its main linguistic characteristics, as well as the challenges involved in translating sarcastic expressions across languages.

2.1 Sarcasm and Related Phenomena

Sarcasm is considered to be a figurative linguistic phenomenon in verbal or written communication that serves to express mockery or humor and which is often applied when the speaker's intent does not correspond to the actual meaning of the statement (Tiwari et al., 2024; Ducret et al., 2020; Camp, 2012). Usually sarcastic expressions are used in order to convey criticism, evaluation or ridicule of a person or object by means of negative or exaggerated statements and often unintentionally causing negative emotions for the recipient (Pradhan et al., 2024).

Sarcasm is closely related to irony, although their relationship remains subject to debate in the scientific literature, as some researchers consider irony a broader category with sarcasm as its subtype (Zhu and Filik, 2023; Loakman et al., 2023) or argue for the clearer distinction of those phenomena (Camp, 2012). A commonly accepted view is that irony refers to a general mismatch between expectations and reality, and does not necessarily imply negative evaluation of the situation (Joshi et al., 2016; Loakman et al., 2023; Khan et al., 2025). In this context, sarcasm is often considered a specific type of verbal irony (Ducret et al., 2020; Kreuz and Caucci, 2007; Zhu and Filik, 2023).

However, the key distinction between irony and sarcasm lies in their communic-

2 Sarcasm as a Linguistic Phenomenon

ative function. While irony can be neutral or even positive and may exist without targeting a specific individual or object, sarcasm is typically more explicit and carries a clear negative or critical intent and it is frequently directed at a specific target in order to express mockery or ridicule (Joshi et al., 2016; Loakman et al., 2023; Khan et al., 2025). In this sense, sarcasm can be described as a more aggressive and negatively charged instance of irony (Loakman et al., 2023), although the boundary between the two remains flexible and context-dependent (Camp, 2012).

In addition to irony, sarcasm is also related to other broader phenomena such as humor and satire. Humor generally arises from incongruity between expectations and reality and is highly dependent on context, background knowledge, and individual interpretation (Loakman et al., 2023). Although both humor and sarcasm rely on incongruity between expectations and reality, they differ in their communicative purpose. While humor primarily aims to entertain, sarcasm uses incongruity as a means of expressing criticism, mockery, or negative evaluation (Chen et al., 2022). Satire, in turn, is a form of humor that employs devices such as irony and exaggeration to criticize social norms, individuals, or political contexts (Nandy et al., 2024).

Taken together, sarcasm can be understood as a distinct form of figurative language that is closely related to irony, humor, and satire. However, it differs from these phenomena in its explicit evaluative function, as it typically conveys negative or critical intent directed at a specific target (Joshi et al., 2016; Loakman et al., 2023; Khan et al., 2025). Recognizing these distinctions is particularly important for both theoretical analysis and practical applications.

2.2 Linguistic Features of Sarcasm

Understanding sarcasm requires considering linguistic features as well as special context and background knowledge, as perception of a sarcastic expression may vary across different cultures (Zawawi and Maghfiroh, 2020; Helal et al., 2024).

From a linguistic perspective, sarcasm is characterized by several recurring features identified in previous research. One of the most prominent features is incongruity, which involves a contrast between the literal meaning and the contextual or intended interpretation, often expressed as a contradiction between sentiments or between sentiment and situation (Chen et al., 2022; Joshi et al., 2015). Another common feature is exaggeration, where strongly affective or intensified expressions are used to convey an implicit negative or ironic meaning (Chen et al., 2022).

In addition, sarcasm may be signaled through various lexical and pragmatic cues, including the use of interjections, punctuation like exclamation marks or rhetorical questions, formulaic expressions, and repetition, which can indicate non-literal

2.2 Linguistic Features of Sarcasm

intent (Kreuz and Caucci, 2007). These features demonstrate that sarcasm does not rely on a single linguistic marker, but rather emerges from a combination of lexical, pragmatic, and contextual signals.

In spoken communication, sarcasm is typically expressed by various cues, including intonation, pauses, speech rate, pitch and facial expressions. These signals often make the speaker’s intended meaning more straightforward and easier for the recipient to interpret, even when the content of the expression contradicts its literal meaning (Pradhan et al., 2024; Tiwari et al., 2024).

The following sentences were created to illustrate typical examples of sarcastic expressions:

- (1) *Of course we should get up at 5 a.m. each morning to go for a run.*
- (2) *Let’s buy this product with a 1.5 rating.*
- (3) *It’s a great idea to watch Titanic during our first cruise.*

In a face-to-face setting, these sentences would rarely be interpreted literally by the recipients, as paralinguistic and contextual cues would guide them toward a sarcastic interpretation and trigger the appropriate reaction.

However, if the same examples are transferred to plain text instead of spoken communication, these disambiguating signals would no longer be available. As a result, the interpretation of sarcastic written communication must rely on interpreting the implied meaning, contextual cues, and world knowledge rather than literal interpretation of statements. Importantly, the sarcastic intent in these examples does not follow a specific lexical or syntactic pattern, but rather comes from a pragmatic mismatch between the literal meaning of the statement and the situational context.

Examples (1)-(3) illustrate how sarcastic meaning emerges from this contrast between the literal statement and contextual expectations. In (1) even if the literal meaning of the sentence suggests going for a run, the sarcastic intent could be deduced from rather early time and especially with the additional phrase *Of course*. In (2), suggesting to buy a product with a 1.5-star rating implies sarcasm, since such a low rating typically indicates poor product quality. The sarcastic interpretation arises from the contrast between the literal suggestion and the commonly shared understanding that products with extremely low ratings are usually avoided.

Similarly, in (3), watching *Titanic* during a first cruise is unlikely to be perceived as a genuinely positive suggestion, as the film is associated with the sinking of a ship and drowning. This mismatch between the literal statement *It’s a great idea* and the contextual implications creates the sarcastic effect.

This lack of clearly identifiable and consistent patterns makes sarcasm particularly challenging for written communication and specifically for tasks like information extraction or sentiment analysis, where paralinguistic cues are absent (Pradhan et al.,

[2024; Ducret et al., 2020]). This difficulty arises from the fact that sarcasm relies on a combination of linguistic, contextual, and cultural features, including pragmatic mismatch, situational context, and shared background knowledge (Zawawi and Maghfiroh, 2020; Helal et al., 2024).

2.3 Sarcasm classification

Camp (2012) states that sarcasm involves a form of targeted insincerity, where the speaker deliberately distances themselves from the literal meaning of the expression. She suggests that sarcastic meaning operates on different levels, depending on the scope of the intended interpretation. Based on this, sarcasm can be classified into several types according to the level at which the sarcastic effect is applied, which are propositional sarcasm, lexical sarcasm, ‘like’-prefixed sarcasm and illocutionary sarcasm. The examples presented below were specifically constructed by the author for illustrative purposes based on the classifications and definitions discussed by (Camp, 2012).

(4) *What lovely weather today to go camping.*

Propositional sarcasm is a type of sarcasm where the speaker expresses a proposition but actually intends the opposite meaning. The speaker pretends to express something that is often positive while implicitly communicating a contrary meaning, usually negative. This intended meaning is not directly stated but could be guessed by the listener, which allows the speaker to avoid full responsibility for it and keep a degree of deniability. In example (4), if someone looks outside and sees heavy rain, the literal meaning is that the weather is nice, but the intended meaning is that the weather is actually very bad to go out.

(5) *If she is so capable, then she can manage the whole project.*

Lexical sarcasm operates at the level of individual words or expressions, where the sarcastic effect is achieved through the inversion of a specific lexical item rather than the entire proposition, typically it can be an evaluative word such as *brilliant* or *genius*, which is often accompanied by comparative expressions like *so*, *such a*, or *like that*. Example (5) uses *so capable* sarcastically to suggest that the person is, in fact, not capable at all.

(6) *Like she’s really going to arrive on time for our flight.*

‘Like’-prefixed sarcasm is a type of sarcasm where *like* or *as if* is used before a sentence to explicitly reject a proposition. It targets the whole statement and expresses direct denial, not just an implied opposite meaning, so the speaker is

clearly committed to rejecting it and has little deniability. It typically relies on a specific context where someone has suggested the proposition, which the speaker dismisses as highly unlikely. Example (6) sarcastically rejects the idea that the person will actually do as suggested.

(7) *Thanks for all your help.*

Illocutionary sarcasm is applied when the speaker sarcastically performs an entire speech act such as thanking, asking, or complimenting, rather than just inverting a word or proposition. The speaker pretends to make a sincere statement that would be appropriate in an opposite situation, meanwhile highlighting the contrast with reality and expressing an opposite attitude. Example (7) is a response to the situation when someone forgets to do their part of a group project, in which case the sentence expresses frustration with the colleague rather than gratitude.

Another classification is proposed by [Zawawi and Maghfiroh \(2020\)](#), who categorizes sarcasm according to its communicative goals, which are highly context-dependent. Unlike [Camp \(2012\)](#)'s approach, where it is based on the operating level of the sarcastic intent, this classification focuses on the pragmatic functions sarcasm may serve in interaction. These include mockery or ridicule, satire, the expression of proximity, and humor. It should also be noted that some of these categories may also be considered broader or related phenomena in comparison to sarcasm. However, within this Zawawi's classification they are treated as communicative purposes that could be fulfilled by sarcasm.

Mockery or ridicule involves direct and often harsh ridicule, where the speaker openly insults or uses bitter words to hurt the recipient. It is typically explicit and confrontational. The following example demonstrates how the speaker is trying to purposely hurt the other person: "You are shrimp - brained! stupid!" ([Zawawi and Maghfiroh, 2020](#), p.255).

Satire involves a more indirect use of sarcasm, where criticism is conveyed through apparent praise or neutral statements that imply the opposite meaning. Unlike direct mockery, the sarcastic intent is less explicit and relies more heavily on contextual interpretation. The following example demonstrates a sentence with a neutral surface meaning, however with negative evaluation: "It is so good for you to stay here (for this long); it is completely free" ([Zawawi and Maghfiroh, 2020](#), p.255).

The expression of proximity refers to the use of sarcasm as a marker of social closeness, where speakers use mildly insulting sarcastic expressions in a playful way to signal familiarity rather than hostility. For example, *Well done, idiot.* said jokingly between close friends.

Sarcasm can also serve a humorous function, where it is used for entertainment and comedic effect. The goal is to amuse rather than seriously criticize by often

exaggerating someone’s behavior or situation. For example, *Wow, you are clearly a morning person*, said jokingly after oversleeping and rushing in chaos.

Overall, these classifications show that sarcasm can operate at different levels of meaning and fulfill various communicative functions. [Camp \(2012\)](#)’s approach focuses on how sarcasm is structured within a sentence, while [Zawawi and Maghfiroh \(2020\)](#)’s classification highlights its pragmatic roles in interaction. Together, they provide a more comprehensive understanding of how sarcasm functions depending on context and speaker intention.

2.4 Sarcasm and Translation Techniques

Perception of sarcasm is directly connected to a specific culture or domain, as the interpretation of a specific expression may differ and require additional contextual information or deeper understanding of the language ([Khan et al., 2025](#); [Tiwari et al., 2024](#); [Zawawi and Maghfiroh, 2020](#)). Therefore, sarcastic sentences that are clear for the first language speakers, could be translated literally into another language if the appropriate translation techniques are not applied ([Zawawi and Maghfiroh, 2020](#)) which, in turn, could lead to misunderstanding.

Therefore, choosing the appropriate translation technique is very important to successfully convey the message of the original sentence into the translation. Research by [Zawawi and Maghfiroh \(2020\)](#) shows that translation techniques, such as compensation, modulation, adaptation and transposition are most frequently used to successfully translate sarcasm while preserving the original meaning. Table [1](#) demonstrates the most common translation techniques used for sarcasm translation ([Newmark, 1988](#); [Zawawi and Maghfiroh, 2020](#); [Molina and Albir, 2002](#)), together with their definitions and English-Russian translation examples that were specifically constructed by the author for illustrative purposes based on the classifications and definitions discussed by ([Newmark, 1988](#); [Zawawi and Maghfiroh, 2020](#); [Molina and Albir, 2002](#)).

The translation techniques presented in Table [1](#) differ in the types of modifications they introduce during the translation process. Each technique addresses a particular aspect of translation and may involve lexical, grammatical, semantic, or pragmatic changes depending on the characteristics and challenges of the source text. Together, these techniques provide translators with different approaches to preserving meaning and communicative effect in the target text.

Table 1: Translation techniques adapted from (Newmark, 1988; Zawawi and Magh-firoh, 2020; Molina and Albir, 2002)

Translation technique	Definition	Example (EN → RU)
Compensation	Moves an element from the Source Text (ST) that cannot be reproduced in the Target Text (TT) to a different place in the TT	I give thee my word → Я даю тебе честное слово
Adaptation	Replaces a ST cultural element with one from the target culture	They met at a diner after the game → Они встретились в кафе после игры
Modulation	Changes the point of view, additionally the meaning could also be changed	You are going to be a father → У тебя будет ребёнок
Transposition	Changes grammatical classes without changing the meaning	Her explanation was clear → Она объяснила понятно
Reduction	Deletes a ST information item in the TT	She wore a long, blue silk dress to the party → Она пришла на вечеринку в платье
Addition	Adds details that are not formulated in the ST	He drives a Tesla → Он водит Tesla — электромобиль известной американской марки
Description	Replaces a term or expression with a description of its form or/and function	We had a full English breakfast → У нас сегодня была на завтрак яичница с сосисками, фасолью и тостами
Borrowing	Places a word from another language directly into other languages	They are playing football → Они играют в футбол
Equivalence	Uses a term or expression that is recognized (by dictionaries or language in use) as an equivalent in the TT	It's raining cats and dogs → Льёт как из ведра

2 Sarcasm as a Linguistic Phenomenon

Translation technique	Definition	Example (EN → RU)
Calque	Translates and places a foreign word or phrase into another language	Long time no see → Давно не виделись
Explication	Introduces information from the ST that is implicit from the situation or context	She lowered her head when greeting the elder → Она склонила голову при приветствии старшего, показывая уважение
Literal translation	Word-for-word translation	I like chocolate → Я люблю шоколад

In practice, human translators may choose these strategies individually or in combination, depending on their competence and on the linguistic and contextual cues in the source text (Qian et al., 2023). MT systems, in contrast, still face significant challenges in handling figurative language, culturally specific expressions, and subtle emotional cues, which can lead to weakening or loss of the originally intended sarcastic meaning (Wang et al., 2024). Therefore, successful translation of sarcasm requires careful attention to linguistic, cultural, and contextual factors, which help to transfer its intended meaning.

3 Fundamentals of Machine Learning

This chapter introduces the main concepts of machine learning that are relevant to automatic sarcasm detection. Since the sarcasm detection task is formulated in this thesis as a supervised text classification task, the chapter first explains the main learning paradigms, then it outlines the development from traditional Machine Learning (ML) to deep learning and explains the role of datasets and labels, and the use of ML in Natural Language Processing (NLP). Finally, it presents the evaluation metrics that are commonly used for model evaluation.

3.1 Core concepts of Machine Learning

ML is defined as a subfield of computer science that is focused on the development and application of algorithms which allow computer systems to learn patterns from data in order to solve practical problems by analyzing data, identifying patterns or making predictions (Bhatt and Gaikwad, 2023; Burkov, 2019). ML algorithms are commonly used in the field of NLP for such tasks as text classification, sentiment analysis and named entity recognition (Magnini et al., 2020). ML methods differ in the way models learn from data and the type of information that is available during the training process. This chapter discusses five commonly used learning paradigms: supervised learning, unsupervised learning, semi-supervised learning, self-supervised learning, and reinforcement learning (Burkov, 2019).

