



978-9934-619-30-4

Trends in AI from Red and Blue Team Perspectives:

Synthetic Data in a Data-Driven Society vs Sentiment Analysis

Published by the
NATO Strategic Communications
Centre of Excellence

mission NATO STRATCOM
alternative paper strategies diplomacy
virtual data global broad
plan
solutions
expertise
goals
communication
LATVIA RIGA
future
CENTRE OF EXCELLENCE

Synthetic Data in a Data-Driven Society vs Sentiment Analysis

[illegible]

ISBN: 978-9934-619-30-4

Authors: Gundars Bergmanis-Korāts, Alec Winshel, James Ostrowski, Kevin Zerussen, Omar Ali, Anastasija Cuikina, Alexander Gegov, Djamila Ouelhadj

Contributors: Marija Isupova, Daina Ozoliņa

Project Manager: Gundars Bergmanis-Korāts

Design: Inga Ropša

NATO STRATCOM COE

11b Kalnciema Iela

Riga LV1048, Latvia

www.stratcomcoe.org

Facebook: [stratcomcoe](https://www.facebook.com/stratcomcoe)

Twitter: [@stratcomcoe](https://twitter.com/stratcomcoe)

Cover image generated by the DALL-E 2.

This publication does not represent the opinions or policies of NATO or NATO StratCom COE.

© All rights reserved by the NATO StratCom COE. Reports may not be copied, reproduced, distributed or publicly displayed without reference to the NATO StratCom COE. The views expressed here do not represent the views of NATO.

Executive Summary

In today's world, much of what we do, from exercising to texting, is measured and tracked. This data is constantly harvested by data brokers and social media to create behavioural profiles, later used by artificial intelligence (AI) powered services. The overwhelming abundance of data has ushered in an age of analytics, and the rise of AI has enabled big-data decision making. However, the training of AI models is often challenging due to a supervised training strategy that requires a large amount of labelled data. Synthetic data is a possible solution to various challenges, including data labelling and, more importantly, data privacy problems. Synthetic data can be generated by using advances in rendering pipelines, generative adversarial models, and fusion models. It is predicted that most online-generated content will be AI created.

In a narrower and more technical context, an increasing number of the deep learning (DL) models each year are either partly or fully adopting the so-called *transformer* architecture that has enabled a significant boost to the field of natural language processing (NLP). Many open-source models allow us to estimate the emotion of spoken/written text, but, as often occurs with community-based models

and data, the majority are trained on the same data sets. The vendors of social media monitoring (SMM) tools are following the trend and implementing new AI-based systems, as well as improving their existing ones, including sentiment estimation strategies. However, we foresee that sentiment is not as clear cut and classifiable as we thought it would be. In this report we assess and explain the limitations of open-source sentiment estimation models/methods to understand the current capabilities and drawbacks when applied directly on scraped open-text data. Additionally, we went beyond sentiment, emotion, and toxicity estimation and developed a metric of micro-aggressions.¹ We focused on structure-based embedding to help understand the structure of toxic/microaggressive messages.

The aim of this report is to inform the general community of AI practitioners and enthusiasts about the risks (red team perspective) and opportunities (blue team perspective) synthetic data brings to digital content generation which can be used to support the disinformation actors in their activities, including context-based content generation. Then we continue with the blue team perspective and investigate how

open-source AI (machine learning and deep learning) could be used to carry out and understand hostile communications, including how AI has enabled a new generation of synthetic operations, how current sentiment analysis lags behind, and how more granular text analysis enables microaggressive and aggressive text patterns to be searched, thus allowing the identification of hostile communications in various online and social media data.

We see that current cutting-edge AI models shape the way we interact with and manipulate information. Widely used and accessible open-source narrow AI models show state-of-the-art performance in various tasks;

however, current NLP lacks capabilities in small languages, which is crucial when monitoring the information environment in multilingual space. Therefore, regaining the trust of text-processing pipelines such as more qualitative language translation and sentiment analysis in SMM tools is an important long-term investment.

CONTENTS

Executive Summary	3
The Double-Edged Sword of Synthetic Data in the Data-Driven Society	6
The Future Is Digital and Largely Synthetic	6
Risks of Synthetic Data	8
NATO and the Metaverse	9
From Useful Digital Avatars to Malicious and Harmful Deepfakes	11
Use of AI-manipulated data by pro-Kremlin actors on VK	13
Conclusions and Recommendations	13
Understanding Communications by Seeing Beyond Polar Emotions: A Reassessment of Traditional Sentiment Analysis	15
Introduction	15
The big picture and relevance to NATO	15
Relevance to practitioners and technical professionals	16
Understanding the Emotions in Communication	18
Sentiment analysis	18
Toxic language detection	19
Assessing the technical challenges in multilingual space: A translation bias of open-source models	20
Beyond polar and toxic emotions: Detecting harmful expressions via a microaggression framework	26
Microaggression detection	26
Motivation for MA detection	26
Composition of MAs	27
Adding Structure to Unstructured Data	29
Introducing RST	29
Introducing DP	32
Example of dependency parsing (Google-translated VK post)	33
Sentiment analysis using DP	34
Case Studies	37
Test design	37
Sentiment and RST	38
Stage 1: sentiment accuracy scores baseline (before RST segmentation + weighting)	38
Stage 2: sentiment accuracy scores (RST segmentation)	38
Stage 3: sentiment accuracy scores (RST segmentation + RST weighting)	39
MAs and RST	40
Stage 1	40
Stage 2	41
Stage 3	41
Summary	42
Recommendations to Data Scientists and AI Enthusiasts in NATO	43
Conclusions and Recommendations	44
AI steers towards intelligent user interfaces.	44
NLP lacks capabilities in small languages.	44
Regaining trust in sentiment analysis using SMM tools is an important long-term investment.	45
Endnotes	47

The Double-Edged Sword of Synthetic Data in the Data-Driven Society

The Future Is Digital and Largely Synthetic

Much of what we do today is measured and tracked. We create data points when we exercise, go on dates, read the news, or look up recipes. Industry, governments, and academia now have access to vast amounts of collected, user-generated, or measured sensorial data. This data is used to do everything from strategic decision making to creating individual behavioural profiles.² The overwhelming abundance of data has ushered in an 'age of analytics', wherein data is fundamental to the day-to-day operations of governments, corporations, and individuals.³

All this data is not useful by itself. It must first be interpreted. Increasingly, artificial intelligence is used for this purpose. AI models are trained on data to solve specific problems. Gigantic, billion-parameter models have demonstrated state-of-the-art performance in numerous tasks, such as curating a social media feed or guiding a self-driving car. At the same time, the industry is squeezing the largest and most powerful machine-learning models into much smaller software that can be run on edge devices, like intelligent kitchen appliances or wearables. Our lives are increasingly interwoven with AI, and our interactions with technology are becoming increasingly personalised. Digital assistants and chatbots are becoming more context-aware and

capable of human-like conversation, learning our habits and personalities, and in the future might even adapt their style of communication according to the personalities of the people they are engaging with.⁴

While the unprecedented availability of training data has fuelled the success of AI models across multiple domains, it presents a new set of challenges. AI models are only as good as the data they are trained on. Models often use a supervised training strategy that requires a large amount of labelled data.⁵ Data labels identify meaningful and informative tags in raw data so that machine-learning models can learn from it. Labels might indicate whether an image contains a person or what words were spoken in a recording. Often, training data does not contain important labels. Annotating data is a time-intensive and usually manual task. Almost every AI engineer and student at some point in their career has thought, 'If only I had a better, less biased, and bigger data set, I could achieve better results.'