In supervised learning, a model is trained on a dataset that consists of labelled examples, where each input is represented by a set of features and is associated with a predefined output label (Burkov, 2019). The main goal of this learning paradigm is to learn the relationship between inputs and outputs in order to use this knowledge to make accurate predictions on previously unseen data (Bhatt and Gaikwad, 2023; Yacoub et al., 2024). It is also commonly applied to classification and regression tasks, where classification involves predicting discrete categories or labels, while regression involves predicting continuous numerical values (Burkov, 2019; Gedela et al., 2024). In the field of NLP, classification is used for tasks such as sentiment analysis, spam detection, topic classification, and also sarcasm detection (Magnini et al., 2020).

In contrast to supervised learning, unsupervised learning works with unlabelled data and is focused on discovering hidden structures, relationships, or patterns

without seeing predefined target values (Bhatt and Gaikwad, 2023; Burkov, 2019). Common applications of unsupervised learning include clustering, dimensionality reduction, and anomaly detection (Bhatt and Gaikwad, 2023). At the same time, semi-supervised learning combines both labelled and unlabelled data during training. In this setting, the dataset contains a relatively small number of labelled examples and a much larger number of unlabelled examples aiming to improve model performance by using information from both types of data (Burkov, 2019).

Self-Supervised Learning (SSL) is a ML paradigm in which models learn from unlabeled data by generating supervisory signals from the input data itself instead of relying on manually annotated labels (Burkov, 2019; Gui et al., 2024). During self-supervised pre-training, the model is trained to solve automatically constructed prediction tasks, enabling it to learn meaningful representations from large amounts of unlabeled data (Gui et al., 2024). As explained by Burkov (2019), this learning paradigm is referred to as self-supervised because the training examples are derived directly from the unlabeled data. In NLP, SSL has become a popular approach for pre-training language models, ranging from early models such as BERT (Devlin et al., 2019) and GPT (Brown et al., 2020) to modern large language models (Gui et al., 2024).

Reinforcement learning differs from other learning paradigms in that an agent learns by interacting with an environment and receiving feedback in the form of rewards or penalties for its actions, with the objective of learning a policy that maximizes the total reward over time (Burkov, 2019). A modern application of reinforcement learning is Reinforcement Learning from Human Feedback (RLHF), which is used during the post-training of large language models. In RLHF, human preference data are used to train a reward model, and reinforcement learning is then applied to optimize the language model, enabling it to generate responses that align better with human preferences (Lambert, 2025).

In the context of the present study, sarcasm detection is treated as a supervised binary classification task, where each text instance is assigned one of two labels, sarcastic or non-sarcastic, and the model learns to distinguish between these categories based on patterns observed in the training data. The performance of the models can be evaluated using standard evaluation metrics, which will be discussed in Section 3.4.

3.2 From Traditional Machine Learning to Deep Learning

Common traditional ML algorithms include Decision Trees, Naïve Bayes, Logistic Regression, Random Forests, and Support Vector Machines (SVMs) and they

3.2 From Traditional Machine Learning to Deep Learning

usually rely on manually engineered features extracted from data (Pradhan et al., 2024; Gedela et al., 2024; Tiwari et al., 2024). However, in order to get the necessary features, textual information must first be converted into numerical representations that can be processed by ML algorithms (Tiwari et al., 2024). Common feature extraction techniques include tokenization, stemming, lemmatization, stopword removal, n -grams, and Term Frequency-Inverse Document Frequency (TF-IDF) representations (Tiwari et al., 2024). These features are then used by ML algorithms to perform various tasks, such as classification (Pradhan et al., 2024; Gedela et al., 2024). Traditional ML approaches often rely on manually engineered linguistic features that are specifically designed to capture lexical, semantic, and syntactic characteristics relevant to a specific task (Pradhan et al., 2024). Such handcrafted features can capture information related to negation, contrast, exaggeration, and other linguistic patterns that are used as input for ML models (Pradhan et al., 2024).

Despite their effectiveness, traditional ML methods often depend on the quality of feature engineering (Pradhan et al., 2024; Tiwari et al., 2024). Since features often need to be manually selected and extracted, these methods may miss complex patterns in the data, which has led to the adoption of Deep Learning (DL) approaches that are capable of learning feature representations directly from data (Pradhan et al., 2024; Sukhavasi et al., 2025). DL is an approach within the field of ML that uses artificial neural networks with multiple hidden layers to learn increasingly complex representations of input data (Burkov, 2019; Bhatt and Gaikwad, 2023; Alpaydin, 2014; Russell and Norvig, 2020). Unlike traditional ML approaches, DL models automatically learn relevant features during training and, eliminating the need for manual feature engineering (Burkov, 2019; Bhatt and Gaikwad, 2023; Gedela et al., 2024). Neural networks trained using DL methods, or more specifically with more than one layer between input and output, are commonly referred to as deep neural networks (Burkov, 2019).

Common deep learning architectures include Generative Adversarial Networks (GANs) (Goodfellow et al., 2014), Convolutional Neural Networks (CNNs) (LeCun et al., 1998), Recurrent Neural Networks (RNNs) (Elman, 1990), and Transformer-based models (Vaswani et al., 2017). GANs (Goodfellow et al., 2014), which presents learning as a game between two networks where a generator produces outputs and a discriminator learns to distinguish them from real data (Koehn, 2020). CNNs (LeCun et al., 1998) are particularly useful for extracting local features from sequential inputs such as text or speech. They apply convolutional kernels over the input to detect patterns in local regions, such as short sequences of words or speech signals (Jurafsky and Martin, 2026). RNNs (Elman, 1990) process input sequences step by step while maintaining a hidden state that captures information from previous time steps, which in turn allows them to model sequential dependencies

(Jurafsky and Martin, 2026). Transformer-based models (Vaswani et al., 2017) rely on self-attention mechanisms that allow each token to attend to all other tokens in the sequence and build contextual representations that capture long-range dependencies (Jurafsky and Martin, 2026). These architectures have been applied to a wide range of NLP tasks including text classification, language modeling, and sequence labeling (Bhatt and Gaikwad, 2023; Sukhavasi et al., 2025).

DL has become one of the valuable approaches in NLP because of its ability to learn complex patterns from large amounts of textual data (Yacoub et al., 2024). RNNs were among the first DL architectures to achieve strong performance on text classification tasks because of their ability to process sequential information and capture contextual dependencies between words (Yacoub et al., 2024; Russell and Norvig, 2020). However, they have largely been superseded by Transformer-based models, which provide better scalability and more effective capturing of long-range dependencies (Vaswani et al., 2017). Compared with traditional count-based text representations, such as bag-of-words and TF-IDF, which represent text as sparse vectors, DL models learn dense distributed representations in which semantically similar words have similar vector representations. This enables DL models to reduce the impact of data sparsity (Gedela et al., 2024; Burkov, 2019). The development of DL approaches in the field of NLP eventually led to the introduction of Transformer-based models and contextual representations, which are discussed in the following section (Khan et al., 2025; Russell and Norvig, 2020).

3.3 Machine Learning in Natural Language Processing

The field of NLP is an interdisciplinary field that focuses on enabling computers to process and understand human language, by applying ML methods NLP aims to improve communication between humans and computers (bin Subait et al., 2025; Russell and Norvig, 2020). NLP combines computational linguistics and machine learning techniques to perform tasks such as text classification, sentiment analysis, machine translation, question answering and also speech recognition (bin Subait et al., 2025; Russell and Norvig, 2020).

Before text can be processed by ML models to perform some task, it must be converted into a numerical representation (Burkov, 2019). One of the earliest approaches is the Bag-of-Words model, which represents a document as a vector of words from a predefined vocabulary, however fails to capture semantic relationships or word order (Burkov, 2019; Alpaydin, 2014). A common extension of this approach is TF-IDF that assigns weights to words based on their importance within a document collection, reducing the influence of very common terms across

3.3 Machine Learning in Natural Language Processing

documents and highlighting more informative words (Tiwari et al., 2024).

Traditional text representation methods were later complemented by word embeddings, which represent words as dense low-dimensional vectors learned from data, capture semantic and syntactic relationships between words and provide a richer representation of words and their similarities (Russell and Norvig, 2020). Word embeddings differ from sparse count-based representations because they use dense vectors with relatively low dimensionality, which helps improve generalization and reduces sparsity issues (Jurafsky and Martin, 2026).

Two widely used approaches for learning word embeddings are Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014). Word2Vec learns word representations by training a model to predict surrounding words from a target word by using a local context window over large text corpora (Jurafsky and Martin, 2026; Gedela et al., 2024). In contrast, GloVe builds a word-word co-occurrence matrix from the entire corpus and learns word representations from it (Jurafsky and Martin, 2026; Gedela et al., 2024).

Although Word2Vec and GloVe capture semantic similarities between words, they generate a single representation for each word regardless of context (Gedela et al., 2024; Russell and Norvig, 2020), which could be a disadvantage for words with multiple meanings. To address this limitation, researchers developed contextual embeddings, which generate different representations for the same word depending on its context within a sentence and which in turn results in a model capturing more detailed linguistic information and improving language understanding (Russell and Norvig, 2020).

A major development in the field of NLP was the introduction of Transformer-based models. The transformer architecture, proposed by (Vaswani et al., 2017), uses a self-attention mechanism that allows the model to capture relationships between words regardless of their position in a sequence (Russell and Norvig, 2020; Khan et al., 2025). Unlike earlier RNNs, Transformers process all tokens simultaneously, making them more efficient for long texts which helped them to achieve strong performance for text classification task and particularly for sarcasm detection (Khan et al., 2025; Russell and Norvig, 2020).

A widely used transformer-based model is Bidirectional Encoder Representations from Transformers (BERT), introduced by (Devlin et al., 2019). BERT generates contextual embeddings by considering both the left and right context of a word. It is pretrained by using a Masked Language Modeling (MLM), where a proportion of input tokens are randomly masked and the model learns to predict the original tokens using both left and right context, rather than predicting the next word in a sequence (Devlin et al., 2019). After pretraining, BERT-based models are fine-tuned on downstream NLP tasks by adding a neural network classifier on top of the pretrained transformer and training it on labeled data for specific tasks such

as named entity recognition or natural language inference (Jurafsky and Martin, 2026; Devlin et al., 2019).

By further building on the BERT architecture, multilingual transformer models have been developed to support cross-lingual NLP tasks, with one of the most successful being XLM-RoBERTa (Conneau et al., 2020) that was pretrained on data from 100 languages and follows the RoBERTa (Liu et al., 2019) training strategy, which instead of Next Sentence Prediction (NSP) that was used in the original BERT model and instead relies solely on the masked language modeling objective trained on much larger and more diverse datasets and demonstrated strong performance on multilingual benchmarks because of its ability to learn cross-lingual representations.

Contextual understanding is usually important for tasks that involve processing of figurative language and implicit meaning. As a result, Transformer-based models have become the foundation of many modern NLP systems and are widely used in automatic sarcasm detection research (Khan et al., 2025).

3.4 Evaluation metrics

The performance of classification models for the sarcasm detection task is commonly evaluated by using metrics such as accuracy, precision, recall, and F1-score (Burkov, 2019; Pradhan et al., 2024; Sukhavasi et al., 2025; Pradhan et al., 2024; Khan et al., 2025; Jeremy et al., 2024; Fatima et al., 2024). These metrics capture different aspects of model performance and help provide a more comprehensive evaluation of classification results. In binary classification settings, these metrics are fundamentally based on four outcomes derived from the confusion matrix: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) (Burkov, 2019).

Accuracy measures the proportion of correctly classified instances among all predictions made by the model and it is calculated by dividing all correct predictions by the total number of predictions (Burkov, 2019; Bharati et al., 2024). Manning et al. (2008) also note that this evaluation metric is especially suitable for classification problems, however it may be less informative when the classes are imbalanced.

Other frequently used evaluation metrics are precision and recall, where precision measures the proportion of segments predicted as belonging to a particular class and that actually belong to that class (Burkov, 2019; Manning et al., 2008; Alpaydin, 2014). Precision is calculated as TP divided by the sum of TP and FP (Burkov, 2019). For instance, in sarcasm detection, precision indicates how many texts predicted as sarcastic are actually sarcastic. Recall measures the proportion of actual examples of a class that are correctly identified by the model (Burkov, 2019; Manning et al., 2008; Alpaydin, 2014). Recall is calculated as TP divided by the

sum of TP and FN (Burkov, 2019). In the context of sarcasm detection, recall shows how many sarcastic texts the model is able to find correctly.

F1-score combines precision and recall into a single metric by calculating their harmonic mean and is commonly used to evaluate classification performance using a single value (Manning et al., 2008). This evaluation metric is frequently used in automatic sarcasm detection studies, and higher F1-scores indicate better overall performance of a classification model, particularly for sarcasm detection (Pradhan et al., 2024; Helal et al., 2024).

For more detailed analysis and also visualization of classification tasks, the confusion matrix is commonly used (Burkov, 2019; Russell and Norvig, 2020; Manning et al., 2008). A confusion matrix is presented in the form of a table and it compares how often each category is classified and misclassified as the opposite category to summarize model performance (Burkov, 2019; Russell and Norvig, 2020). In case of binary classification, the confusion matrix consists of TP, TN, FP, FN, which were introduced earlier (Burkov, 2019). These values show how many instances were classified correctly or incorrectly. In case of multiclass classification, the confusion matrix can contain as many rows and columns as there are classes (Burkov, 2019; Manning et al., 2008), which helps to identify patterns of misclassification and to reveal which classes are most frequently confused with one another (Burkov, 2019). Precision and recall can also be calculated from the confusion matrix (Burkov, 2019). Overall, confusion matrices provide a useful visual representation of classifier performance and are commonly used alongside quantitative metrics to evaluate classification systems (Burkov, 2019; Russell and Norvig, 2020; Manning et al., 2008).

The proposed method in this master's thesis uses accuracy, precision, recall, F1-score, and confusion matrices to evaluate and visualize performance of the trained classification models. Particular attention is paid to F1-score that provides a balanced assessment of model performance, while confusion matrices are used to visualize results and analyze patterns of correct and incorrect classifications across classes.

4 Approaches to Machine Translation

Machine Translation (MT) has evolved significantly over the past decades, progressing from rule-based systems to statistical and neural approaches. Modern Neural Machine Translation (NMT) systems have improved translation quality and are widely used today. This chapter introduces the main principles of NMT, including encoder-decoder framework, attention mechanisms, and Transformer models, as well as contemporary multilingual MT model such as No Language Left Behind (NLLB) (Team et al., 2022). It also discusses approaches to machine translation evaluation and highlights the key challenges that continue to affect translation quality and system performance.

4.1 Evolution of Machine Translation Approaches

MT research began in the 1940s and has since evolved significantly over the past decades, progressing from rule-based systems to statistical and neural approaches (Koehn, 2020). However, research progress slowed after the release of ALPAC Report in 1966 that concluded that MT systems did not yet provide sufficient quality or economic advantages in comparison to human translation (Koehn, 2020). Rule-Based Machine Translation (RBMT), the dominant machine translation paradigm until the early 2000s, relied on manually created dictionaries and grammatical rules in order to generate translations by mapping structures and vocabulary from the source language to the target language (Koehn, 2020; Kenny, 2022; Moorkens et al., 2025). While successful in restricted domains and practical applications, RBMT required extensive expert effort, was difficult to scale, and struggled with the complexity of real-world language (Moorkens et al., 2025; Kenny, 2022).

The limitations of RBMT encouraged the development of corpus-based approaches that learned translation patterns from large collections of parallel texts rather than relying on manually crafted linguistic rules. This led to the development of Statistical Machine Translation (SMT), which became the dominant MT paradigm throughout the 2000s. SMT includes both word-based and phrase-based approaches and relies on a translation model, which learns translation probabilities from parallel corpora, and a language model, which helps produce fluent transla-

tions. Unlike RBMT, SMT learns translation patterns automatically from data rather than relying on manually crafted linguistic rules (Koehn, 2020; Kenny, 2022; Moorkens et al., 2025).

However, SMT often struggled with complex languages and was prone to errors, such as word omission, and inconsistent translations of the same source-language word, sometimes even within a single sentence (Kenny, 2022). To address some of these limitations, researchers incorporated linguistic information into SMT systems, and (Goldwater and McClosky) (2005) showed that morphological analysis, including lemmatization and morphological tagging, improved Czech-English translation quality by reducing data sparsity, particularly for highly inflected languages.