One promising solution to this challenge is computer-generated artificial (also called *synthetic*) data. This emerging practice leverages advances in rendering pipelines, generative adversarial models, and fusion

models to generate new, annotated training data. A Gartner 'Predict', published in the *Wall Street Journal*, says: 'By 2024, 60% of the data used for the development of AI and analytics projects will be synthetically generated'.⁶

A similar view is shared also by Schick,⁷ and later mentioned in a Europol report,⁸ which states that as much as 90 per cent of online content may be synthetically generated by 2026. However, according to Wim Kees Janssen, CEO and founder of Syntho.ai, despite the rapid development of technologies, these predictions may be slightly overestimated. But we are already witnessing the imminent growth of synthetic data. Janssen said:

I would not fully agree yet with Gartner's prediction. My estimate for today is that currently the synthetic data volume is less than 10% and that by 2030 we will reach numbers close to 60%, as the year 2024 comes way too early. Many people have not even heard about synthetic data yet and most organisations are struggling with how to apply synthetic data in practice. Next to that, we sometimes also see that synthetic data is applied in cases where one should not use synthetic data, so there are also some misconceptions and over-promises in the market. One should actually be careful with that, because it can harm the adoption of synthetic data. Of course, companies like Syntho aim to accelerate the adoption of synthetic data by sharing knowledge and use cases via an easy-to-use platform.⁹

If the prediction of dramatic synthetic data increase is true, the ability to create synthetic data will significantly transform machine-learning (ML) training strategies, and in the future may become a part of fully automated AI model training pipelines.

The synthetic data industry is not only promising to increase and improve data sets. It also offers a solution to the complex issue of data privacy, which is a major topic in the field of AI ethics. It is crucial to ensure privacy when training on user-generated data, as such data sets risk exposing user data and can be subject to biases.¹⁰ Legal restrictions¹¹ govern the processing of biometric data and provide guidelines for its ethical use. This is a major topic of interest for big tech companies.

Today's vision of tomorrow states that only a fraction of actual data need be provided to a synthetic data vendor, which then uses algorithms to create entirely new and artificial data that cannot be traced back or reverse-engineered to the original data. As a result, synthetic data is not subject to privacy regulations like Europe's General Data Protection Regulation (GDPR), making it a solution to data privacy challenges. The ability to generate new information also makes synthetic data a solution for when there is not enough data, when you want to up-sample edge cases.¹² As with most technological advancements, this approach goes both ways—it can be used in malicious and non-malicious activities.

Later in this paper we will briefly summarise what the technology of today and tomorrow will bring to our digital lives: from analysing deepfakes to assessing emotion classification in textual data.

Risks of Synthetic Data

Malicious actors use synthetic data to mislead by passing it off as the real thing. For example, they could create bot social media accounts with realistic, computer-generated profile information. Such accounts can then be programmed using rule-based systems to respond with pre-scripted answers to specific keywords and sentiment, and to engage with prescription or completely computer-generated data. Perhaps the most sophisticated cases involve AI-generated realistic¹³ and dynamic content,¹⁴ which allow us a glimpse into the possible risks these technologies pose now and in the near future.

“ Malicious actors use synthetic data to create fake social media accounts as bots or sock puppets.

There is no doubt that the organisations able to create fake posts, articles, images, and videos that look real can be harmful. Hostile and technically advanced countries will probably use synthetic data to boost their information activities, including creating false narratives or toxic posts and comments about political subjects. This could have a significant negative impact on democracy, political stability, and security. However, our research indicates that we are not yet at that point. Generating unstructured data at scale with current AI capabilities is impractical and resource-intensive in

most cases. Disinformation researcher Tim Hwang writes that ‘online propagandists are pragmatists’.¹⁵ These bad actors aim to wield the greatest influence at the lowest possible cost. To that end, they have relied more on crude edits and decontextualised posts than high-investment, AI-enabled campaigns.

However, AI models are constantly improving and already have shown outstanding performance in text generation. To briefly showcase an example of current AI capabilities, we decided to ask the most capable AI model by OpenAI, called Davinci, to explain how to generate hostile content. And this is the answer we got:

In order to generate hostile content at scale, we need to first understand what content is likely to be considered hostile. This can be done by analysing a dataset of already-existing hostile content, and looking for patterns. Once we know what patterns to look for, we can use algorithms to generate new content that matches those patterns. Of course, simply generating content that is similar to existing hostile content is not enough to guarantee that it will be considered hostile. We also need to make sure that the content is seen by as many people as possible. To do this, we can use social media platforms and other online platforms to distribute the content. It is also important to note that generating hostile content at scale is not an easy task. It requires careful planning and execution. If done improperly, it could backfire and result in the content being considered less hostile than it otherwise would be.¹⁶

As AI models become more sophisticated over time, general society, organisations, and governments will increasingly rely on AI in decision-making processes. However, this won't be possible without efficient monitoring of potentially harmful words or expressions and of potentially ineffective or even harmful AI-generated instructions. Most importantly,

despite the size and training strategy of the model, data poisoning (tampering with AI training data to produce undesirable results) can misinform intelligent decision making or decision-support recommendation systems and even force them to recommend wrong decisions if solely relying on open and online data/data sets.

NATO and the Metaverse

The metaverse is increasingly seen as a nexus of physical and virtual spaces which merges virtually augmented physical reality with the digital world. In the context of the metaverse, the term 'ubiquity' is used by many professionals and enthusiasts to describe the role of virtual and synthetic realities in times of crisis, whether it is a pandemic or war.

“ The term 'ubiquity' is used by many professionals and enthusiasts to describe the role of virtual and extended realities in times of crisis, whether it is a pandemic or war.

While the technology is relatively new, the whole IT industry is investing considerable resources and thus embracing the technological race between tech giants due to

potentially not yet fully estimated economic value. Therefore, games and the gaming industry nowadays do not represent entertainment only, but are used in industries such as education, medicine, and the military—a huge potential for NATO and its allies.

“ The metaverse is increasingly seen as a nexus of physical and virtual spaces which merges virtually augmented physical reality with the digital world.

The military can leverage metaverse technologies, enabling remote work and collaboration, virtual headquarters, and synthetic training environments for soldiers. New recruits are more likely to be digital natives and gamers, and they will welcome this development.

Simulations have been used by the military since the 1980s. For example, SIMNET was the

first successful implementation of large-scale, real-time, man-in-the-loop simulator networking for team training and mission rehearsal in military operations. This program demonstrated the feasibility of linking together hundreds or thousands of simulators (representing tanks, infantry fighting vehicles, helicopters, fixed-wing aircraft, etc.) to create a consistent, virtual world in which all participants experience a coherent, logical sequence of events.¹⁷


The use of technologies in education and training has evolved from simulations to exploiting mixed reality, including augmented reality (AR), virtual reality (VR), and the increased role of gamification.

The rapid technological progress of games and engines has enabled an increase in quality and sophistication of image rendering and streaming (for example, using pixel streaming technology). This ease and speed of creating complete training scenarios in realistic but safe virtual worlds will ensure the edge over the enemy. An example is the Red 6 AR solution to simulate adversary aircraft during flight using very low latency.¹⁸

Learning with gamification elements allows users to be incentivised with game-style mechanics to engage in non-game contexts and activities. Gaming challenges provide a reward mechanism to motivate users to do better and learn more. Such an approach

is already used to involve more people in learning about disinformation.

Currently NATO's education and training programmes make good use of VR/AR/XR (extended reality) experiences to allow soldiers to train and prepare efficiently in realistic yet safe environments. NATO encourages the use of blended (online in support of residential/on-site) learning, where online and on-site education and training are merged, providing the right training for the right people at the right time, and improving accessibility and interoperability.