The next step in the evolution of MT was NMT, which became practical due to advances in DL and computing power and translates text using a neural network model rather than multiple separate components as in SMT (Koehn, 2020; Kenny, 2022). The introduction of RNNs, Sequence-to-Sequence models, and attention mechanisms improved translation quality, allowing NMT to surpass SMT in most evaluation tasks and become the dominant machine translation approach by the mid-2010s (Koehn, 2020; Kenny, 2022; Moorkens et al., 2025).

Transformer architectures, which rely on attention mechanisms and enable more efficient parallel processing of linguistic information, further advanced MT and now form the basis of modern MT systems, further advancing multilingual language technologies (Koehn, 2020; Moorkens et al., 2025). The connection between NMT and Transformer architecture will be further discussed in the next section.

Transformer architectures, which rely on attention mechanisms and enable more efficient parallel processing of linguistic information, further advanced MT and now form the basis of modern MT systems and multilingual language technologies (Koehn, 2020; Moorkens et al., 2025). More recently, Large Language Models (LLMs), which are also based on the Transformer architecture, have been applied to MT. These models are pretrained on very large text corpora and can perform translation as one of many language tasks, even though they are not specifically designed for MT (Moorkens et al., 2025). The relationship between NMT and the Transformer architecture is discussed in more detail in the next section.

4.2 Neural Machine Translation and Transformers

NMT became the dominant MT paradigm after 2016 when it replaced Statistical Machine Translation (SMT) as the state-of-the-art approach (Johnson et al., 2016; Koehn, 2020; Kenny, 2022). Unlike SMT, which relies on separately trained statistical components such as translation and language models trained on parallel corpora, NMT uses artificial neural networks that are composed of interconnected neurons. By considering the entire source sentence, NMT creates contextualized

word representations and captures complex linguistic relationships, producing more fluent and natural translations (Koehn, 2020; Kenny, 2022).

Early NMT systems were typically based on RNNs and followed an encoder-decoder framework, in which the encoder converts the source sentence into a vector representation, while the decoder generates the target sentence word by word from this representation (Koehn, 2020). Although those early encoder-decoder models performed well on short sentences, they often struggled with longer inputs because of the limitations of fixed-length sentence representations, a problem that was improved by the introduction of the attention mechanism, which allowed models to consider relevant source words as contextual clues when selecting target-language words and significantly improved translation quality (Koehn, 2020; Moorkens et al., 2025).

Building on this idea, Vaswani et al. (2017) proposed the Transformer, an encoder-decoder architecture that uses multi-head self-attention mechanisms to capture contextual information. This architecture replaced recurrent processing with self-attention and became the foundation of modern MT systems, enabling more efficient parallel computation while better capturing long-range dependencies (Koehn, 2020; Moorkens et al., 2025). Transformers model relationships between all words in a sequence simultaneously through self-attention, producing richer contextual representations and improving the handling of complex linguistic structures (Koehn, 2020; Moorkens et al., 2025). Due to these advantages, Transformers became the foundation of modern machine translation systems and are now widely used across a variety of NLP tasks (Koehn, 2020; Moorkens et al., 2025).

The Transformer follows an encoder-decoder architecture composed of multiple stacked encoder and decoder layers. Each encoder layer consists of a multi-head self-attention mechanism followed by a position-wise feed-forward network, which together produce contextualized representations of the input sequence. The decoder layers have a similar structure but additionally include masked self-attention and encoder-decoder attention, which allows the model to generate the target sequence autoregressively while attending to relevant parts of the source sentence (Vaswani et al., 2017; Moorkens et al., 2025).

An important example of a Transformer-based multilingual NMT system is NLLB (Team et al., 2022), which was designed with the objective of improving MT for low-resource languages, and its NLLB-200 version supports 200 languages, including many that were previously underrepresented in MT research (Buscaldi and Rosso, 2023). This makes NLLB particularly relevant for MT research, even though the research revealed a significant underperformance on endangered languages, such as Ligurian (Buscaldi and Rosso, 2023).

Overall, the development of NMT from early RNN-based encoder-decoder models to Transformer architectures has significantly improved translation quality through

better contextual understanding and the processing of long-range dependencies. As a result, Transformer-based systems now form the basis of most modern MT systems and multilingual language technologies (Koehn, 2020; Moorkens et al., 2025; Kenny, 2022).

4.3 Evaluation of Machine Translation

Evaluation of MT quality is an important area of MT research because, despite their considerable progress, MT systems still can produce errors from minor inaccuracies to serious mistranslations, prompting further research into quality assessment, error analysis, and methods for improving MT output (Kenny, 2022). MT evaluation can be broadly divided into human and automatic evaluation. Although human evaluation enables the assessment of translation quality in terms of adequacy, fluency, and meaning preservation, it can be time-consuming and resource-intensive, which has motivated further development of automatic evaluation metrics (Rossi and Carré, 2022).

One of the most commonly used automatic metrics is Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002). BLEU measures the overlap of n -grams between a machine-generated translation and one or more reference translations by calculating the precision of matching n -grams, usually from unigrams to 4-grams, and by capturing lexical similarity and aspects of word order (Rossi and Carré, 2022). BLEU also incorporates a brevity penalty in order to discourage excessively short translations, and is calculated at the corpus level rather than for individual sentences (Rossi and Carré, 2022).

Despite the development of many newer metrics, BLEU remains widely used in MT evaluation research, although its scores can be influenced by language typology, particularly in highly inflected languages where minor differences in word forms may result in lower scores (Lo et al., 2023; Koehn, 2020). However, more recent MT literature shows that BLEU is increasingly complemented or replaced by learned metrics such as COMET, which tend to correlate more strongly with human judgments (Rei et al., 2020b).

To overcome such limitations, embedding-based metrics have been developed such as BERTScore (Zhang et al., 2019), that relies on contextual embeddings from BERT-based pretrained language models, and COMET (Rei et al., 2020a), an XLM-R-based evaluation framework trained to predict human quality assessments from source-translation-reference tuples (Lo et al., 2023). By incorporating the source sentence alongside the translation and reference, COMET can better capture semantic adequacy and meaning preservation. This framework shows strong correlation with human judgments, demonstrating the effectiveness of learned evaluation metrics for MT quality assessment (Rei et al., 2020b).

Recent research has also explored more advanced approaches to translation Quality Estimation (QE) that aims to assess translation quality without requiring a reference translation. Qian et al. (2024) showed that by jointly training sentence-level and word-level QE tasks within a multi-task learning framework improves performance in comparison to separate training. Their findings further indicate that sentence-level QE can support emotion classification, and particularly in settings when training data is limited and imbalanced.

4.4 Challenges of Machine Translation

Despite the significant progress achieved in NMT and more recently in LLMs, several challenges still continue to affect translation quality, reliability, and practical applicability. While earlier NMT research mainly focused more on improving model architectures and addressing limitations related to context representation, data availability, and translation accuracy, current research shows that LLMs have shifted the nature of these challenges toward issues of adaptation, reliability, controllability, and evaluation (Koehn and Knowles, 2017; Koehn, 2020; Pang et al., 2025; Budimir, 2025; Polakova et al., 2026).

Wan et al. (2022) categorized common translation error types in NMT into over-translation, under-translation, mistranslation, and disorder. Their findings show that shorter texts are especially prone to mistranslation errors due to the limited amount of contextual and semantic information available. Furthermore, the authors demonstrate that an imbalanced distribution of training data across different sentence-length categories can contribute to over-translation errors, which subsequently reduce overall translation quality. These findings highlighted the importance of sufficient context and balanced training data for achieving reliable MT performance.

With the emergence of LLM-based MT, several long-standing limitations of traditional NMT systems have been mitigated. Pang et al. (2025) demonstrate in their study that LLMs require less parallel training data than conventional encoder-decoder models and achieve strong performance in translating long sentences and document-level contexts. Nevertheless, they continue to experience difficulties with domain mismatch and the accurate translation of rare and domain-specific terminology. In addition to these challenges, LLM-based MT also introduces new issues, including the imbalance of multilingual pretraining resources, high inference costs, and the need for improved evaluation methods that better reflect human judgments, since the existing metrics do not always match with human assessments of translation quality.

Culture-specific items (CSIs) also remain particularly challenging for contemporary MT systems. Budimir (2025) demonstrate that, although recent LLM-based

4 Approaches to Machine Translation

models generally outperform the evaluated NMT system in translating culturally bound expressions, they still struggle to achieve consistent cultural adaptation, particularly when translating proper names and other CSIs in morphologically rich, low-resource languages. The authors show that MT outputs often fail to preserve the communicative and cultural functions of CSIs and also highlight that human translators continue to play an essential role in culturally sensitive and literary translation.

Similar challenges are observed in the translation of specialized educational content. Polakova et al. (2026) show that, despite substantial improvements achieved by LLM-based systems over pre-LLM MT systems, educational materials remain difficult to translate because they combine domain-specific terminology, interactive exercises, and multimodal elements. In addition to lexical and terminological errors, the authors identify several error typical for LLM-based translation, including hallucinations, unintended correction of counterfactual statements, and completion of unfinished text. These findings suggest that improving MT for educational content requires not only advances in translation models but also specific solutions for terminology translation, context-dependent disambiguation, and the preservation of structured markup.

Overall, recent research suggests that the main challenges of MT have shifted from improving model architectures to ensuring reliable, context-aware, and application-specific translation. Although LLM-based systems have significantly advanced translation quality, they still continue to face difficulties related to domain adaptation, specialized terminology, and the evaluation of translation quality. Moreover, the introduction of LLMs has brought new challenges, such as hallucinations and context-related limitations, which also highlight the continued importance of human expertise and task-specific adaptation in order to achieve stable translation performance across different languages and domains.

5 Sarcasm Detection Approaches

This chapter provides an overview of the main approaches to automatic sarcasm detection. It discusses the challenges of sarcasm detection and reviews traditional feature-based methods as well as deep learning approaches, including recurrent neural networks and Transformer-based models.

5.1 Automatic Sarcasm Detection

Automatic sarcasm detection is considered a challenging task in the field of NLP because sarcastic expressions are often subtle and implicit and due to absence of paralinguistic cues (Pradhan et al., 2024; Khan et al., 2025). Additional challenges include the wide range of emotions associated with sarcastic communication, domain-specific interpretations, multilingualism, and the informal and unstructured language in online texts (Tiwarei et al., 2024). Unlike regular sentences, sarcasm usually shows a mismatch between the words used and the speaker’s intended message, which has to be deducted from context rather than from the literal meaning of the sentence (Ducret et al., 2020). For this reason, models that do not take context into account often perform poorly, as understanding sarcasm requires additional information beyond the sentence itself (Helal et al., 2024).

The importance of context has been further highlighted in several studies. Ducret et al. (2020) show that including contextual information, such as preceding or following tweets, significantly improves model performance. Their experiments demonstrate that classifiers considering both the target text and its context achieve higher F1 scores. Similarly, Helal et al. (2024) report that adding dialogue history or metadata helps models improve accuracy and generalizability across different datasets. Tiwarei et al. (2024) also found that contextual approaches outperform lexical and pragmatic approaches for sarcasm detection, highlighting the importance of contextual incongruity in identifying sarcastic expressions. Additionally, Nimase and Hong (2024) demonstrated that richer contextual representations, particularly sentence-level embeddings and combinations of multiple contextual embeddings, consistently improve sarcasm detection performance.

Given the complexity of sarcastic language, a variety of computational approaches have been proposed for its automatic detection. These approaches range from traditional feature-based ML methods to deep learning models. The following sections

first discuss feature-based approaches and then neural approaches, beginning with recurrent architectures and concluding with more recent Transformer-based models.

5.2 Feature-Based Approaches

Feature-based approaches represent one of the traditional methods for automatic sarcasm detection and rely on manually engineered linguistic features that are used as input for machine learning classifiers rather than being learned automatically from data (Ducret et al., 2020; Pradhan et al., 2024). Traditional machine learning algorithms such as Random Forests, Decision Trees, Support Vector Machines (SVM), Logistic Regression, and Gradient Boosting are commonly applied for sarcasm detection tasks (Ducret et al., 2020; Pradhan et al., 2024; Tiwari et al., 2024).

Several studies have proposed different types of handcrafted features for sarcasm detection (Ducret et al., 2020; Tiwari et al., 2024). In their research (Tiwari et al., 2024) distinguish between lexical, pragmatic, and contextual features. Lexical features include n -grams and part-of-speech information, pragmatic features capture stylistic cues such as capitalization, repeated characters, and emojis, while for contextual features model inconsistencies between the literal meaning of a statement and its broader context. These features are then used as input for SVM classifiers, with contextual features producing the strongest results, highlighting the potential importance of contextual information for sarcasm detection (Tiwari et al., 2024).

Similarly, (Ducret et al., 2020) grouped sarcasm-related features into complexity, stylistic, and psychological categories. In addition to TF-IDF and n -grams, the authors incorporated sentiment and emotion-related information, including Valence, Arousal, Dominance (VAD) scores (Warriner et al., 2013) and Linguistic Inquiry and Word Count features (LIWC) (Tausczik and Pennebaker, 2010). Their experiments demonstrated that contextual information significantly improves sarcasm detection, since sarcastic texts often exhibit emotional or semantic incongruity with their preceding context. The highest performance was achieved by using a Random Forest classifier with a combination of all feature types and contextual information (Ducret et al., 2020).

Further evidence for the effectiveness of feature-based approaches is provided by (Pradhan et al., 2024), who proposed a sarcasm detection framework based on 13 handcrafted features capturing text meaning, lexical diversity, readability, and grammatical composition. These features were used to train several traditional machine learning models, including Random Forest, Decision Tree, Gradient Boosting, SVM, and Logistic Regression classifiers. The authors also reported that an ensemble model built on combination of linguistic features and best performing classifiers, which included Gradient Boosting, Decision Tree and Random Forest,

outperformed several deep learning baselines, demonstrating that carefully designed linguistic features can be effective for sarcasm detection tasks (Pradhan et al., 2024).

Overall, feature-based approaches demonstrate the usefulness of manually engineered features for sarcasm detection. The reviewed studies used a wide range of linguistic, stylistic, psychological, pragmatic, and contextual features, which were successfully used by traditional ML classifiers for sarcasm detection (Ducret et al., 2020; Pradhan et al., 2024; Tiwari et al., 2024).

5.3 Deep Learning Approaches

Deep learning approaches became important in sarcasm detection because they can automatically learn semantic and contextual patterns from textual data without relying on manually engineered features (Yacoub et al., 2024; Sukhavasi et al., 2025). Tiwari et al. (2024) note that DL approaches generally outperform earlier methods when sufficient training data are available because they can automatically learn relevant linguistic patterns from data, although feature engineering remains valuable in some settings (Tiwari et al., 2024). Similarly, Sukhavasi et al. (2025) argue that traditional ML methods often suffer from limited learning capability, lower accuracy, and difficulties in handling complex sarcastic expressions, which has promoted the application of DL models for sarcasm detection.

Among the most widely used deep learning architectures for sarcasm detection are RNNs, CNNs, Long Short-Term Memory networks (LSTMs), and Bidirectional LSTMs (Bi-LSTMs) (Sukhavasi et al., 2025; Yacoub et al., 2024; Khan et al., 2025). According to Yacoub et al. (2024), DL models, particularly RNNs, are effective for text classification due to their flexibility and ability to model complex data structures. Recent studies have successfully applied a variety of neural architectures, including CNNs, LSTMs, and Bi-LSTMs, to sarcasm detection tasks (Yacoub et al., 2024; Sukhavasi et al., 2025).