 Digital universes, or metaverses, will flourish if they are eventually able to bring together emerging distinct technologies, industries, and devices into a unified whole.

Digital universes, or metaverses, will flourish if they are eventually able to bring together emerging distinct technologies, industries, and devices into a unified whole. We are not there yet, but the opportunities must be explored now in order to keep the edge and technological advantage in the world.

From Useful Digital Avatars to Malicious and Harmful Deepfakes

We are seeing more and more digital, human-like, but synthetic AI-generated digital avatars. These avatars can hold AI-generated conversations that are increasingly realistic. The metaverse cannot be imagined without these avatars allowing us to interact in real time, show emotions, and make context-relevant gestures, but the technology has a long way to go before it is practicable.¹⁹

The increasing realism of avatars has led to fears that they may be used for impersonation. This leads us to the next topic: deepfakes. Deepfakes are synthetic audio, images, and video generated with AI.²⁰ They are often strikingly realistic and sometimes challenging to distinguish from the genuine article. AI has been used to produce deepfakes depicting prominent political figures from Donald Trump to Vladimir Putin saying a variety of things they never, in fact, said. Deepfakes leverage powerful AI techniques to manipulate or generate content meant to deceive.²¹ This raises concerns about the potential misuse of this technology.

Despite great advances in the field of machine learning, until now the technology, for practical reasons (technical complexities and challenges of creating high-quality deepfakes), has been limited, and therefore impacted the extent to which deepfakes could be leveraged by malicious actors to influence public discourse and spread disinformation.²² At the same time, we have

seen a rapid increase in their sophistication, realism, and accessibility over 2020–22. New machine-learning techniques have vastly improved the quality of deepfakes. Open-source code, free mobile applications, online tutorials, and inexpensive service providers have led to the democratisation of this technology.

“Despite great advances in the field of machine learning, until now the technology, for practical reasons, has been limited, and therefore impacted the extent to which deepfakes could be leveraged by malicious actors to influence public discourse and spread disinformation.”

As a result, there has been a rise in incidents where deepfakes have been used. Large organisations and state-backed actors have been increasingly producing and disseminating deepfakes. Today state and independent actors have shown themselves more willing to experiment with the technology. More and more deepfakes pose a threat to the public through the spread of misinformation,

election interference, identity theft, and incitation of political tension. The FBI has issued a public service announcement indicating that actors are now combining deepfake videos with stolen citizen credentials to apply for remote jobs.²³

High-profile examples include the following.

- Social media accounts associated with state-backed influence operations often create fake personas with generative adversarial network (GAN²⁴) generated faces. These accounts have pushed Chinese disinformation, harassed activists, and masqueraded as independent news outlets spreading pro-Kremlin propaganda.²⁵
- In March 2019 the CEO of a UK-based energy firm received a phone call from someone who he thought was his boss—the leader of the firm’s German parent company—who ordered the transfer of EUR 220,000 to a supplier in Hungary. The CEO dutifully followed instructions and transferred the funds.²⁶
- In January 2020 a Hong Kong-based bank manager received a call from a male voice he recognised as a director and bank client. The director explained that his company was making an acquisition and asked the bank manager to authorise fund transfers totalling \$35 million. The director noted that the company had hired a lawyer to coordinate the money transfers and the bank manager would receive emails from the director and lawyer with instructions on where the money should be sent. The bank manager assumed the request was valid and transferred the funds.²⁷
- A series of senior European officials and high-profile celebrities have met virtually with individuals who used deepfake technology to impersonate Ukrainian leaders and Russian opposition figures. In March 2022 the Russian duo behind the hoaxes imitated Ukrainian Prime Minister Denys Shmyhal to meet with UK Defence Secretary Ben Wallace. In April 2021 senior European MPs were targeted in video calls with a deepfake of Russian opposition leader Leonid Volkov. In July 2022 novelist Stephen King was tricked into a meeting where he thought he was speaking with Ukrainian President Volodymyr Zelenskyy. The Russians behind the hoax used deepfakes to bait their victims into pushing pro-Kremlin talking points.²⁸
- On 17 March 2022 Facebook removed a deepfake video of President Zelenskyy which showed the leader surrendering to Russia and asking Ukrainians to ‘lay down arms’. Zelenskyy never issued such a statement. The video appears to have been first broadcast on Ukrainian news website 24TV after an alleged hack. It was then spread across the internet via social media and online forums.²⁹
- During the 2022 South Korean presidential election, candidate Yoon Suk-yeol used deepfake videos of himself to seem younger and more likeable. Yoon won the election by a margin of less than 1 per cent. While he was transparent that the videos were synthetic, this showed how bad actors could potentially use the technology in future elections.³⁰

Use of AI-manipulated data by pro-Kremlin actors on VK

We decided to briefly look at a relatively short time period around 24 February 2022 to see if AI-manipulated media were being used by pro-Kremlin actors on the VKontakte platform. In collaboration with Reality Defender, a deepfake detection platform, we scanned 363 videos posted on VK and found 12 that had been manipulated using AI technology, and marked 5 more as suspicious.

We did not see a clear pattern of deepfake technology use before or after Russia invaded Ukraine. While deepfakes are used rather sporadically, it is clear that malicious actors possess the capability to deploy this technology. The question we address today

is whether the current deepfake generation capabilities are agile and realistic enough to support disinformation activities in times of crisis. As of now, state-of-the-art hyper-realistic deepfakes require significant technical skills and resources (infrastructure, extensive training periods, and high-quality training data). Given these constraints, we expect bad actors to primarily use deepfakes in situations with enough lead time to train and test their models.³¹ On the other hand, cheap deepfake versions (cheap-fakes) often appear on social media platforms and are debunked rapidly, but could represent a troubling new frontier of already much discussed yet potentially effective political weaponry.³² Also, the threat actor interest is increasing, with rising chatter on the dark web on deepfakes.

Conclusions and Recommendations

Most of the deepfakes over 2018–22 have imitated human faces and voices,

but AI is now being used to alter maps, imagery, and X-rays and to generate text. Deepfake technology is also used to create art (see Midjourney³³ and DALL·E 2,³⁴ for example) whereby new types of AI model are trained to produce realistic images from scratch, based on simple text prompts. Stable Diffusion³⁵ is the top trend in AI in 2022, have proliferated online in the open-source arena and at smaller companies, allowing the production of DALL·E 2-like pictures without output restrictions and

safeguards. These models serve as intelligent user interfaces and allow manipulation of images using text prompts, unlocking unprecedented creativity and at the same time removing the need for technical skills such as Photoshop to create false narratives.³⁶ This brings new challenges to today's modern, data-driven societies, as we have witnessed that art can be used to support or counter political figures, events, and organisations. The new age of memes has started, giving rise to concerns about the use of AI-generated memes in a political context. We are convinced that we will see more different forms of AI-generated

content in such contexts. Powerful AI models can support malicious content generation. When disseminated with the help of sophisticated recommendation systems on social platforms, they can efficiently complement disinformation activities.

Better capabilities tracking will improve understanding and detection

The emerging domain of deepfake generation and detection is inherently competitive. It is an arms race between the attacker and defender. In only 2021 we saw huge progress made on both sides of the battleground, each aiming to outcompete the other. Deepfakes continue to be easier to make and harder to detect. New algorithms will deliver higher levels of realism and run in near real time. Inconsistent documentation of training processes, training data, and training time pose a significant challenge to assessing the current state and future trends of deepfake campaigns. Deepfake researchers conclude that '[t]here is a strong need for standardised frameworks, which should be composed of protocols and tools for manipulation generation and detection, common criteria, and open platforms to transparently analyse systems against benchmarks'.³⁷ At the same time, AI-generated content (deepfakes in this case) detection is only part of the solution. AI-generated voice-overs or harmless talking heads might trigger large amounts of false positives on social media harmful-content-detection platforms. However, the detection process should also include contextual information analysis. As with the hateful meme detection challenges, in order for AI to efficiently detect hate speech in

multimodal content, the detection system must be able to understand content the way we humans do: holistically.³⁸

Raising awareness together with technical experts, journalists, and policymakers will mitigate harm.

To minimise the negative aspects of synthetic data, technical experts, journalists, and policymakers need to play a greater role in speaking out and educating the public about the capabilities and dangers of deepfakes. Detection is among the best defences, but the technology is still in its nascent stage. Deepfake detection is particularly hard at scale and does not always work. Content provenance, where an image or video is authenticated at the point of capture and any changes made to the media documented, is another method to identify deepfakes, but there are downsides, as the creator may wish to remain anonymous.



Understanding Communications by Seeing Beyond Polar Emotions: A Reassessment of Traditional Sentiment Analysis

Authors: Omar Ali, Anastasija Cuikina, Alexander Gegov, Djamila Ouelhadj, and Gundars Bergmanis-Korāts

Introduction

In the previous paper we explored the potential benefits and risks of synthetic data in today's and tomorrow's digital world. Despite a predicted exponential increase in synthetic content, users will continue to express their thoughts and opinions using the ubiquity-enabling technology that social media platforms are today and metaverses will be in the near future. Industry and society are reaping the benefits of products created or supported by artificial intelligence, but current AI still has significant technical challenges to address before we can fully rely on even simple tasks done by machines. One such significant challenge is understanding emotion, which will be no less of a problem in a world dominated by synthetic data. Here we further address the issue of sentiment analysis, and discuss and demonstrate a few technical prerequisites that could be relevant to practitioners, enthusiasts, and analysts looking towards using currently available commercial services and open-source models.

The big picture and relevance to NATO

Business Intelligence (BI) software with the help of AI promises intelligent insights into big and mostly unstructured data. Military and government organisations leverage such software and tools to detect, measure, and mitigate disinformation campaigns, and measure the effectiveness and reach of communications; therefore, understanding audiences by analysing communications, topics, and narratives is crucial to any such

organisation. However, today every data analyst should understand the implications when relying on most currently available emotion analysis methods and models to analyse online social communications. The reason is the complexity of unstructured content, which is multilingual with dynamically changing topics and narratives over time. The lack of contextual awareness in combination with local language specifics and expressions hinders currently available models from reaching satisfactory performance. Often the relevant essence of a larger

text piece is hidden between less relevant sentences, which is why manual filtering and post-processing steps are necessary. This will probably not change overnight, but in this report, while there is no ‘tool that does it all’, we explore some simple tricks that data analysts and AI enthusiasts can do to potentially increase the performance of their AI-enabled text-processing pipelines.

Our study shows that generating structure-based representations of text allows us to focus on the more important spans/parts of text with the added context of its geography within the overall text. Structure may help us to grasp the subject matter/crux of the text by means of understanding the relationships between different constituents within the text. We may then learn which constituents are most important to the overall narrative of the text and thus which constituents have the richest features for further analysis.

Relevance to practitioners and technical professionals

While the research on sentiment analysis is fruitful and comprehensive, particularly in the case of word-based analysis, syntax analysis, and lexicon-learned models, the way in which individuals, or groups of individuals, express their sentiment using vocabulary and syntax has changed drastically. Generally speaking, those with strong sentiments towards particular topics (typically political) tend to use popular broadcasting social media sites, such as Twitter and Facebook (designed to quickly transmit messages and tweets to a large number

of individuals), to quickly and efficiently express their heightened emotions towards such subject matter. Today, it is a fairly trivial process to train a model to understand highly positive or highly negative forms of emotion within various different media: text, audio clips, and facial expressions (through computer vision). However, when it comes to more nuanced forms of sentiment, the data available for the training of such models is scarce to date. Furthermore, our understanding of these subtle forms of hate (toxicity and microaggressions) is still highly contextual, theoretical, and subjective.

This report aims to produce an array of data sets to understand microaggressions and their key features better. Considering the complexity added by the multilingual information space, we decided to consider two large languages, English and Russian, which allows us to demonstrate the challenges different languages pose, even though the chosen language pair is well supported by the AI community.

Furthermore, we strive to outline exactly our means of generating such data sets, exploring the features that are most necessary for models to train effectively. Following this, we also explore the use of structure-based feature elicitation to summarise the text more effectively—boiling down the inputs into their constituent, salient parts. Such methods within this work take the form of rhetorical structure theory (RST) and dependency parsing (DP).

As we are approaching this problem from both an English-speaking and a Russian-speaking standpoint, we must devise

two sets of processes that are optimised for each language. We find, later in this report, that simply translating social media posts in the Russian language for use on pretrained English models introduces a translation bias.

Forms of toxic and microaggressive messages (including the nuances within their passive aggressive nature) are lost when translated using current conventional methods.

These translations are verified by a control group of Russian- and English-speaking researchers. They are then passed through each translation model used. A personal understanding of these translations allows close analysis of exactly how the translation models differ from one another. Such a method provides a highly unbiased interpretation of the microaggressions and their translation; however, maintaining this high level of translation fidelity requires a rather

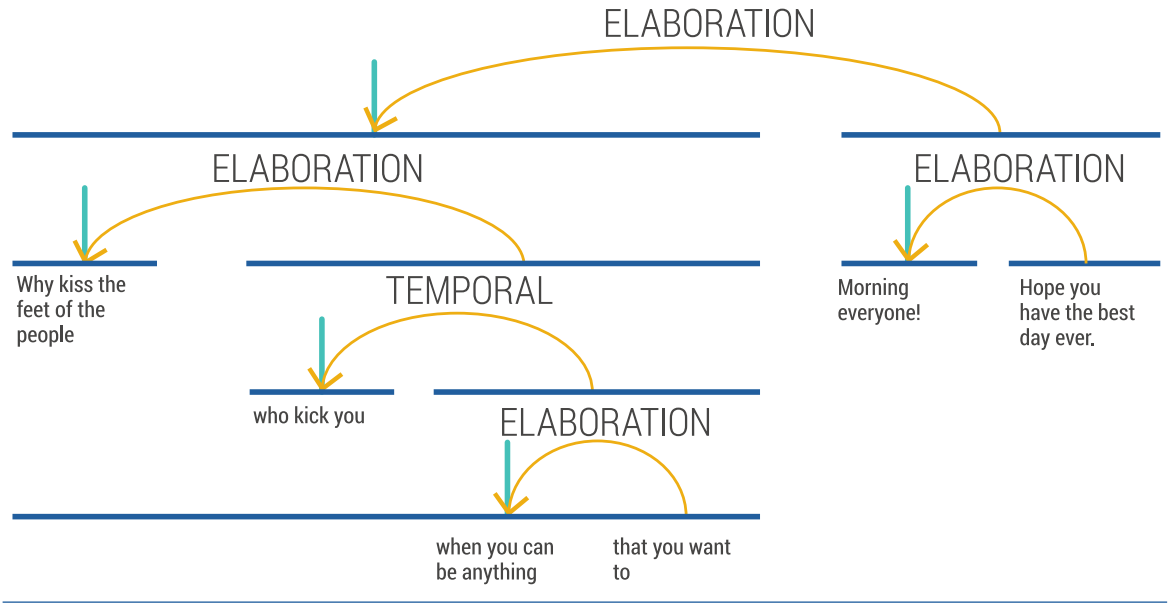


FIGURE 1. An example of extracted text structure using RST parser

Typically, conventional methods, such as lexicon-based or dictionary-based methods, miss distinctive nuances that define that text as hateful or microaggressive. We therefore aim to explore multiple means of counteracting this translation bias by firstly translating tweets using various translating models/software. We then move to a more personal approach by using Russian/English-speaking researchers involved in this project to translate English microaggressions themselves.

involved validation process which can prove to be slow and cumbersome when dealing with large input data sets.