Recent research has also explored ensemble architectures that consist of different model combinations. Gedela et al. (2024) proposed a voting-based ensemble framework incorporating CNN, Bi-LSTMs, hybrid CNN-BiLSTM architectures, and BERT embeddings. The best performance was achieved by a sequential CNN-BiLSTM model built on top of BERT embeddings and combined with an SVM classifier. Additionally, Sukhavasi et al. (2025) proposed an optimized Bi-LSTMs model combined with a Group Teaching Optimization (GTO) algorithm and reported superior performance in comparison to other DL approaches.

Another important direction in sarcasm detection research is the application of Transformer-based models. As discussed in the previous chapter, Transformer architectures rely on self-attention mechanisms that enable them to capture contextual

5 Sarcasm Detection Approaches

information more efficiently (Vaswani et al., 2017; Jurafsky and Martin, 2026). One example is UrduSarcasmNet, proposed by Khan et al. (2025), a transformer-based architecture that combines cascaded and multi-grouped attention mechanisms to capture both local and global contextual information. The proposed model demonstrated improved sarcasm detection performance in the low-resource Urdu language, while ablation studies showed that these attention components contribute substantially to the detection of subtle sarcastic patterns (Khan et al., 2025).

The growing importance of transformer-based approaches is also highlighted by Yacoub et al. (2024), who argue that advances in transformer architectures are expected to improve the detection of implicit sarcasm and support the development of models capable of identifying sarcastic content without explicit indicators. They also emphasize the increasing importance of transfer learning, external knowledge integration, and pre-trained language models such as BERT (Devlin et al., 2019) for future sarcasm detection research (Yacoub et al., 2024). Transfer learning is a method that adapts a model pre-trained on large-scale corpora to a specific task by fine-tuning it on a smaller dataset, which enables models to achieve strong performance even when only limited annotated data are available (Burkov, 2019).

Gregory et al. (2020) compared several DL architectures for sarcasm detection, including LSTM, GRU, and transformer-based models. The study incorporated contextual information and experimented with pre-trained transformers such as BERT, RoBERTa (Liu et al., 2019), XLNet (Yang et al., 2019), and ALBERT (Lan et al., 2019). The results showed that transformer representations captured the most relevant information for sarcasm detection and consistently outperformed LSTM-based models. The highest performance was achieved by an ensemble of transformer models (Gregory et al., 2020).

However, despite their strong performance, transformer models typically require substantial computational resources and large amounts of training data, which may limit their applicability in some scenarios (Khan et al., 2025). Overall, the evolution from recurrent neural architectures to Transformer-based models has substantially improved sarcasm detection by enabling models to learn richer semantic and contextual representations. Nevertheless, important challenges remain for multilingual sarcasm detection, including cross-lingual generalization and cultural variation, as discussed in the following section.

5.4 Challenges of Multilingual Sarcasm Detection

Despite recent progress, sarcasm detection research still has important limitations related to language coverage. Existing studies are more often conducted in monolingual settings, where a single model is trained and evaluated on data from one language (Гурийн and Жуков, 2023; Khan et al., 2025; Tasnim and Nath, 2024).

5.4 Challenges of Multilingual Sarcasm Detection

By focusing on one language, these settings may not consider linguistic and cultural differences that influence how sarcasm is expressed across languages (Tasnim and Nath, 2024). Tasnim and Nath (2024) highlight the potential of multilingual approaches for cross-language sarcasm detection. However, their findings also suggest that performance may still vary across languages, potentially due to differences in language-specific model optimization, dataset characteristics, preprocessing quality, and cultural variation.

An additional challenge of multilingual sarcasm detection is that sarcastic expressions often rely on language-specific linguistic and pragmatic cues that cannot always be transferred across languages. Irony markers, idiomatic expressions, humor conventions, and culturally shared knowledge may differ substantially between languages, making it difficult for multilingual models to recognize sarcastic intent consistently. As a result, multilingual models trained on one language may not generalize equally well to another language, even when based on shared multilingual representations (Tasnim and Nath, 2024; bin Subait et al., 2025). These challenges become particularly important in cross-lingual applications, where differences in linguistic structure and cultural background may influence both sarcasm interpretation and model performance.

Another challenge concerns the complexity of sarcasm itself. Unlike literal language, sarcastic expressions often involve a mismatch between the surface meaning of words and the speaker’s intended message, making them difficult to identify automatically (Gedela et al., 2024). Contextual information is often necessary for accurate sarcasm detection, for instance in social media interactions where sarcastic posts may depend on preceding messages to fully understand the message (Sukhavasi et al., 2025). Additional challenges arise from various culturally specific forms of sarcasm that may differ across languages and dialects (bin Subait et al., 2025).

Data availability and annotation quality also remain important limitations. Existing studies report a lack of large, high-quality sarcasm datasets for many languages, particularly low-resource languages (bin Subait et al., 2025), as well as precisely labeled datasets for more efficient classifier training (Sukhavasi et al., 2025). In addition, datasets collected from social media platforms may often contain noisy labels and informal language, which can negatively affect model performance and generalization (Sukhavasi et al., 2025). Finally, recent research has identified other directions that remain insufficiently explored, including cross-domain generalization and the incorporation of multimodal information such as emojis and images that frequently accompany sarcastic content in social media (bin Subait et al., 2025).

Overall, multilingual sarcasm detection remains an open research challenge despite recent advances in multilingual language models. Existing approaches continue to face difficulties related to data availability, contextual dependencies, cultural

5 Sarcasm Detection Approaches

variation, and model generalization across languages and domains (Sukhavasi et al., 2025; bin Subait et al., 2025). Addressing these limitations is essential for improving the robustness and generalizability of multilingual sarcasm detection systems.

6 Sarcasm and Machine Translation

This chapter examines the relationship between sarcasm and MT. It discusses the main challenges of translating sarcastic content, the difficulties involved in evaluating machine-translated sarcasm, and the potential impact of MT on automatic sarcasm detection.

6.1 Challenges of Sarcasm Translation

Figurative and emotionally loaded user generated content (UGC) remains one of the major challenges in the field of MT. This type of content, especially social media posts, is characterized by emotions, slang, cultural references, and non-literal expressions such as irony and sarcasm (Saadany et al., 2023; Qian et al., 2023, 2024). Previous studies show that MT systems often fail to translate such content and that standard evaluation metrics often do not capture errors related to meaning shifts that affect interpretation and speaker’s original intent (Saadany et al., 2023; Qian et al., 2024).

Recent research of UGC demonstrates that emotion-carrying words, such as slang, and culture-specific expressions are among the most frequent sources of MT errors (Qian et al., 2023). These findings are also relevant to sarcasm, which often depends on subtle emotional cues and a mismatch between literal form and intended meaning (Pradhan et al., 2024; Helal et al., 2024; Zawawi and Maghfiroh, 2020). While professional human translators are able to choose between the appropriate translation techniques and could eventually produce texts that differ from each other for UGC, such an option is not available for MT that could not produce good solutions for UGC (Qian et al., 2023).

Sarcasm is particularly difficult to translate because its interpretation often depends on contextual incongruity, shared background knowledge, and cultural information rather than on the literal meaning of words (Tiwari et al., 2024; Zawawi and Maghfiroh, 2020). In addition, sarcastic meaning in spoken communication is frequently supported by paralinguistic cues such as intonation and facial expressions, which are absent in written text and therefore unavailable to MT systems (Pradhan et al., 2024; Helal et al., 2024).

Cross-lingual differences could also complicate the translation of sarcasm, since languages differ in their structure and richness, and sarcastic expressions sometimes

lack direct equivalents in the target language. Additionally, sarcastic meaning often depends on culturally specific or contextual cues (Tiwari et al., 2024; Zawawi and Maghfiroh, 2020; Khan et al., 2025), which MT systems may fail to transfer, leading to weakening or loss of sarcasm.

6.2 Evaluation of Machine-Translated Sarcasm

From an evaluation perspective, Qian et al. (2023) show that MT evaluation frameworks mainly operate at the sentence level and do not explicitly consider emotional information, which is often important for translating pragmatic phenomena such as sarcasm. MT output may appear fluent and semantically correct while failing to transfer the original intended meaning (Saadany et al., 2023).

This limitation is particularly relevant for automatic evaluation metrics such as BLEU, which primarily measure lexical overlap between MT output and reference translations rather than the preservation of pragmatic meaning (Rossi and Carré, 2022). As a result, a translation may receive a relatively high score despite failing to preserve the sarcastic intent of the source text. Consequently, semantic adequacy does not necessarily imply pragmatic equivalence, especially in the case of figurative language such as sarcasm.

More recent neural evaluation metrics, such as COMET, use contextual representations of the source and target texts to estimate translation quality and generally provide evaluation results that are closer to human judgments than lexical overlap metrics such as BLEU (Rei et al., 2020a; Moorkens et al., 2025). Later versions of the COMET introduced reference-free QE models that function without reference translations, which in turn enables translation QE in scenarios where reference translations are unavailable (Rei et al., 2020b).

Despite these advances, automatic evaluation metrics sometimes struggle to fully assess whether pragmatic phenomena such as sarcasm have been successfully preserved. For this reason, MT research continues to combine automatic evaluation with human assessment (Moorkens et al., 2025; Jurafsky and Martin, 2026). Human evaluation traditionally relies on the adequacy metric, which evaluates how well a translation preserves the original meaning and intent of the source text, and the fluency metric, which evaluates how natural and grammatically correct the translation is in the target language (Moorkens et al., 2025; Jurafsky and Martin, 2026). In the case of sarcasm, adequacy is particularly important because some translation may be fluent and semantically accurate but still can fail to preserve the speaker’s intended sarcastic meaning. Therefore, the evaluation of machine-translated sarcasm needs to consider not only overall translation quality but also whether the intended sarcastic meaning of the original text has been preserved.

6.3 Impact of Machine Translation on Sarcasm Detection

Previously mentioned challenges may potentially have implications for downstream NLP tasks. Changes in the text introduced by MT can negatively affect automatic sarcasm detection, as translated texts may no longer contain the cues on which detection models usually rely.

Evidence from related NLP tasks supports this assumption. For example, [Dmonte et al. \(2025\)](#) investigated the effect of MT on offensive language detection and found that offensive language identification models generally perform better on original texts than on machine-translated content. The authors connect this performance drop with translation errors and changes in linguistic features that were introduced during MT, which can alter the information required for successful offensive language identification.

Automatic sarcasm detection models typically rely on lexical, emotional, pragmatic, and contextual cues to identify sarcastic expressions ([Ducret et al., 2020](#); [Pradhan et al., 2024](#); [Tiwari et al., 2024](#)). Recent research further demonstrates that sarcasm recognition performance can be improved by using richer contextual information, which highlights the importance of preserving contextual information for accurate classification ([Nimase and Hong, 2024](#)). Consequently, if these lexical, emotional, contextual, and pragmatic cues are altered or modified during MT, the translated text may no longer contain sufficient information for accurate sarcasm detection.

This master's thesis investigates how MT affects the preservation of sarcastic meaning and how this further influences the performance of sarcasm detection models. Although recent research has made progress in multilingual sarcasm detection and translation of figurative language, studies that specifically examine the interaction between MT and automatic sarcasm detection remain limited. Existing work has primarily investigated these topics separately, leaving the influence of MT on downstream sarcasm detection underexplored. To the best of my knowledge, no detailed study has yet investigated this issue for the English-Russian language pair, making the present research a relevant contribution at the intersection of Translation Studies (TS) and NLP.

7 Related Work

Sarcasm has been studied from multiple perspectives, including linguistic analysis, translation studies, and computational approaches for sarcasm detection. [Zawawi and Maghfiroh \(2020\)](#) examine sarcasm from a translation perspective by analyzing sarcastic expressions in Mark Manson’s *The Subtle Art of Not Giving A Fuck* and their translation techniques to Bahasa Indonesia. Authors analyze 145 sarcastic sentences by using a qualitative descriptive approach and identify four main types of sarcasm in the source text: ridicule, satire, proximity, and humor. Their findings show that translators rely on various techniques, such as compensation, adaptation, modulation, and transposition, to preserve sarcastic meaning in translation. The results suggest that the successful transfer of sarcasm largely depends on the appropriate selection of translation techniques and careful consideration of contextual meaning.

In contrast to qualitative linguistic analyses, sarcasm detection in social media has also been studied using computational methods. For example, [Ducret et al. \(2020\)](#) investigate linguistic and contextual features used for sarcasm detection in Twitter posts. The authors use a Twitter Corpus ([Ghosh et al., 2020](#)) consisting of 5,000 sarcastic and non-sarcastic tweets and analyze stylistic, complexity, and psychological features, as well as sentiment and emotion related indicators. Their experiments with a Random Forest classifier demonstrate that incorporating conversational context significantly improves performance, achieving F1-score of approximately 70%. The study also shows that sarcastic tweets often exhibit emotional or semantic incongruity with their surrounding context.

Recent research has also explored the problem of sarcasm detection across multiple languages. [Tasnim and Nath \(2024\)](#) investigate cross-language sarcasm detection using English ([Misra and Arora, 2023](#)) and Bangla ([Lora et al., 2024](#)) datasets and evaluate 27 model configurations using multilingual models such as XLM-RoBERTa ([Conneau et al., 2020](#)), mBERT ([Devlin et al., 2019](#)), and mT5 ([Xue et al., 2021](#)) with different classifiers, including SVM ([Cortes and Vapnik, 1995](#)), Random Forest ([Ho, 1995](#)), and LSTM ([Hochreiter and Schmidhuber, 1997](#)). Their results reveal no statistically significant difference in performance between languages with respect to the F1-score, according to the Wilcoxon test ([Wilcoxon, 1945](#)), while accuracy shows a significant difference, with all evaluated models achieving higher accuracy on the English dataset than on the Bangla dataset. This suggests that overall sarcasm detection ability may remain relatively consistent across languages,

7 Related Work

although performance can still vary depending on specific factors. However, authors emphasize that further research is needed to assess the applicability of these findings to other languages.

Recent studies have also investigated the effectiveness of transformer-based language models for sarcasm detection across different domains. Oprea and Băra (2025) conduct a systematic comparative study of four transformer-based models, namely RoBERTa-large (Liu et al., 2019), RoBERTa-base (Liu et al., 2019), DistilBERT-base-uncased (Sanh et al., 2019), and DistilBERT-SST2 (Sanh et al., 2019), using supervised fine-tuning on two publicly available datasets containing sarcastic news headlines and Amazon product reviews. Their methodology includes group-aware data splitting, label smoothing, macro-F1-based early stopping, and cross-domain evaluation to assess model robustness and cross-domain generalization. Their results show that RoBERTa-base achieves the highest macro F1-score of 0.943 on the news headline dataset, while DistilBERT-SST2 demonstrates the strongest cross-domain performance, reaching a macro F1-score of 0.878 on the Amazon review dataset and shows the most consistent performance on unseen sarcastic texts despite its smaller size. The authors conclude that domain-aware fine-tuning and appropriate model selection are important for effective sarcasm detection.

Other studies have focused on benchmarking transformer-based models for sarcasm detection in specific languages. Abu Farha and Magdy (2021) conduct a comprehensive evaluation of transformer-based language models for Arabic sentiment analysis and sarcasm detection using the ArSarcasm-v2 dataset of annotated tweets (Abu Farha et al., 2021). The authors compare 24 models, including multilingual models such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) as well as Arabic-specific models. Their results demonstrate that monolingual Arabic models outperform multilingual ones, particularly when trained on large datasets containing social media text. Among the evaluated models, MARBERT (Abdul-Mageed et al., 2021), which was trained on a 128 GB corpus, half of which consisted of tweets, achieves the highest F1-score of 0.584 on the sarcasm detection task, slightly outperforming the next best Arabic-specific model, AraBERT-large (v2) (Antoun et al., 2020), which achieves an F1-score of 0.579. The authors attribute MARBERT’s performance to the diversity of its training data, which enables the model to better handle variations in dialectal Arabic.