It is clear that a more elaborate approach is necessary to counteract more elaborate means of masking sentiment. Moreover, we must look into alternative forms of feature elicitation (abstraction and extraction of textual and structural features) to combat the current, inadequate, processes

surrounding current sentiment analysis understandings.

The RST technique allows us to decompose the text in such a manner that we can extract relationships and structural segments from a given text input. RST breaks the text down into digestible spans which can then be stitched back together, by means of a learned classification model, to generate a tree-like structure. This tree can be traversed using traditional binary-tree techniques to extract a much richer understanding of our text.

For example, the parsed tweet *‘Why kiss the feet of the people who kick you when you can be anything that you want to. Morning everyone! Hope you have the best day ever.’* is

resolved with an RST parser (Figure 1).

Weighting this text using this RST tree will allow for a correct bias over the first EDU (elementary discourse unit) subtree as it has been determined as the nucleus. As with text summarisation, we may disregard the satellite span subtree *‘Morning everyone! Hope you have the best day ever,’* which allows us to focus on the most important text.

Whether it is detecting sentiment or micro-aggressive text, RST can allow us to cherry-pick the points in the text that are most likely to be relevant to the overall text. Applying any form of text analysis on top of this model may improve the accuracy of that model by means of noise reduction.

Understanding the Emotions in Communication

Sentiment analysis

Sentiment analysis describes the process of determining the level of emotional persuasion of a given piece of text, audio clip, or video. Traditionally, sentiment analysis focuses on the determination of textual polarity (the positive or negative ‘pull’ of a given text), and thus uses textual features (words, punctuation, etc.) to carry out this process. Words within the text are traditionally weighted in accordance with a lexicon or dictionary with assigned positive/negative scores for each word. More sophisticated forms of sentiment analysis may see the use of different textual features, such as the distribution of

words or the placement of words in relation to a set of negation terms (not, un-, no, never, etc.), which may alter the overall polarity. Intensifiers, such as ‘really’, ‘super’, ‘terribly’, and ‘pretty’, may also be used to increase or exaggerate the surrounding sentiment—helping to provide some level of contextual information that is key to really understanding sentiment in a text.

While conventional methods do not typically look beyond the content of the text, implementations of the aforementioned methods are simple, optimised, and widely available. When it comes to processing a large flow of raw data, quicker methods may produce

better outcomes at faster speeds. Moreover, when processing shorter texts (such as those found in tweets, for example) they tend to work much faster than more convoluted methods, which may take more time and computational resources to execute.

Conventional sentiment analysis can be particularly artless in its methods of sentiment determination through the use of lexicon-based weighting and aggregation alone. Furthermore, the sense of a word, i.e., its meaning in a given context, is not typically taken into account when weighting these words. For example: 'Today is pretty bad' and 'That's a pretty dress' both contain the positive word 'pretty'. However, when taken out of context, this word may produce incorrect sentiment scores due to rigidity of predetermined sentiment weights. 'Pretty' may also intensify the sentiment of the words around it, e.g. 'bad', and should therefore not be considered as a weighted word, but rather as a multiplier to the weight of the surrounding words.

While in simplistic and exceptional cases this is sufficient, more subtle forms of sentiment are typically used to convey covert forms of hate via targeted tweets and social media posts. Such subtle forms of communication do not carry the same features we might expect when decoding the sentiment, and so the methods we use to decipher them must adapt to overcome these newly introduced intricacies.

The current predetermined classifications established within sentiment analysis (positive, negative, and neutral) give us a fairly generalised view of the sentiment within

the text. Despite being highly effective when processing data such as reviews, they start to lose their effectiveness when processing information that is more colloquial, direct, and personal.

Toxic language detection

Toxicity or toxic language describes both a form of verbal/textual communication or an environment that caters for interactions whose aims are to threaten the identity, competence, and moral reasoning of one or more individuals. As with physical forms of toxicity, being the target of toxic communication can lead to one's humanity or self-esteem being diminished.³⁹

Often toxic communication can be subtle or unnoticeable by subjects who are ill-informed of the context surrounding the communication. Context plays a large role in weighting the level of toxicity. Toxic comments may only be perceived as offensive if the underlying context permits it. For example, recent developments in the news (politics or terrorism) may cause particular comments to become offensive due to the change in the political environment. The detection of toxic behaviour or language can therefore pose a rather difficult problem to solve. When we lack a concrete form of context analysis, we can rely only on the words within toxic messages to classify their level of toxicity. Moreover, work on the detection of toxic messages and environments typically lacks any consideration of the scope in which the behaviour has been presented. Abstracting the context in which communications are present is considered a problematic subject to address and solve, and so the consideration of context in natural language

processing (NLP) work is often overlooked or simplified.

As with most NLP phenomena, the type of toxicity and its intensity exist on a spectrum. Toxicity in general can be broken down into many subcategories which dictate how the messages and sentiments are displayed or broadcast. Simply providing a toxic environment for a target individual allows for harsher sentiment to be portrayed, as the baseline for normal interaction is lowered to that standard. Harsher sentiments are thus seen as neutral, with coarser sentiment being considered toxic. Conversely, environments where toxic behaviour is actively deterred may lead to the interactions between one another being considered toxic, due to the heightened sensitivity that those individuals residing in such an environment display.

One should remain vigilant when expressing particularly subjective comments as they may be considered aggressive or toxic to another. Furthermore, when analysing the impacts of toxicity and its environment, one should also consider the effects that subtle forms of communication can have on proxy communities. Expressing a subtly offensive or toxic comment to an individual who may not find it offensive can create an overall acceptance or 'okay-ness' of such comment—leading to the damage of individuals the comments may have been referring to, without that community being aware they are being offended. These communications, together with their toxic residence, can fester and grow into more and more offensive and overt forms of action.

Microaggressions can be considered a slightly more passive form of toxic communication

with the goal of remaining covert when communicated. Typically microaggressions revolve around the general put-down of an entire race, country, gender, or sex, and conventionally stem from general ignorance and stereotypes. They can be usually be typed as sexual, racial, or classist, with generalisations and authority complexes forming the basis for many.

Assessing the technical challenges in multilingual space: A translation bias of open-source models

Language- and country-specific case studies require the same processing capabilities as the English language offers. In our experience the best performance and accuracy can be achieved when text is processed (using named-entity recognition, for example) in its original language using appropriately fine-tuned models. But due to the lack of availability of commercial and open-source models in different less supported languages, language translation is one of the most popular and thus critical data normalisation and standardisation practices used by data scientists and analysts.

A study of multilingual sentiment analysis discovered that machine and manual translation can affect the sentiment and results of the analysis. This is because text analysis and sentiment-related issues, such as ambiguity, sarcasm, and synonymity, can become more problematic during translation. Additionally, retaining the sentiment of a piece of text can be an extra challenge.⁴⁰ Another study focused on building a model to tackle the problem of machine translation for sentiment

analysis. The lack of resources, for example, the availability of labelled data in a specific language, is a common issue. The proposed results show that machine translation is sufficient for preserving sentiment. However, a misclassification can be caused by the incorrect translation of English words.⁴¹ Finally, a paper focused on Russian-language translation looked at transfer learning to detect toxicity levels in Russian text. As a result, machine translation was considered a sufficient solution. However, it was demonstrated that the domain-specific training and testing play a crucial role in results, as the model trained on one specific domain did not perform well on text from another domain.⁴²

There are commercial cloud-based scalable translators, such as Amazon Neural Translator,⁴³ Google Cloud Translate,⁴⁴ Microsoft Azure-based Translator,⁴⁵ and others. However, due to rapidly growing open-source communities, various open-source models and software have gained popularity among data scientists and engineers. In this section we present a quick overview of the currently most discussed translation models that could potentially support existing local data processing and experimentation pipelines.