Similarly, Гурин and Жуков (2023) investigate sarcasm detection in Russian Twitter posts. Their research combines sentiment analysis with sarcasm detection and compares a convolutional neural network using Word2Vec embeddings with the Transformer-based model RuBERT (Kuratov and Arkhipov, 2019). For sentiment classification, the RuBERT-based model achieves the highest accuracy of 79.28%, outperforming the CNN using Word2Vec embeddings, which reaches 74.23% accuracy. A separate sarcasm detection model achieves an accuracy of 69.19%,

highlighting the greater difficulty of sarcasm detection. The authors conclude that sarcasm detection in Russian remains challenging due to contextual ambiguity and linguistic variability, emphasizing the need for larger manually annotated datasets.

Overall, these studies show that sarcasm has been explored from both linguistic and computational perspectives, highlighting its complexity and context-dependence. While computational approaches have achieved promising results, sarcasm detection remains challenging, particularly across languages and contexts.

8 Methodology

This thesis presents an experimental approach to investigate whether sarcasm can be preserved after being translated by MT systems and then recognized by sarcasm detection models. The study uses an English (Misra and Arora, 2023) and a Russian dataset (Kostya165, 2023). The English dataset is translated into Russian using MT, and sarcasm detection models are applied to both original and translated data to analyze the effect of translation on sarcastic meaning.

Firstly, a multilingual transformer-based model XLM-R (Conneau et al., 2020) will be fine-tuned for sarcasm detection by using annotated English and Russian datasets. The following training setups will be applied, including fine-tuning on each language separately, fine-tuning a combined English-Russian dataset, and also sequential fine-tuning where the model will be first trained on English dataset and then on Russian. Model performance will be evaluated using such classification metrics as precision, recall, and F1-score in order to select the best configuration.

Secondly, the English sarcasm dataset will be translated into Russian using the MT model NLLB (Team et al., 2022). The translated dataset will still have the original sarcasm annotations that would allow a direct comparison between original and translated segments. The sarcasm detection model that performed best will then be applied to the data translated into Russian to examine how translation affects sarcasm detection performance.

Finally, a qualitative analysis will be performed for the translated data, focusing on cases where sarcasm is correctly or incorrectly detected after translation. This analysis will examine which translation techniques were used in the MT output for translation of sarcastic expressions. This approach would allow to analyze the impact of MT on sarcasm detection.

8.1 Datasets Selection and Preprocessing

The study uses an English dataset (Misra and Arora, 2023) and a Russian dataset (Kostya165, 2023). The English dataset is based on the `Sarcasm_Headlines_Dataset`, which contains news headlines that are already annotated for binary sarcasm classification. The corpus includes 11 724 sarcastic and 14 985 non-sarcastic headlines collected from online news sources. The dataset is provided in a case-normalized form, where all text, including named entities such as company names and proper nouns,

8 Methodology

is converted to lowercase. The Russian dataset is based on the `ru_emotion_dvach corpus`, which contains user-generated text samples labeled across multiple emotional categories, including sarcasm, aggression, anxiety, neutral, and positive emotions, with an overall number of 59 054 segments.

Table 2 shows the main information about each dataset and presents the language, label distribution, and reference source. While the English dataset already provides binary sarcasm annotations and is rather balanced, the Russian dataset required additional preprocessing in order to be better adapted for the sarcasm detection task.

Table 2: Overview of the selected datasets used for sarcasm detection

Dataset	Language	Original distribution	Reference
Sarcasm_Headlines_Dataset	English	14 985 non-sarcastic	Misra and Arora (2023)
		11 724 sarcastic headlines	
ru_emotion_dvach corpus	Russian	11 693 sarcastic	Kostya165 (2023)
		47 361 other emotions	

To ensure compatibility between datasets, both datasets were converted into a binary classification format consisting of text and label, while any potential missing values were removed during preprocessing. For the Russian dataset, sarcastic examples were assigned label 1, while all remaining emotional categories were treated as non-sarcastic and assigned label 0.

Since the Russian dataset originally contained substantial class imbalance, additional preprocessing was necessary. All sarcastic examples were extracted, and an equal number of non-sarcastic segments were randomly sampled from the remaining categories to create a balanced binary classification dataset with 11 693 examples for both sarcastic and non-sarcastic classes. The English dataset was also balanced to ensure equal numbers of sarcastic and non-sarcastic samples with 11 724 examples for each class. This minimal preprocessing approach was chosen to preserve the original lexical and stylistic properties of the data, which are important for sarcasm detection, as sarcastic cues may be expressed through diverse linguistic forms.

To maintain comparable experimental conditions across languages, both datasets were reduced to 20 000 samples each, with approximately balanced class distributions with 10 009 non-sarcastic and 9 991 sarcastic for English dataset and 10 024 sarcastic and 9 976 non-sarcastic for Russian dataset. This design choice was selected in order to minimize any potential bias in case of unequal dataset sizes

8.2 Model Selection and training configurations

and allowed a fairer cross-lingual comparison.

Each dataset was divided into training, validation, and test subsets using stratified sampling to preserve class balance and a fixed random seed of 42 was used for reproducibility. The data was split into 70% training, 15% validation, and 15% test sets to ensure a sufficient amount of data for model training while still providing adequately sized validation and test sets for model selection and evaluation. The training set contained 14 000 examples, with 7 007 non-sarcastic and 6 993 sarcastic instances for the English dataset, and 7 016 sarcastic and 6 984 non-sarcastic instances for the Russian dataset and both the validation and test sets containing 3 000 examples each, with approximately balanced class distributions of about 50% per class. Table 3 summarizes the final dataset statistics across all splits.

Table 3: Final dataset statistics across splits

Language	Split	Non-sarcastic	Sarcastic	Avg. words
English	Train	7 007	6 993	9.87
	Validation	1 501	1 499	9.85
	Test	1 501	1 499	9.81
Russian	Train	6 984	7 016	26.62
	Validation	1 496	1 504	26.57
	Test	1 496	1 504	26.51

8.2 Model Selection and training configurations

A multilingual transformer-based model, XLM-R (Conneau et al., 2020), was selected for model’s training in this experiment and is fine-tuned for sarcasm detection by using annotated English and Russian dataset. The model was chosen because it supports multiple languages and allows the same model to be applied to both English and Russian data, which is important for cross-lingual experiments. Although pretrained sarcasm detection models exist, they were not used in this study to avoid relying on previous task-specific training. Instead, XLM-R was used as a multilingual baseline for all experiments to ensure a consistent and comparable setup.

Instead of focusing on comparing different model architectures, this part of the experiment focuses on evaluating the results of different training configurations. The model is fine-tuned under four different settings: monolingual training on English data, monolingual training on Russian data, multilingual training on a combined English and Russian dataset, and finally sequential fine-tuning, where the model was first trained on English data and then further fine-tuned on Russian data.

8 Methodology

The following Table 4 demonstrates the fine-tuning configurations that were used during this stage in the current experiment.

Table 4: Overview of models and hyperparameters used in the experiment

Model	Language	Samples	Epoches	Batch	LR
XLM-RoBERTa	EN	14,000	5	16	2e-5
XLM-RoBERTa	RU	14,000	5	16	2e-5
XLM-RoBERTa	EN+RU	14,000	5	16	2e-5
XLM-RoBERTa	EN->RU	14,000	5	16	2e-5

All models were fine-tuned under the same conditions to ensure consistency of the experiment on different datasets. The AdamW optimizer was used with a learning rate of 2e-5 and a weight decay of 0.01. The number of training epochs was fixed to 5 based on preliminary runs, where no further improvement in validation performance was observed beyond this point. A batch size of 16 was used for both training and evaluation.

All models were trained for a binary classification task using the standard cross-entropy loss function. Training loss was computed during training, and validation performance in terms of F1-score was used for model selection. The best model was selected based on validation F1-score.

F1-score was used as the main evaluation metric because it provides a balanced measure of precision and recall, which is important for sarcasm detection, where both false positives and false negatives are relevant. A fixed random seed of 42 was used to ensure reproducibility. For the final evaluation, precision, recall, and F1-score were computed on the test sets to provide a more detailed assessment of model performance.

8.3 Sarcasm Detection Approach

For the current experiment sarcasm detection is performed as a binary text classification task, where each input text is assigned one of the following classes, either sarcastic (1) or non-sarcastic (0). The input data consists of short segments, such as news headlines or user-generated content, which often require contextual interpretation to identify implicit meaning. Each input text was tokenized with the XLM-R tokenizer and processed as a single sequence. The encoded text representation was then passed through the model’s classification layer, which produces scores for the two possible classes. The final label was assigned based on the class with the highest predicted score. A multilingual XLM-R model was used in all experimental settings to ensure consistent processing across English and Russian data.

During this experiment cross-lingual robustness was evaluated by fine-tuning models under different configurations and testing them on both English and Russian datasets. Additionally, the impact of MT was assessed by applying the best trained model to machine translated dataset at the further stage of the experiment in order to analyze whether sarcastic meaning was preserved across languages.

8.4 Machine Translation Approach

In the second stage of the experiment, the English test dataset was translated into Russian using the MT system. The NLLB-200 model (Team et al., 2022) was used for this purpose due to its multilingual capabilities and strong performance across a wide range of language pairs.

The translation was performed at the sentence level, where each English text segment will be translated into Russian. The original sarcasm annotations were preserved to make possible the direct comparison between the original English inputs and their translations into Russian.

The resulting translated dataset was then used to evaluate how machine translation affects sarcasm detection performance. Specifically, the best-performing sarcasm detection model identified in the previous stage was applied to the data translated into Russian. This setup would enable analysis of how MT affects sarcasm recognition performance.

All texts were translated using the same NLLB-200 model and translation setup, with a maximum sequence length of 128 tokens, which was sufficient given that the dataset consists of news headlines with an average length of approximately 9 words. The translated outputs are stored alongside the original data to enable subsequent evaluation and comparison of sarcasm detection performance.

8.5 Evaluation

The performance of each model under different training configurations was evaluated on both English and Russian test sets using standard classification metrics, including accuracy, precision, recall, and F1-score. These metrics were computed based on the comparison between the predicted labels and the true labels.

Accuracy was used as a general measure of overall performance, while precision and recall provide a more detailed view of the model's behavior in terms of false positives and false negatives. F1-score was used as the main evaluation metric, as it balances precision and recall and is particularly informative for sarcasm detection tasks, where both false positives and false negatives can significantly affect the interpretation of results.

Evaluation was performed on both English and Russian test sets in order to assess cross-lingual performance. Each fine-tuned model was tested on both datasets, to observe how well the models generalize across languages. All evaluation metrics were computed using the scikit-learn library based on model predictions obtained through the Hugging Face Transformers library.

In addition to sarcasm detection performance, the quality of machine-translated data was also evaluated. Two reference-free evaluation metrics were used for this purpose, namely multilingual BERTScore and COMET-QE (Rei et al., 2020b). These metrics were applied to the translated dataset to assess the semantic similarity between the original English texts and their Russian translations, providing an indication of how well meaning was preserved during MT. In terms of sarcasm detection, the impact of MT was evaluated by applying the best-performing sarcasm detection model to the translated dataset and comparing its performance with that obtained on the original English and Russian test sets.

8.6 Qualitative Analysis

Finally, a qualitative analysis was performed on the translated dataset, focusing on text segments that are labeled as sarcastic in the original English dataset. The analysis included both cases where the translated Russian text were correctly identified as sarcastic by the best-performing model and cases where sarcasm was not successfully detected after translation. This comparison aimed to identify how MT influences the preservation of sarcastic meaning.

Particular attention was paid to the translation techniques used in the MT output, which was identified through manual qualitative analysis based on translation studies literature, including compensation, modulation, adaptation, transposition, and other strategies described in translation studies literature (Newmark, 1988; Zawawi and Maghfiroh, 2020; Molina and Albir, 2002), in order to examine whether specific translation strategies were associated with the preservation or loss of sarcastic meaning.

This analysis aims to identify which translation techniques are associated with preservation or loss of sarcastic meaning in machine-translated text. By comparing both correctly and incorrectly classified translated examples, the qualitative analysis complements the quantitative evaluation and provides further insight into when sarcastic meaning remains recognizable after MT.

9 Results

This chapter presents the results of the sarcasm detection experiments, including model performance under different training configurations and the application of the best-performing model to the English dataset translated into Russian. The performance of each model was evaluated on both English and Russian test sets using standard evaluation metrics, including accuracy, precision, recall, and F1-score.

Table 5 provides an overview of the evaluation results across all configurations. Overall, the results suggest that model performance varies depending on both the training configuration and the language of evaluation. While the monolingual models achieved the strongest performance on test data in the same language as the training data, the combined and sequential training approaches demonstrated different levels of cross-lingual generalization. The following sections present a detailed analysis of each training setup and its performance in both monolingual and cross-lingual scenarios.

Table 5: Overview of the results of the test set evaluation on different model configurations

Model	Test set	Accuracy	Precision	Recall	F1
EN-only	EN	91.70	92.63	90.59	91.60
EN-only	RU	50.20	51.15	14.76	22.91
RU-only	RU	94.50	95.51	93.41	94.45
RU-only	EN	49.23	45.89	8.94	14.96
Combined	EN	89.40	91.56	86.79	89.11
Combined	RU	93.80	94.65	92.89	93.76
Sequential EN–RU	EN	73.57	87.24	55.17	67.59
Sequential EN–RU	RU	94.13	93.57	94.81	94.19

Figure 1 presents a comparison of F1-scores across different training configurations and test sets. The visualization highlights clear differences between monolingual, combined, and sequential training approaches. Monolingual models demonstrate strong performance when they are evaluated on the test set with the same language but show a substantial decline in cross-lingual settings. In contrast, the combined model, that was trained on both Russian and English datasets, maintains con-

9 Results

sistently high F1-scores across both English and Russian test sets, which could indicate improved generalization across languages.

Unlike the combined model, which was trained simultaneously on the merged English and Russian datasets, the sequential model was first trained on the English dataset and then further fine-tuned on the Russian dataset. The sequential model achieves high performance on Russian, which was the second training language, but shows reduced performance on English, the first training language, suggesting an imbalance in learned representations. Overall, the chart demonstrates the contrast between the performance of monolingual models and their limited cross-lingual generalization, while it also emphasizes the advantage of the combined training approach. The following sections provide a more detailed analysis of each model configuration.

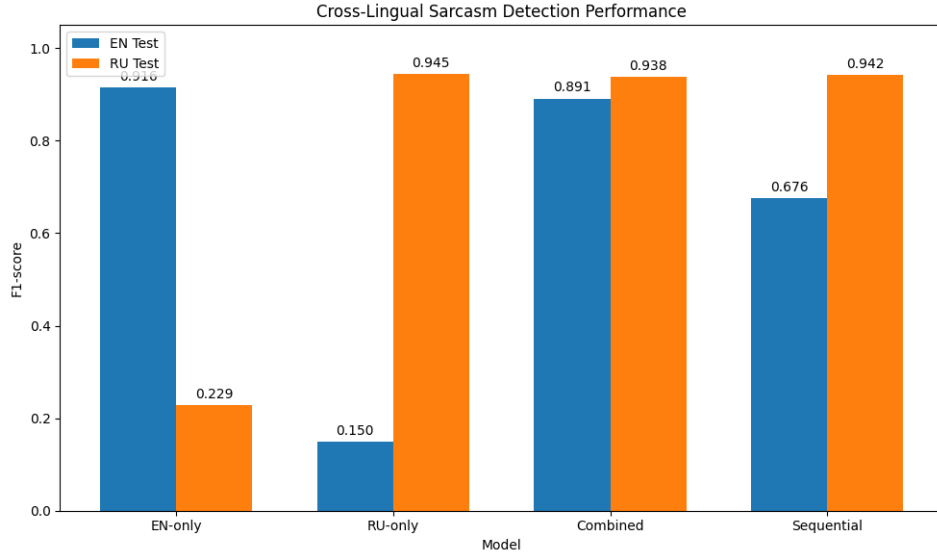


Figure 1: Overview of the F1 score across different models

9.1 Training on English dataset

The model trained on the English dataset demonstrates strong performance when evaluated on the English test set, achieving high scores across all evaluation metrics, with an F1-score of 92.60%. This indicates that the model is able to effectively learn and recognize sarcasm patterns within the same language.

However, when applied to the Russian test set, the model’s performance drops significantly. In particular, recall decreases sharply from 90.59% on the English test set to 14.76% on the Russian test set, resulting in a low F1-score of 22.91%. This

indicates that the model struggles to generalize sarcasm detection across languages when trained only on English data. The substantial decrease in recall further shows that the model fails to correctly identify a large portion of sarcastic instances, revealing a strong bias toward the non-sarcastic class in the cross-lingual setting.

To further analyze the performance of the English-only model, confusion matrices presented in Figure 2 were generated for the evaluation on both English and Russian test sets. These matrices provide insights into the distribution of correct and incorrect predictions across classes.

On the English test set, the model demonstrates balanced performance, with a relatively even distribution of errors across sarcastic and non-sarcastic classes. In contrast, on the Russian test set, the model shows a strong bias toward predicting the non-sarcastic class, consistent with the low recall observed earlier.

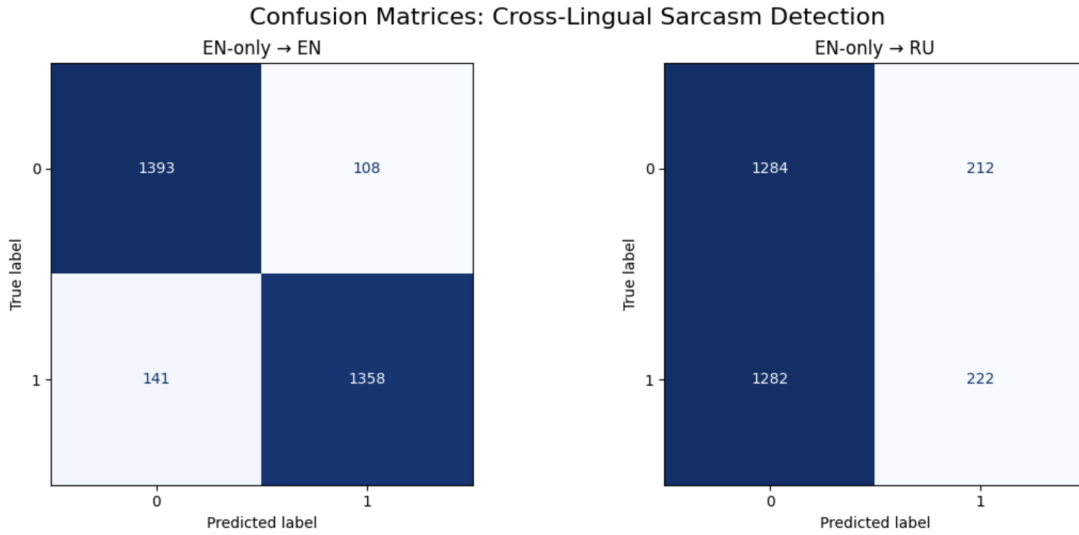


Figure 2: Confusion matrices for English-only model evaluated on the English and Russian test sets

9.2 Training on Russian dataset

The model trained on the Russian dataset achieves strong performance on the Russian test set, with an F1-score of 94.45%, indicating effective sarcasm detection within the same language. The model demonstrates high recall and precision, suggesting that it is able to reliably identify both sarcastic and non-sarcastic instances in Russian.

In contrast, when evaluated on the English test set, the model performs poorly, with a substantial decrease in all evaluation metrics and an F1-score of 14.96%. In

9 Results

particular, recall drops sharply from 93.42% on the Russian test set to 8.94% on the English test set, indicating that the model fails to correctly identify a large portion of sarcastic instances in English. This behavior suggests that the model has difficulty transferring learned representations across languages.

Similar to the English-only configuration, this result highlights limited cross-lingual generalization when the model is trained on a single language, confirming that sarcasm detection patterns learned in one language do not directly transfer to another.

The confusion matrices in Figure 3 further illustrate the differences in performance between monolingual and cross-lingual evaluation for the Russian-only model. The model correctly identifies the majority of both sarcastic and non-sarcastic segments on the Russian test set. However, on the English test set, it shows a strong bias toward predicting the non-sarcastic class that results in a large number of sarcastic instances being misclassified. This pattern is consistent with the substantial decrease in recall and F1-score observed during cross-lingual evaluation.

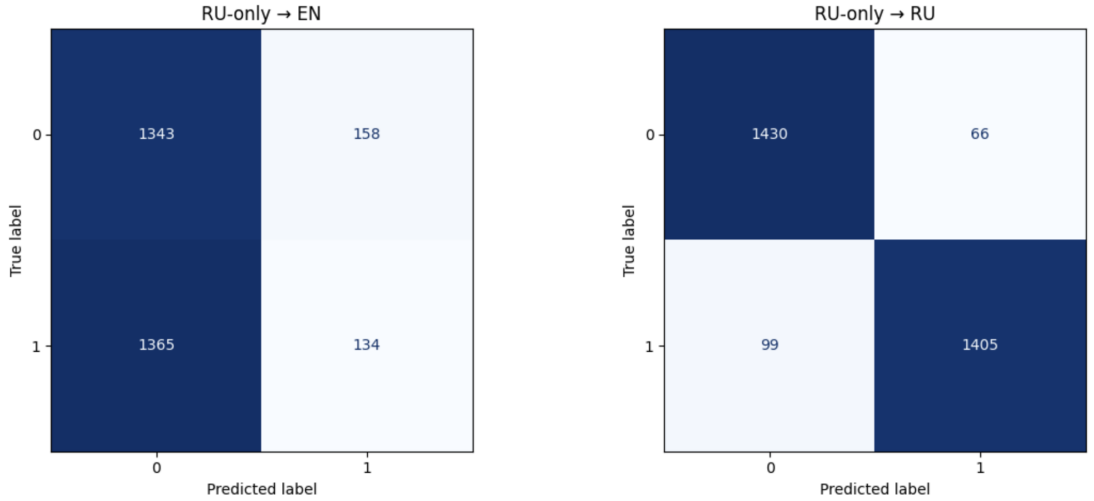


Figure 3: Confusion matrices for the Russian-only model evaluated on the English and Russian test sets

Interestingly, the English-only model demonstrates slightly better cross-lingual performance on the Russian test set than the Russian-only model does on the English test set. While both configurations show limited generalization, this asymmetry indicates that representations learned from English data may transfer more effectively to Russian than vice versa.

9.3 Combined training

The model which was trained on the combined English-Russian dataset demonstrates consistently strong performance across both test sets that is reflected in Table 5. On the English test set, the model achieves an F1-score of 89.11%, while on the Russian test set it reaches 93.76%.

Although the performance on the English data is slightly lower compared to the English-only model, the combined model shows a clear improvement in cross-lingual generalization. This indicates that training on multilingual data allows the model to generalize better across languages while also maintaining high overall performance.

For further examination of model performance, confusion matrices were generated for the combined model on both English and Russian test sets in Figure 4. These matrices provide additional insight into the distribution of correct and incorrect predictions across classes labeled as sarcastic and non-sarcastic.

On both datasets, the combined model demonstrates relatively balanced performance, with a more even distribution of predictions between sarcastic and non-sarcastic classes compared to the monolingual models. The number of false positives and false negatives is reduced, particularly when compared to monolingual models evaluated on a different language than the one used during training.

This behavior also aligns with the overall evaluation metrics, indicating that training on combined multilingual data improves the model’s ability to generalize across languages and reduces class bias observed when monolingual models are evaluated on a language different from the one used during training.

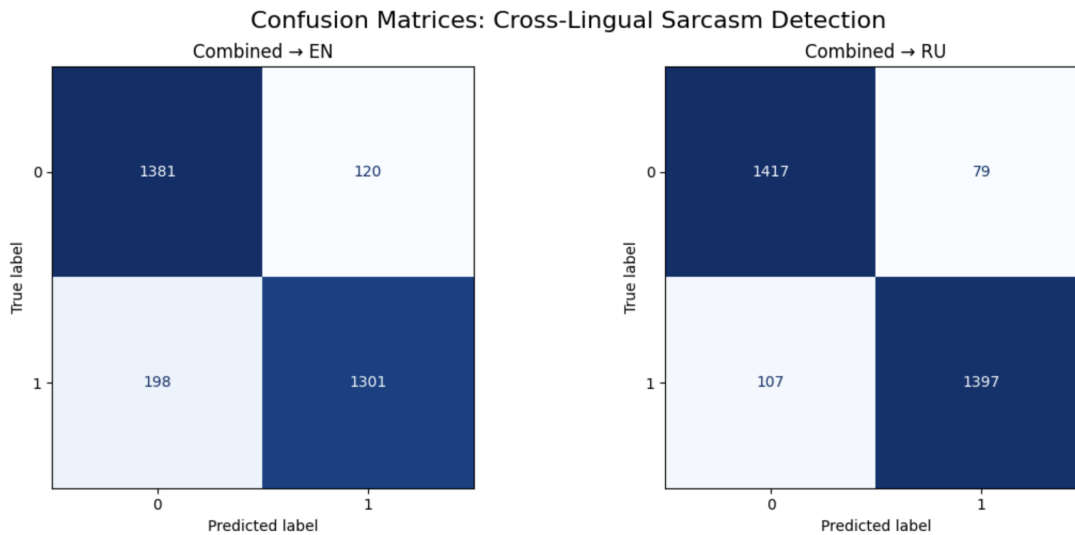


Figure 4: Confusion matrices for the Combined model evaluated on the English and Russian test sets

9.4 Sequential training

The sequentially trained model was at first fine-tuned on the English dataset and then further on the Russian dataset in order to examine how training on multiple languages in sequence affects model performance and cross-lingual generalization. The sequential model demonstrates strong performance on the Russian test set, achieving an F1-score of 94.19%. The model maintains high recall and precision, indicating effective sarcasm detection in Russian after the second stage of training.

However, its performance on the English test set is noticeably lower, with an F1-score of 67.59%. In particular, recall decreases compared to the English-only model, which could suggest that the model becomes less effective at identifying sarcastic segments in English after being additionally fine-tuned on Russian dataset.

This result indicates that the model adapts effectively to the second training language at the later stage but exhibits reduced performance on the original language. Compared to the combined training approach, the sequential setup results in less balanced performance across languages, since the model appears to shift toward the characteristics of the second training dataset.

The confusion matrices in Figure 5 further support the evaluation results. While the model demonstrates strong performance on the Russian test set, the number of incorrectly classified sarcastic instances increases on the English test set. This pattern is consistent with the decrease in recall and F1-score and provides additional evidence that the second stage of fine-tuning affects the model’s performance on previously learned data.

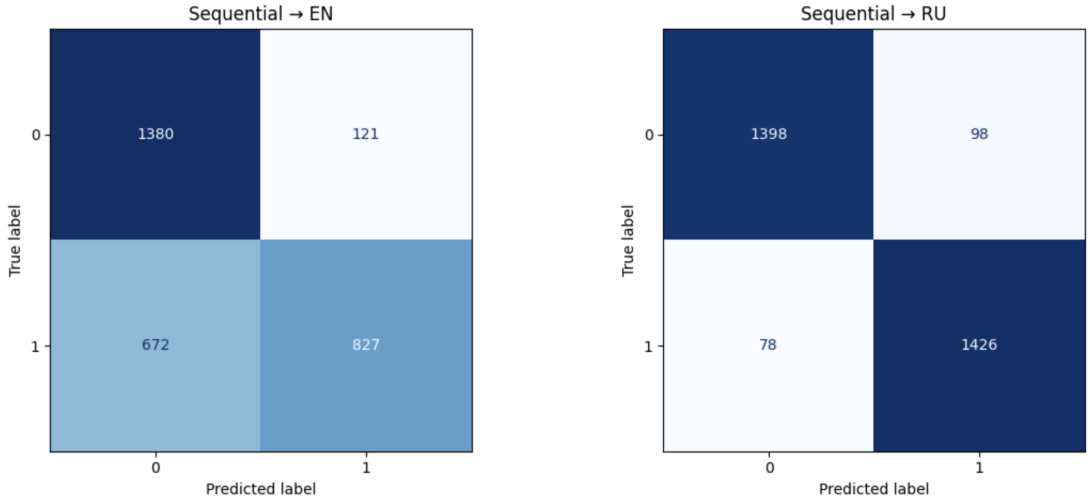


Figure 5: Confusion matrices for the Sequential model evaluated on the English and Russian test sets

This pattern highlights the impact of training order, suggesting that the model prioritizes the most recently seen data. The second stage of fine-tuning on Russian leads to improved performance in Russian, while earlier knowledge acquired from English is partially diminished.

9.5 Sarcasm detection on machine translated dataset

The performance of the combined model on the English dataset in Figure 6 that was machine translated into Russian shows a substantial decline when it is compared to the performance on the original English and Russian datasets. While the model previously achieved high F1-scores of 89.11% on the original English test set and 93.76 on the Russian test set, the results on the translated data achieving F1-score of 36.06% indicate a significant drop in sarcasm detection performance. This decline is particularly evident in the model’s ability to recognize sarcastic instances, with recall for the sarcastic class decreasing to 22.82%. The results suggest a bias toward predicting the non-sarcastic label, which in turn leads to a substantial reduction in the detection of sarcastic content. This pattern suggests that MT may affect the preservation of sarcastic meaning and reduce the visibility of cues that are important for sarcasm recognition.

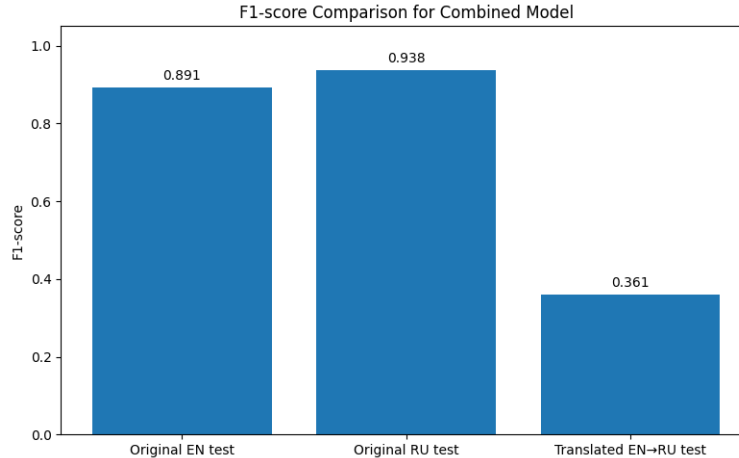


Figure 6: Overview of the F1 scores evaluated on the English, Russian and machine translated test sets

The confusion matrix in Figure 7 provides further insight into the model’s behavior on the machine translated dataset. The model correctly identifies most non-sarcastic instances. However, many instances that were originally labeled as

9 Results

sarcastic in the English dataset are predicted as non-sarcastic after being machine translated into Russian, with the original English labels being retained. This pattern is consistent with the F1-scores reported above and suggests that MT may not fully preserve the linguistic cues associated with sarcasm. Consequently, texts that were originally labeled as sarcastic become less likely to be recognized as such by the model, contributing to the observed decline in performance.

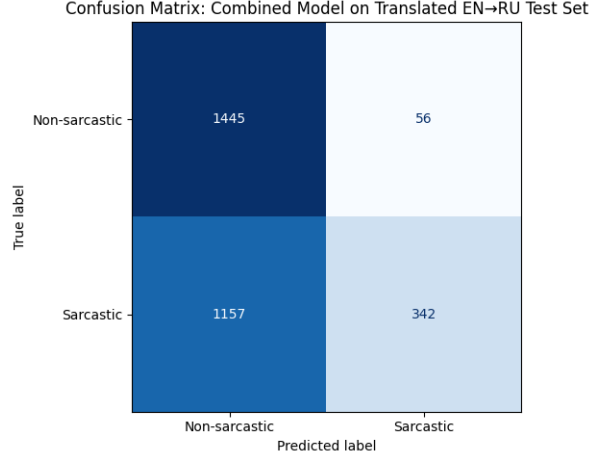


Figure 7: Confusion matrices for the Combined model evaluated on the machine translated dataset

For further assessment of the quality of the machine translated dataset, two reference-free evaluation metrics were applied to the MT output: multilingual BERTScore and COMET-QE. The translated texts achieved an average BERTScore of 71.14% and an average COMET-QE score of 72.19%, which demonstrates a moderate degree of semantic similarity between the original English texts and their Russian translations. The median COMET-QE score of 76.53% further suggests that most translations preserve a substantial portion of the original meaning, although translation quality varies across individual examples during closer manual analysis.