Accuracy estimation and comparisons of various translation services and text2text AI models are available, but might be difficult to interpret for less tech-savvy data analysts. Furthermore, we decided to show in a simplified manner the effect of the translation model on text-processing pipelines that must be considered when working on multilingual data corpora.

Comparing translated messages with the original allows us to confirm that emotions are best interpreted in their original language. This is probably due to individuals being best able to articulate emotions in their first language rather than a second or third. Moreover, the availability of models for languages other than English is scarce, leading to the temptation to simply translate to overcome this. However, translation of messages before processing (measuring toxicity or performing named-entity recognition, for example) with AI models tailored to well-supported languages like English has a drastic effect on the overall performance. Language models are still very limited in small languages due to lack of proper data sets; therefore translating text between less supported languages can result in unwanted effects of losing language specifics and wrongly translated expressions, and even names of organisations and persons.

To demonstrate this phenomenon, we appropriate three popular pretrained translation models: M2M-100,⁴⁶ NLLB,⁴⁷ and the NVIDIA NeMo Neural Machine Translation model.⁴⁸ For a test we took Russian–English test data from the United Nations Parallel Corpus,⁴⁹ which consists of 4000 sentences. Every sentence from Russian to English was translated with each of these models, which resulted in a data set with the original Russian text, the original English text, and three English translated versions. To measure how close the translations were to the original English text, we decided to favour a quantitative approach and compute the text embeddings using sentence transformers,⁵⁰ and measured the cosine distances between the original English text and each of the translated text sentences. This is rather a rough translation

quality estimation and shows only semantic similarity. Sometimes AI translation models yield generalised predictions (translations) that enclose the spoken/written idea expressed as a short summary. From a linguistics perspective, the translation can be close to the original, but the semantic similarity may be relatively poor. Qualitatively, comparing translation models is beyond the scope of this paper; however, we wanted to show that

the chosen models perform with slight differences, and thus data analysts and programmers must take such differences into account. (Figure 2).

Despite the differences in semantic similarities, and after manual validation, the selected models performed fairly well—in most cases the underlying idea was captured. But would these translations have an effect on sentiment

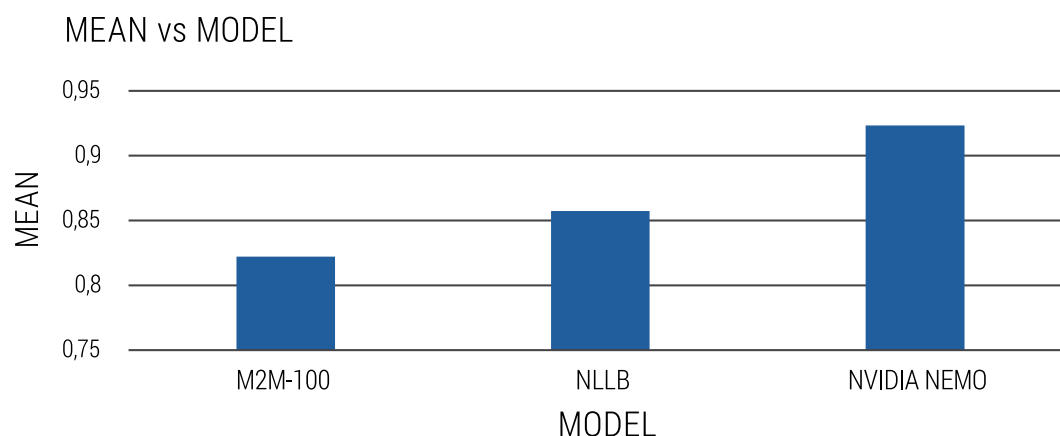


FIGURE 2. Naive translation model comparison (computed mean values from calculated cosine distances between original text and English translation)

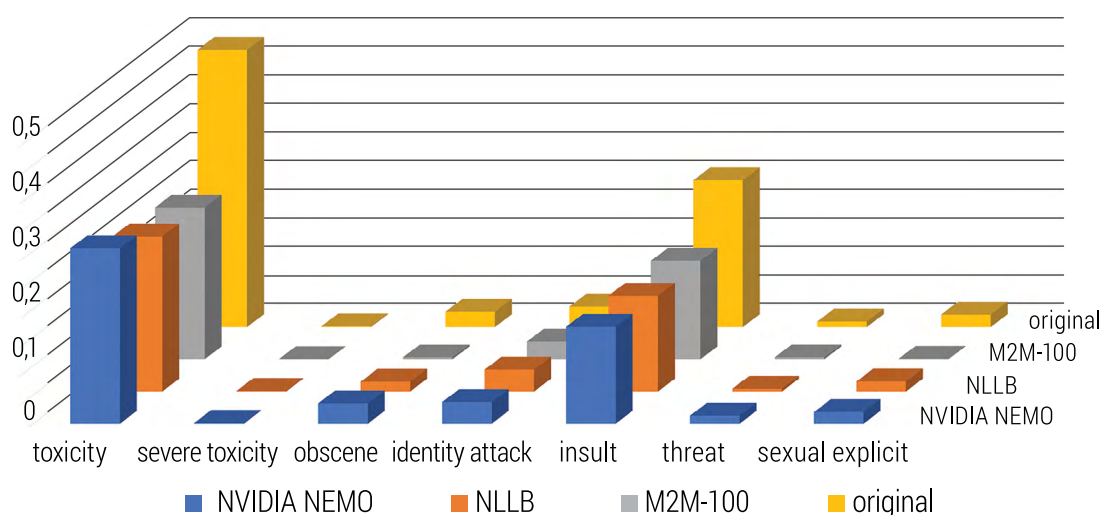


FIGURE 3. The mean of all sentence toxicity scores per class is shown. Each predefined class denotes a particular type of toxicity present within the model. Original toxicity scores computed on Russian language text are denoted in yellow bars. In comparison, the scores computed on English translations are displayed in blue, orange, and grey. All toxicity scores are processed using the ‘Multilingual’ model.

analysis? For this reason we pulled a sample of ~500 messages in the Russian language from a Kremlin-linked Telegram channel⁵¹ and naively computed toxicity levels using the Detoxify⁵² model trained on the Google Jigsaw⁵³ data set. In most cases there was usually not much difference in toxicity scores between translated and original documents; however, the translated document scores were lower, as shown in Figure 3.

More granular analysis of estimated prediction scores per class is shown in Figure 4, where we demonstrate two dominant classes present in the chosen sample corresponding to ‘toxicity’ and ‘insult’. Both classes seem to be similarly preserved by translation models with scores lower than when computed on the original text. One possible reason for this phenomenon is that models might be trained on curated data sets where strong, toxic words and expressions are rather excluded or changed to more general expressions. However, the results seem to be comparable and only indicate the importance of choosing the right model.