Despite these relatively positive translation quality scores, sarcasm detection performance decreases substantially on the translated dataset. This suggests that MT is generally able to preserve the overall semantic content of the source texts, while failing to consistently preserve the pragmatic and contextual cues necessary for sarcasm recognition.

This observation motivates further qualitative analysis of correctly detected sarcastic correctly detected sarcastic instances as well as sarcastic instances incorrectly classified as non-sarcastic segments to examine which translation strategies were used during MT.

9.6 Qualitative analysis of the translation techniques in Machine translated dataset

The qualitative analysis focused on instances that were originally labeled as sarcastic in the English dataset and examined how MT into Russian affected their recognition by the combined model, which achieved strong performance on both the original English and Russian test sets. Particular attention was paid to two types of outcomes after MT: segments that remained correctly classified as sarcastic and segments that were originally labeled as sarcastic but were classified as non-sarcastic in the translated dataset. For each outcome type, approximately 100 examples were examined. The following Table 6 presents the translation techniques observed in cases where sarcasm was successfully recognized after MT.

Table 6: Translation techniques applied for successfully identified sarcastic segments

Original EN text	MT Russian text	Translation Technique
(1) gun show vendor jokes with insane customer about how he hopes he's not insane	Продавец оружия шутит с безумным клиентом о том, как он надеется, что он не сумасшедший	Literal translation
(2) 'no way to prevent this,' says only nation where this regularly happens	"Нет способа этого предотвратить", - говорит только страна, где это происходит регулярно	Literal translation
(3) buddy system responsible for additional death	Система "друг" ответственная за дополнительную смерть.	Calque
(4) nation fondly recalls when just regulating video games seemed like solution to gun violence	Нация с удовольствием вспоминает, когда просто регулирование видеоигр казалось решением насилия с оружием	Literal translation
(5) mother's little angel just made fun of classmate's weight for 30 straight minutes	Маленький ангел матери только что высмеял вес одноклассницы в течение 30 минут подряд	Literal translation
(6) millionaire thinks of self as upper-middle class	Миллионер считает себя средним классом.	Reduction

As shown in Table 6, literal translation was the most frequently observed

9 Results

translation technique among the correctly identified sarcastic segments, while calque and reduction appeared only occasionally. In these examples, MT generally preserved the semantic contrast or mismatch present in the original text. Although some translations contain lexical or grammatical inaccuracies, the cues contributing to sarcastic meaning often remained visible enough for the model to correctly classify the translated segment as sarcastic.

Examples (1) and (4) in Table 6 demonstrate that sarcasm remains recognizable when the translated text preserves the intended absurdity or inconsistency of the situation. Example (1) conveys sarcasm in the form of a gun show vendor joking with a clearly unstable customer about being insane, while in example (4), the sarcastic effect is created by nostalgically recalling a time when video games regulation was perceived as a sufficient solution to gun violence. In both cases, the original sarcastic intent remains understandable and detectable in the MT output. Example (6) in Table 6 is particularly interesting because the reduction of the expression upper-middle class to middle class results in partial information loss, yet the sarcastic contrast between wealth and self-perception remains visible for successful sarcasm recognition.

In examples (2) and (5) in Table 6, the contrast between the literal statement and the implied meaning also remains visible after translation. In example (2), the contradiction between the claim that there is no way to prevent this and the observation that the same event happens regularly continues to convey the original sarcastic intent in the translated text. Similarly, example (5) preserves the contrast between the positive description little angel and the negative behavior that follows.

Nevertheless, successful sarcasm recognition does not appear to depend on a fully accurate translation. Example (3) contains an awkward calque and grammatical inaccuracies, however, the translated text still contains enough contextual and semantic cues for the model to classify it as sarcastic. Similar observations can be made for several other examples in Table 6, where sarcasm remains detectable despite small translation imperfections. This suggests that fully accurate translation is not always necessary for successful sarcasm recognition in machine translated texts.

However, not all sarcastic instances remained recognizable after being machine translated. A substantial number of segments that were originally labeled as sarcastic were actually classified as non-sarcastic by the model. Additionally, approximately 100 examples of false negative predictions were investigated as well for a more comprehensive analysis. Table 7 presents examples of sarcastic segments that were incorrectly classified as non-sarcastic after MT.

In contrast to the examples presented in Table 6, the segments shown in Table 7 demonstrate cases in which sarcastic segments were mistakenly identified as non-sarcastic by the model after MT. After identifying these misclassified cases,

9.6 Qualitative analysis of the translation techniques in Machine translated dataset

Table 7: Translation techniques applied for unsuccessfully identified sarcastic segments

Original EN text	MT Russian text	Translation Technique
(7) pfizer breaks psychological need to always seek fda’s approval	Психологическая потребность всегда искать одобрение FDA.	Reduction
(8) surviving miner ordered back to work	Выжившим шахтерам приказано вернуться к работе	Literal translation
(9) guinea pig returned for store credit	Гвинея вернулась для покупки кредита.	Literal translation + Reduction
(10) army commander depressed after reading facebook comments on latest raid	Командир армии обеспокоен после прочтения комментариев на фейсбуке о последнем рейде.	Modulation
(11) lucky bastard gets to be in coma	Бедный ублюдок попал в коме.	Modulation

the translated segments were qualitatively analyzed to determine which translation techniques were present and whether they were associated with changes in how sarcastic meaning is expressed. While the overall meaning of the source text is often preserved, in some cases the sarcasm becomes less explicit or the overall clarity of the message is reduced, making it difficult for the model to recognize sarcastic intent. At the same time, other examples preserve the sarcastic meaning of the original text but are nevertheless classified as non-sarcastic by the model. The following examples illustrate the translation techniques identified in these misclassified cases and examine how the translated segments preserve or alter the expression of sarcastic meaning.

Examples (7) and (9) in Table 7 contain cases in which information loss is observed during translation. In example (7), the omission of the subject pfizer removes an important element of the original statement, while in example (9), the expression guinea pig is incorrectly translated as the name of the country, resulting in both information loss and semantic distortion. In these examples, the translated segments contain less explicit sarcastic cues, which may make sarcastic intent more difficult for the model to recognize.

Examples (10) and (11) in Table 7 include instances in which modulation is observed. In example (10), the stronger emotional state depressed is translated

9 Results

as the less expressive concerned, reducing the evaluative impact of the statement. Similarly, in example (11), the sarcastically positive adjective lucky is replaced by the word with negative meaning poor, removing the contrast that contributes to the sarcastic effect. In both examples, the modulation is associated with a weaker expression of the original sarcastic meaning, although this alone may not directly explain the model’s incorrect prediction.

Example (8) in Table 7 shows that preserving the original meaning through literal translation does not necessarily guarantee successful sarcasm recognition. The translated segment preserves both the semantic content and the sarcastic intent of the original statement. Nevertheless, the model classifies the segment as non-sarcastic, indicating that sarcastic meaning may remain present in the translation while still not being detected by the model.

Overall, the analysis suggests that unsuccessful sarcasm recognition cannot only be explained by the translation techniques observed in the MT output. Among the misclassified examples, some translations exhibit information loss, semantic distortions, or polarity shifts that could be associated with less explicit sarcastic cues. In contrast, other examples preserve both the semantic content and the sarcastic intent of the source text but are nevertheless classified as non-sarcastic by the model. These findings indicate that the presence of a particular translation technique alone does not determine whether sarcasm will be successfully recognized after MT.

10 Discussion

This chapter discusses the key findings of the current master’s thesis and compares them with the existing research on automatic sarcasm detection and MT. Furthermore, it also addresses the limitations of the study and outlines possible directions for future research and further development of this field.

The results of this master’s thesis demonstrate that sarcasm detection performance is strongly influenced by both the training configuration and the language of evaluation. Both monolingual models, English-only and Russian-only, achieved high performance when they were evaluated on test sets in the same language as the training data, but their performance decreased substantially in cross-lingual settings. This result suggests that sarcasm detection patterns learned from one language do not effectively transfer to another in the English-Russian language pair, despite the use of the multilingual transformer-based model XLM-R (Conneau et al., 2020). While Tasnim and Nath (2024) reported no significant differences in F1-scores, when they investigated cross-language sarcasm detection using English and Bangla datasets, the present experiment demonstrated a significant performance drop in cross-lingual evaluation. These findings suggest that cross-lingual sarcasm detection is dependent on the language pair and evaluation setting and also support Tasnim and Nath’s conclusion that further research is needed to assess the applicability of cross-language sarcasm detection to other language pairs.

Among all evaluated training configurations, the combined model achieved the most balanced performance across both English and Russian test sets. While both monolingual models produced slightly higher results with their training languages during evaluation on test sets, the combined approach demonstrated better cross-lingual generalization. This observation partially supports the findings of Abu Farha and Magdy (2021), who demonstrated superior performance of language-specific models. In the present study, monolingual models also achieved the highest scores on their respective languages. However, the combined model provided a better balance between monolingual performance and cross-lingual generalization which suggests that combined multilingual training may be beneficial when balanced performance across multiple languages is important.

The sequential training configuration examined in this study involved initial training on the English dataset followed by further fine-tuning on the Russian dataset. Although the model achieved strong results on the Russian test set after the second stage of fine-tuning, its performance on English considerably decreased.

This finding suggests that, in the English to Russian training configuration, sequential fine-tuning resulted in less balanced performance across languages than the combined training approach.

Another important finding of this master’s thesis concerns the impact of MT on sarcasm detection. Although the fact that translation quality achieved moderate scores according to BERTScore and COMET-QE, sarcasm detection performance decreased significantly on the machine translated dataset. This result indicates that preserving the overall semantic content of a text does not necessarily guarantee the preservation of sarcastic meaning. These findings support the observations of [Zawawi and Maghfiroh \(2020\)](#), who emphasize the importance of translation techniques for successful sarcasm transfer, while at the same time the results of the current experiment also demonstrate that MT systems face challenges when processing figurative and context-dependent language.

The qualitative analysis of the MT output further showed that the preservation of sarcastic meaning and further sarcasm detection do not depend only on the applied translation techniques. Literal translation, reduction and modulation were observed both in segments that remained correctly classified as sarcastic after MT and in segments that were originally sarcastic but were classified as non-sarcastic after translation. Furthermore, several translated segments appeared to preserve the sarcastic meaning of the source text but were nevertheless classified as non-sarcastic by the model. This observation is consistent with the findings of [Ducret et al. \(2020\)](#), who highlight the importance of context for sarcasm detection and also suggest that successful sarcasm detection depends not only on semantic content but also on contextual and pragmatic cues.

Several limitations of the present study should be further acknowledged. First, the experiments were conducted using only one English-Russian language pair. Although these languages differ linguistically and culturally, the findings from the current experiment cannot be directly applied to other language pairs. Cross-lingual sarcasm detection may be affected by language-specific characteristics, dataset composition, and differences in cultural norms that are associated with the use and interpretation of sarcasm. Therefore, further research involving additional languages would be beneficial to evaluate the broader applicability of the observed results.

Second, the MT analysis was limited to a single MT system, namely NLLB-200 ([Team et al., 2022](#)). Different MT systems may apply different translation strategies and actually vary in their ability to preserve sarcastic meaning, pragmatic information, and contextual cues. Therefore, the observed drop in sarcasm detection performance may vary across different MT systems. Future studies could compare multiple MT systems and study whether other architectures are more effective at preserving sarcasm during translation.

The qualitative analysis also had several limitations. Due to time and scope constraints, only a limited number of machine translated segments were examined manually. While these examples provided valuable insights into the interaction between translation techniques and sarcasm preservation, they may have not captured other patterns present in the dataset. A larger-scale qualitative analysis could provide a more comprehensive understanding of how MT influences sarcasm recognition.

Finally, the current study analyzed sarcastic sentences without applying additional conversational or situational context. Previous research has shown that contextual information often plays an important role in sarcasm interpretation. This limitation is also supported by several examples that were studied during the qualitative analysis, where sarcastic meaning appeared to remain clear for human readers but was not successfully identified by the model. Future research could therefore explore whether application of additional contextual information improves sarcasm detection performance in multilingual and machine-translated settings.

11 Conclusion

This thesis investigated the interaction between MT and automatic sarcasm detection in the English-Russian language pair. The study combined experimental setup with qualitative analysis in order to examine whether sarcastic meaning remains detectable after MT and to identify the translation techniques used when translating sarcastic content.

The results of the experiment demonstrated that the sarcasm detection performance is strongly influenced by the training configuration and the evaluation setting. The monolingual English-only and Russian-only models achieved high performance when evaluated on test sets in the same language as the training data, while their performance was lower in cross-lingual settings. This comparison was used primarily to select the best performing model for further evaluation on machine-translated data. For this reason, among all evaluated configurations, the combined model, which was fine-tuned simultaneously on both English and Russian datasets, achieved the most balanced performance across English and Russian test sets and was therefore selected for the analysis of the machine-translated dataset, as it provided a more stable baseline across both languages.

The evaluation on the machine translated dataset revealed a significant drop in sarcasm detection performance in comparison to the original English and Russian test sets. Although the translated texts achieved moderate BERTScore and COMET-QE scores, the model successfully identified only a relatively small proportion of sarcastic segments after MT. These results indicate that the preservation of general semantic content during MT does not necessarily guarantee the preservation of sarcastic meaning. Consequently, these findings support the initial assumption that MT may negatively affect sarcasm detection by altering or weakening cues that are important for sarcasm detection.

Additional qualitative analysis of the MT output highlights two complementary findings. First, in some cases MT appears to weaken or alter linguistic and pragmatic cues that contribute to sarcastic meaning, which may reduce the explicitness of sarcasm in the translated text. Second, there are also cases in which the sarcastic meaning remains largely preserved in the machine-translated output, yet the model still classifies these instances as non-sarcastic. This indicates that misclassification of sarcastic segments as non-sarcastic after MT cannot be attributed solely to translation quality or translation techniques. Instead, this may also be related to limitations of the model in capturing contextual and pragmatic aspects of sarcasm.

11 Conclusion

These findings further indicate that additional contextual information may be required to reliably capture sarcastic intent in multilingual and machine-translated settings.

Overall, the findings in this thesis highlight the challenges of processing sarcasm in multilingual NLP settings. While modern MT systems are often capable of preserving the general meaning of a text, they still may fail to preserve pragmatic and contextual cues that are important for automatic sarcasm detection. Future research could extend this work by investigating additional language pairs, comparing different MT systems, and further exploring the role of contextual information in multilingual sarcasm detection and translation.

Bibliography

- Muhammad Abdul-Mageed, AbdelRahim A. Elmadany, and El Moatez Billah Nagoudi (2021). [ARBERT & MARBERT: deep bidirectional transformers for arabic](#). *CoRR*, abs/2101.01785.
- Ibrahim Abu Farha and Walid Magdy (2021). [Benchmarking transformer-based language models for Arabic sentiment and sarcasm detection](#). In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 21–31, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.
- Ibrahim Abu Farha, Wajdi Zaghouani, and Walid Magdy (2021). [Overview of the WANLP 2021 shared task on sarcasm and sentiment detection in Arabic](#). In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 296–305, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.
- E. Alpaydin (2014). [Introduction to Machine Learning](#). Adaptive Computation and Machine Learning series. MIT Press.
- Wissam Antoun, Fady Baly, and Hazem Hajj (2020). [AraBERT: Transformer-based model for Arabic language understanding](#). In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resources Association.
- Rajesh Bharati, Wasudeo Rahane, P.D. Patil, Shubham Tapkeer, Vijay Waghmare, Jaydeep Patil, and Simran Desai (2024). [Svm-based sarcasm detection system: Nlp using heuristic approach](#). In *2024 8th International Conference on Computing, Communication, Control and Automation (ICCUBEA)*, pages 1–6.
- Bhagirath Bhatt and Anjali Gaikwad (2023). *The Fundamentals of Machine Learning*. LAP LAMBERT Academic Publishing.
- Wala bin Subait, Mashael M. Asiri, Muhammad Swaileh A. Alzaidi, Meshari H. Alanazi, Menwa Alshammeri, Ayman Yafoz, Raed Alsini, and Alaa O. Khadidos (2025). [Artificial intelligence-based natural language processing for sarcasm detection and classification on arabic corpus](#). *Alexandria Engineering Journal*, 125:320–331.

Bibliography

- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei (2020). [Language models are few-shot learners](#). *CoRR*, abs/2005.14165.
- Bojana Budimir (2025). [The challenge of translating culture-specific items: Evaluating MT and LLMs compared to human translators](#). In *Proceedings of Machine Translation Summit XX: Volume 1*, pages 455–467, Geneva, Switzerland. European Association for Machine Translation.
- Andriy Burkov (2019). *The Hundred-Page Machine Learning Book*, volume 1. Andriy Burkov, Quebec City, QC, Canada.
- Davide Buscaldi and Paolo Rosso (2023). [How good is NLLB for low-resource languages? a study on the genoese language](#). In *Proceedings of the Ninth Italian Conference on Computational Linguistics (CLiC-it 2023)*, pages 490–493, Venice, Italy. CEUR Workshop Proceedings.
- Elisabeth Camp (2012). [Sarcasm, pretense, and the semantics/pragmatics distinction](#). *Noûs*, 46(4):587–634.
- Wangqun Chen, Fuqiang Lin, Xuan Zhang, Guowei Li, and Bo Liu (2022). [Jointly learning sentimental clues and context incongruity for sarcasm detection](#). *IEEE Access*, 10:48292–48300.
- Zheng Lin Chia, Michal Ptaszynski, Marzena Karpinska, Juuso Eronen, and Fumito Masui (2024). [Initial exploration into sarcasm and irony through machine translation](#). *Natural Language Processing Journal*, 9:100106.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov (2020). [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- Corinna Cortes and Vladimir Naumovich Vapnik (1995). [Support-vector networks](#). *Machine Learning*, 20:273–297.

- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Alphaeus Dmonte, Shrey Satapara, Rehab Alsudais, Tharindu Ranasinghe, and Marcos Zampieri (2025). [On the effects of machine translation on offensive language detection](#). *Soc. Netw. Anal. Min.*, 14(1).
- Martina Ducret, Lauren Kruse, Carlos Martinez, Anna Feldman, and Jing Peng (2020). [Linguistic features in automatic sarcasm detection](#). In *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020)*, pages 132–138, Bologna, Italy. CEUR Workshop Proceedings.
- Jeffrey L. Elman (1990). [Finding structure in time](#). *Cognitive Science*, 14(2):179–211.
- Eman Fatima, Hira Kanwal, J. A. Khan, et al. (2024). [An exploratory and automated study of sarcasm detection and classification in app stores using fine-tuned deep learning classifiers](#). *Automated Software Engineering*, 31(1):69.
- Ravi Teja Gedela, Utpal Baruah, and Biplab Soni (2024). [Deep contextualised text representation and learning for sarcasm detection](#). *Arabian Journal for Science and Engineering*, 49(3):3719–3734.
- Debanjan Ghosh, Avijit Vajpayee, and Smaranda Muresan (2020). [A report on the 2020 sarcasm detection shared task](#). In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 1–11, Online. Association for Computational Linguistics.
- Sharon Goldwater and David McClosky (2005). [Improving statistical MT through morphological analysis](#). In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 676–683, Vancouver, British Columbia, Canada. Association for Computational Linguistics.
- Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio (2014). [Generative adversarial networks](#).
- Hunter Gregory, Steven Li, Pouya Mohammadi, Natalie Tarn, Rachel Draelos, and Cynthia Rudin (2020). [A transformer approach to contextual sarcasm detection](#)

Bibliography

- in Twitter. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 270–275, Online. Association for Computational Linguistics.
- Jie Gui, Tuo Chen, Jing Zhang, Qiong Cao, Zhenan Sun, Hao Luo, and Dacheng Tao (2024). A survey on self-supervised learning: Algorithms, applications, and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9052–9071.
- Nivin A. Helal, Ahmed Hassan, Nagwa Lotfy Badr, and Yasmine M. Afify (2024). A contextual-based approach for sarcasm detection. *Scientific Reports*, 14.
- Tin Kam Ho (1995). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE.
- Sepp Hochreiter and Jürgen Schmidhuber (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.
- Nicholaus Jeremy, Matthew Alam, Justin Tristan Tirtawijaya, and Jonathan Samuel Lumentut (2024). Optimizing sarcasm detection through various input methods in text data. In *2024 2nd International Conference on Technology Innovation and Its Applications (ICTIIA)*, pages 1–5.
- Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean (2016). Google’s multilingual neural machine translation system: Enabling zero-shot translation. *CoRR*, abs/1611.04558.
- Aditya Joshi, Vinita Sharma, and Pushpak Bhattacharyya (2015). Harnessing context incongruity for sarcasm detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 757–762, Beijing, China. Association for Computational Linguistics.
- Aditya Joshi, Vaibhav Tripathi, Pushpak Bhattacharyya, Mark Carman, Meghna Singh, Jaya Saraswati, and Rajita Shukla (2016). How challenging is sarcasm versus irony classification?: A study with a dataset from English literature. In *Proceedings of the Australasian Language Technology Association Workshop 2016*, pages 123–127, Melbourne, Australia.
- Daniel Jurafsky and James H. Martin (2026). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd edition. Online manuscript released January 6, 2026.

- Dorothy Kenny (2022). *Machine translation for everyone : Empowering users in the age of artificial intelligence*. Language Science Press,, [s.l.] :.
- Shumaila Khan, Iqbal Qasim, Wahab Khan, Khursheed Aurangzeb, Javed Ali Khan, and Muhammad Shahid Anwar (2025). [A novel transformer attention-based approach for sarcasm detection](#). *Expert Systems*, 42(1):e13686.
- Philipp Koehn (2020). [Neural Machine Translation](#), 1 edition. Cambridge University Press, New York.
- Philipp Koehn and Rebecca Knowles (2017). [Six challenges for neural machine translation](#). In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.
- Kostya165 (2023). Русский датасет эмоций (e-kulture). https://huggingface.co/datasets/Kostya165/ru_emotion_e-kulture.
- Roger Kreuz and Gina Caucci (2007). [Lexical influences on the perception of sarcasm](#). In *Proceedings of the Workshop on Computational Approaches to Figurative Language*, pages 1–4, Rochester, New York. Association for Computational Linguistics.
- Yuri Kuratov and Mikhail Y. Arkhipov (2019). [Adaptation of deep bidirectional multilingual transformers for russian language](#). *CoRR*, abs/1905.07213.
- Nathan Lambert (2025). [Reinforcement learning from human feedback](#). *ArXiv*, abs/2504.12501.
- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut (2019). [ALBERT: A lite BERT for self-supervised learning of language representations](#). *CoRR*, abs/1909.11942.
- Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner (1998). [Gradient-based learning applied to document recognition](#). *Proc. IEEE*, 86:2278–2324.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov (2019). [Roberta: A robustly optimized BERT pretraining approach](#). *CoRR*, abs/1907.11692.
- Chi-kiu Lo, Rebecca Knowles, and Cyril Goutte (2023). [Beyond correlation: Making sense of the score differences of new MT evaluation metrics](#). In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 186–199, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Bibliography

- Tyler Loakman, Aaron Maladry, and Chenghua Lin (2023). [The iron\(ic\) melting pot: Reviewing human evaluation in humour, irony and sarcasm generation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6676–6689, Singapore. Association for Computational Linguistics.
- Sanzana Lora, G. M. Shahariar, Tamanna Nazmin, Noor Rahman, Rafsan Rahman, Miyad Bhuiyan, and Faisal Shah (2024). [Ben-sarc: A self-annotated corpus for sarcasm detection from bengali social media comments and its baseline evaluation](#). *Natural Language Processing*, pages 1–26.
- Bernardo Magnini, Alberto Lavelli, and Simone Magnolini (2020). [Comparing machine learning and deep learning approaches on NLP tasks for the Italian language](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2110–2119, Marseille, France. European Language Resources Association.
- Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze (2008). [Introduction to Information Retrieval](#). Cambridge University Press, Cambridge, UK.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean (2013). [Efficient estimation of word representations in vector space](#).
- Rishabh Misra and Prahal Arora (2023). [Sarcasm detection using news headlines dataset](#). *AI Open*, 4:13–18.
- Lucía Molina and Amparo Albir (2002). [Translation techniques revisited: A dynamic and functionalist approach](#). *Meta: Journal des traducteurs*, 47:498.
- Joss Moorkens, Andy Way, and Séamus Lankford (2025). [Automating translation](#). Routledge introductions to translation and interpreting. Routledge, Abingdon, Oxon New York, NY.
- Abhilash Nandy, Yash Agarwal, Ashish Patwa, Millon Madhur Das, Aman Bansal, Ankit Raj, Pawan Goyal, and Niloy Ganguly (2024). [***YesBut***: A high-quality annotated multimodal dataset for evaluating satire comprehension capability of vision-language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16878–16895, Miami, Florida, USA. Association for Computational Linguistics.
- Peter Newmark (1988). [A Textbook of Translation](#). English language teaching. Prentice-Hall International.

- Ojas Nimase and Sanghyun Hong (2024). [When do “more contexts” help with sarcasm recognition?](#) In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17537–17543, Torino, Italia. ELRA and ICCL.
- Simona-Vasilica Oprea and Adela Bâra (2025). [LLM-as-a-judge for sarcasm detection using supervised fine-tuning of transformers](#). *J. King Saud Univ. - Comput. Inf. Sci.*, 37(10).
- Jianhui Pang, Fanghua Ye, Derek Fai Wong, Dian Yu, Shuming Shi, Zhaopeng Tu, and Longyue Wang (2025). [Salute the classic: Revisiting challenges of machine translation in the age of large language models](#). *Transactions of the Association for Computational Linguistics*, 13:73–95.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2002). [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Jeffrey Pennington, Richard Socher, and Christopher Manning (2014). [GloVe: Global vectors for word representation](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.
- Lucie Polakova, Miroslav Hrabal, Věra Kloudová, Michal Novák, Mariia Anisimova, and Martin Popel (2026). [Challenges in machine translation of interactive multimodal exercises](#). In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, pages 141–152, San Diego, California, USA. Association for Computational Linguistics.
- Jitesh Pradhan, Rajshree Verma, Sumit Kumar, and Varun Sharma (2024). [An efficient sarcasm detection using linguistic features and ensemble machine learning](#). *Procedia Computer Science*, 235:1058–1067.
- Shenbin Qian, Constantin Orăsan, Félix do Carmo, and Diptesh Kanojia (2023). [Challenges of human vs machine translation of emotion-loaded Chinese microblog texts](#). In *Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track*, pages 217–236, Macau SAR, China. Asia-Pacific Association for Machine Translation.
- Shenbin Qian, Constantin Orasan, Diptesh Kanojia, and Félix Do Carmo (2024). [A multi-task learning framework for evaluating machine translation of emotion-loaded user-generated content](#). In *Proceedings of the Ninth Conference on*

Bibliography

- Machine Translation*, pages 1140–1154, Miami, Florida, USA. Association for Computational Linguistics.
- Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie (2020a). [COMET: A neural framework for MT evaluation](#). *CoRR*, abs/2009.09025.
- Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie (2020b). [Unbabel’s participation in the WMT20 metrics shared task](#). In *Proceedings of the Fifth Conference on Machine Translation*, pages 911–920, Online. Association for Computational Linguistics.
- Caroline Rossi and Alice Carré (2022). [How to choose a suitable nmt solution?: Evaluation of mt quality](#). In *Machine Translation for Everyone*, pages 51–79. Language Science Press.
- Stuart J. Russell and Peter Norvig (2020). *Artificial Intelligence: A Modern Approach*. Pearson.
- Hadeel Saadany, Constantin Orasan, Rocio Caro Quintana, Felix Do Carmo, and Leonardo Zilio (2023). [Analysing mistranslation of emotions in multilingual tweets by online MT tools](#). In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 275–284, Tampere, Finland. European Association for Machine Translation.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf (2019). [Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter](#). *CoRR*, abs/1910.01108.
- Venkata Sukhavasi, V. P. K. Sistla, and Venkata Dondeti (2025). [Sarcasm detection using optimized bi-directional long short-term memory](#). *Knowledge and Information Systems*, 67(7):2771–2799.
- Nuzat Tasnim and Rudra Pratap Deb Nath (2024). [A cross-language analysis on sarcasm detection](#). In *2024 27th International Conference on Computer and Information Technology (ICCIT)*, pages 998–1003.
- Yla R. Tausczik and James W. Pennebaker (2010). [The psychological meaning of words: Liwc and computerized text analysis methods](#). *Journal of Language and Social Psychology*, 29:24 – 54.
- Nllb Team, Marta Ruiz Costa-jussà, James Cross, Onur cCelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Alison Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John

- Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon L. Spruit, C. Tran, Pierre Yves Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco (Paco) Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang (2022). [No language left behind: Scaling human-centered machine translation](#). *CoRR*, abs/2207.04672.
- Swati Tiwari, Vivek Shukla, and Abhishek Shukla (2024). [Sarcasm detection in text and speech: A comprehensive study](#). In *2024 International Conference on Artificial Intelligence and Quantum Computation-Based Sensor Application (ICAIQSA)*, pages 1–6.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin (2017). [Attention is all you need](#). *CoRR*, abs/1706.03762.
- Yu Wan, Baosong Yang, Derek Fai Wong, Lidia Sam Chao, Liang Yao, Haibo Zhang, and Boxing Chen (2022). [Challenges of neural machine translation for short texts](#). *Computational Linguistics*, 48(2):321–342.
- Shun Wang, Ge Zhang, Han Wu, Tyler Loakman, Wenhao Huang, and Chenghua Lin (2024). [MMTE: Corpus and metrics for evaluating machine translation quality of metaphorical language](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11343–11358, Miami, Florida, USA. Association for Computational Linguistics.
- Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert (2013). [Norms of valence, arousal, and dominance for 13,915 english lemmas](#). *Behavior Research Methods*, 45(4):1191–1207.
- Frank. Wilcoxon (1945). [Individual comparisons by ranking methods](#). *Biometrics*, 1:196–202.
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel (2021). [mT5: A massively multilingual pre-trained text-to-text transformer](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.
- Ahmed Derbala Yacoub, Salwa Slim, and Amal Aboutabl (2024). [A survey of sentiment analysis and sarcasm detection: Challenges, techniques, and trends](#). *International Journal of Electrical and Computer Engineering Systems*, 15(1):69–78.

Bibliography

- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le (2019). [Xlnet: Generalized autoregressive pretraining for language understanding](#). *CoRR*, abs/1906.08237.
- Mohamad Risal Zawawi and Devi Laila Maghfiroh (2020). [Sarcasm and the translation quality in the subtle art of not giving a fuck book](#). *LiNGUA: Jurnal Ilmu Bahasa dan Sastra*.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi (2019). [Bertscore: Evaluating text generation with BERT](#). *CoRR*, abs/1904.09675.
- Ning Zhu and Ruth Filik (2023). [Individual differences in sarcasm interpretation and use: Evidence from the UK and china](#). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 49(3):445–463.
- А. А. Гурин and Т. А. Жуков (2023). [Автоматическое определение сарказма в текстах на русском языке](#). *Цифровая экономика*, 1(22):41–44. РЭУ им. Г.В. Плеханова, Москва.