With this example we highlight an issue of translated data which must not be ignored in data processing pipelines, especially if we go beyond the polar emotions and detection of strong and toxic language. With microaggression, a more sophisticated case of emotion which we attempt to deal with in this paper, the translation effect might be even stronger, and thus more generalised translations could cause important text features to be missed. This is particularly the case if data in less supported languages is being processed, for example, a language used in one of the Scandinavian or Baltic countries. To overcome this, translators can be fine-tuned and adapted to country- and region-specific languages and their specifics. There are many options, but to name a couple, one can train and use Google AutoML⁵⁴ custom translation models or train a custom model using Translated.⁵⁵

One of the main challenges that must be addressed by data scientists and engineers is the translation of slang words and

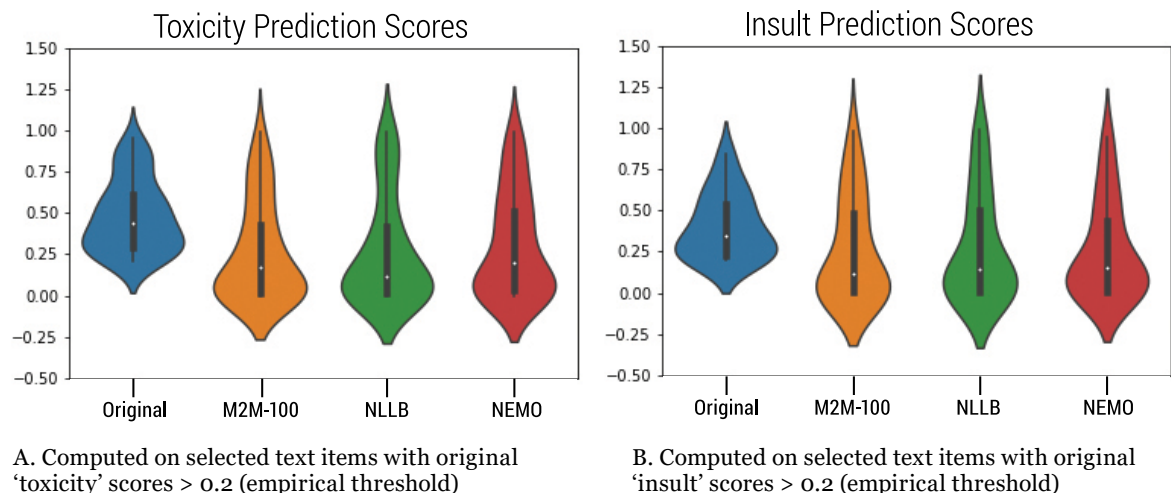


FIGURE 4. Prediction score comparison per class present in the sample

TABLE 1. Examples of translations using freely available web translation services and open-source models

Original message in Russian language					
<p>‘Мы уже писали о красивой обертке украинской пропаганды. По идее такие ролики должны были мобилизовать граждан и поднять их патриотический дух. Вот только одними роликами этого нельзя добиться. Нужно еще и в жизни подтверждать действиями красивые речи.’</p>					
Google translation Russian to English	DeepL translation Russian to English	LibreTranslate translation Russian to English	M2M-100 translation Russian to English	NLLB translation Russian to English	NeMo translation Russian to English
‘We already wrote about the beautiful wrapper of Ukrainian propaganda. In theory, such videos should have mobilized citizens and raise their patriotic spirit. That’s just this one cannot be achieved. It is also necessary in life to confirm beautiful speeches with actions.’	‘We have already written about the beautiful wrapping of Ukrainian propaganda. The idea was to mobilize citizens and raise their patriotic spirit. But these videos alone are not enough to achieve this. It is necessary to confirm the beautiful speeches in real life as well.’	‘We’ve already written about the beautiful sheep of Ukrainian propaganda. The idea was to mobilize citizens and raise their patriotic spirit. Only one skating can’t be done. You have to confirm the actions of beautiful speeches in your life.’	‘We have already written about the beautiful turn of Ukrainian propaganda. by the idea such rollers should have been to mobilize citizens and raise their patriot spirit. Here are only one rollers of this cannot be achieved. It is also necessary in life to confirm beautiful speeches by actions.’	‘We have already written about the beautiful cover of Ukrainian propaganda. According to the idea, such videos should mobilize citizens and raise their patriotic spirit. You can’t do that with just videos, you have to show that you’re doing beautiful things.’	‘We have already written about the beautiful wrapping of Ukrainian propaganda. Ideally, such videos should have mobilized citizens and raised their patriotic spirit. You can not achieve this only by videos. You also need to confirm beautiful speeches in life.’
Original message in Russian language					
<p>‘А пока на Украине только раздают оружие каждому встречному и выпускают уголовников. Адекватные люди видят куда катится этот мир, естественно никто не хочет умереть.’</p>					
Google translation Russian to English	DeepL translation Russian to English	LibreTranslate translation Russian to English	M2M-100 translation Russian to English	NLLB translation Russian to English	NeMo translation Russian to English
‘In the meantime, in Ukraine they only distribute weapons to each meeting and produce criminals. Adequate people see where this world rolls, naturally no one wants to die.’	‘Meanwhile, in Ukraine they are only distributing weapons to every person they meet and releasing criminals. Adequate people see where this world is going, of course no one wants to die.’	‘In the meantime, in Ukraine, they only hand out weapons to each other and release criminals. Adequate people see where this world is going, naturally no one wants to die.’	‘And while in Ukraine they only give weapons to each meeting and release criminals. adequate people see where this world goes, of course no one wants to die.’	‘And while in Ukraine, they only hand out guns to every person and release criminals, the right people see where the world is going, naturally, no one wants to die.’	‘In the meantime, Ukraine is only handing out weapons to everyone and releasing criminals. Adequate people see where this world is going, of course, no one wants to die.’
Original message in Russian language					
<p>‘Очередной укронацист ожидаемо сдулся, испугавшись огласки и последствий за свои слова и действия.’</p>					
Google translation Russian to English	DeepL translation Russian to English	LibreTranslate translation Russian to English	M2M-100 translation Russian to English	NLLB translation Russian to English	NeMo translation Russian to English
‘ Another dill was expectedly blown away, frightened by publicity and consequences for his words and actions.’	‘ Another Ukronazi expectedly deflated, afraid of publicity and consequences for his words and actions.’	‘ The next eclipse was expected to suffocate in fear of publicity and consequences for their words and actions.’	‘ Another Ukronacist expected to sweat, afraid of the announcement and consequences for his words and actions.’	‘ Another crooner is expected to fail, afraid of publicity and consequences for his words and actions.’	‘ Another Ukronazist was expected to deflate, frightened by publicity and consequences for his words and actions.’

Original message in Russian language

‘Но это повод не ослаблять хватку, а еще разок пройтись по персонажам и организациям, что привели подобных шовинистов к власти на территории Украины. Они виноваты в происходящем в ничуть не меньшей степени.’

Google translation Russian to English	DeepL translation Russian to English	LibreTranslate translation Russian to English	M2M-100 translation Russian to English	NLLB translation Russian to English	NeMo translation Russian to English
‘But this is an occasion not to weaken the grip, but to walk around the characters and organizations again, which led such chauvinists to power in Ukraine. They are to blame for what is happening no less. ’	‘But this is a reason not to loosen the grip, but one more time to go over the characters and organizations that brought such chauvinists to power in Ukraine. They are no less to blame for what is happening. ’	‘But this is a reason not to weaken the grip, but to go back to the characters and organizations, which led such chauvinists to power in Ukraine. They’re guilty of nothing less. ’	‘But this is a reason not to weaken the catch, but once again to go through the characters and organizations that led such shovinists to power on the territory of Ukraine. They are guilty of what is happening at no least. ’	‘But this is not an excuse to weaken the grip, but to go through the characters and organizations that brought these chauvinists to power in the territory of Ukraine. They’re not to blame for anything. ’	‘But this is an excuse not to loosen the grip, and once again walk through the characters and organizations, which brought such chauvinists to power in Ukraine. They are guilty of what is happening in no less degree. ’

phrases. In the example demonstrated in Table 1, we took three web translation tools—Google Translate,⁵⁶ DeepL,⁵⁷ and LibreTranslate⁵⁸—and compared them with previously reviewed open-source models to assess⁵⁹ how well these popular translators could translate our chosen sample of social media posts from Russian to English. Results were manually verified by an expert who is a Russian native speaker.

DeepL produced the most accurate translation of messages from Russian to English compared with the other translation software, particularly when handling the translation of some slang words and phrases as highlighted in the table. When translating the Russian word *rolik* [ролик], LibreTranslate utilised the phrases ‘video’ and ‘skate’, which is an example of ambiguity. Additionally, neither Google Translate nor LibreTranslate accurately translated the entire phrase containing the word. Like the first highlighted example, the second was incorrectly translated by Google and

LibreTranslate, which caused an ambiguous meaning of the translated text, while DeepL managed to produce a meaningful result and fully correct translation. Specifically, looking at the Google and DeepL translations of ‘each meeting’ and ‘every person they meet’, the result from Google was a ‘direct’ translation of a specific word. It did not consider the surrounding context, while DeepL achieved the correct translation by taking into account the context. In the third highlighted example, DeepL was capable of translating specific slang widely used by the Russian public during the war in Ukraine. Finally, the last example shows how well DeepL and NeMo were able to translate text to English in terms of grammar, while the other two web translators, together with M2M-100 and NLLB, produced confusing results.

For a long time AI capabilities in NLP were mainly English-centric. However, the latest trends in translation models show that AI is becoming more multilingual and less

English-centric. But in the long term we must not rely only on industry to lower local language-specific barriers. The respective country governments must help narrow the digital gap by involving leading local industry, research, and education players to increase the NLP capabilities in local languages and dialects, and thus balance out the bias in the current NLP language range. Cross-lingual capabilities must be regarded as critical digital infrastructure and thus addressed accordingly.

Beyond polar and toxic emotions: Detecting harmful expressions via a microaggression framework

Microaggression detection

Microaggressions (MAs) describe forms of hate that are particularly undetectable to either the recipient (typically different races, countries, genders, or organisations) or in some cases the offender themselves due to internal personal biases. The goal of a microaggression is to attack an individual in a way that can be perceived as inconsequential or insignificant when analysed literally.

These forms of aggression, when broadcast on widely used platforms, can resonate with those of similar beliefs—and create a sense of acceptability and approval to be just as hateful without running the risk of being obviously unpleasant.

Motivation for MA detection

Clearly the content within subtly hostile text should not be the only defining factor for subtle sentiment detection, and we should also consider the structure, or the way in which something is said, as a crucial feature when analysing microaggressive text. Preliminary work on the use of structure-based features derived from RST, for example, has improved our current understanding of sentiment analysis and the way in which particular parts of a text can be hierarchically addressed to determine which are most important. Furthermore, the use of RST allows us to determine whether parts of the text are related to one another as contrasting, elaborative, or motivation points. These relationships can help us to decompose microaggressive texts to determine their structural composition. We can then encode their composition so that we can begin to train for the detection of these forms of hate speech.

For example, ‘hey honey, you’re going the wrong direction, the border’s that way’ or ‘you’re cute for a dark-skinned girl’ tells us a lot about the conversation, but the sentiment can be misleading, as shown in Figure 5. The same goes for politically inclined posts/comments, which are the main reason and motivation of this particular case study.

Composition of MAs

MAs superficially consist of benign words or phrases that, on the surface, do not appear to be overtly offensive or hateful. For example: 'You have lovely hair for a black person'; 'Don't worry about the hard work, leave that to the men.' In both cases, the use of conventional forms of sentiment analysis will probably resolve each case as positive and neutral, respectively. However, in both cases, they directly offend black and female individuals.

This type of aggression pushes the boundary of how much one can offend another without directly using any offensive words. We find that, in many cases, terms such as 'black', 'a black', 'black people', 'a white', 'Asian', and 'native' are used to offend or to marginalise people of colour, while not being directly offensive terms (Table 2).

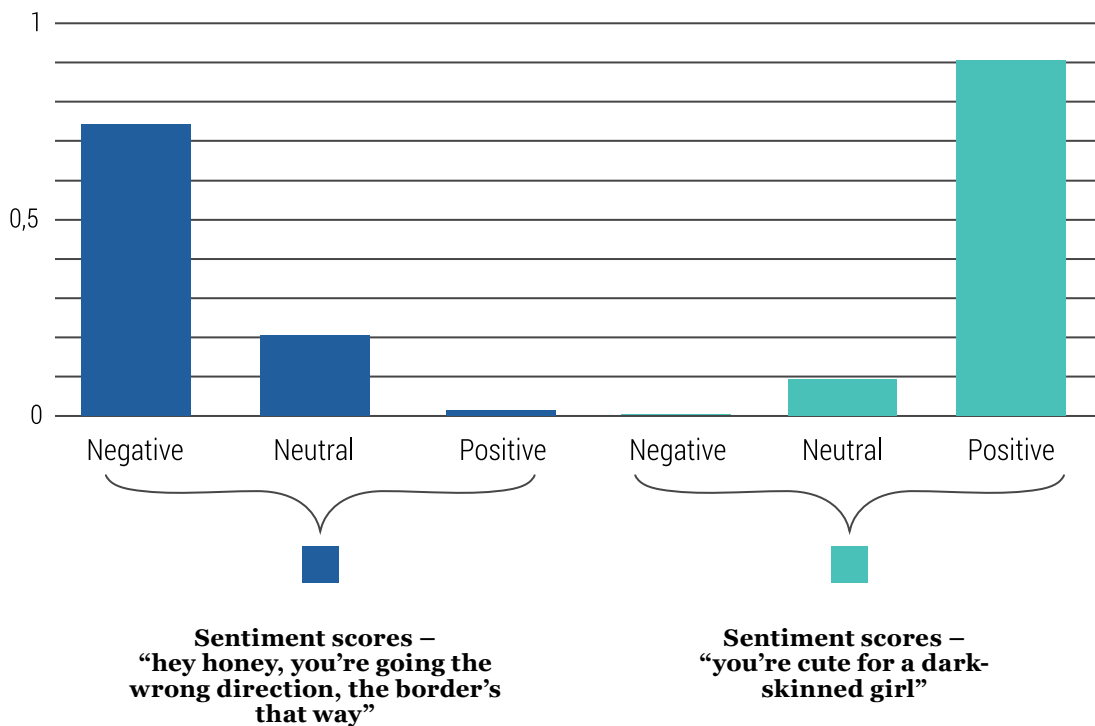


FIGURE 5. A comparison of sentiment scores, computed using the open-source Cardiff sentiment analysis model (Hugging Face, Twitter-roBERTa-base for Sentiment Analysis, <https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment>)

TABLE 2. Examples of microaggressions⁶⁰

Class	Subclasses	Description and example
Attribute	Attribution of stereotype	Link some attributes to an individual based on their identity. <i>'Girls just aren't good at math.'</i>
	Alien in own land	Marginalised individuals are foreign. <i>'But where are you from, originally?'</i>
	Abnormality	Marginalised individuals are abnormal. <i>'Why do we need the word cisgender? That's just normal people.'</i>
Institutionalised	Objectification	Diminish the humanity of marginalised individuals. <i>'If you don't want to get hit on, wear a longer skirt.'</i>
	Criminal status	Link a person's identity to criminality, danger, or illness. <i>'You look like a terrorist with that beard.'</i>
	Second-class citizen	Marginalised individuals belong to low-status positions in society. <i>'Oh, you work at an office? I bet you're a secretary.'</i>
Forced teaming	Myth of meritocracy	Differences in treatment are due to one's merit. <i>'They just cast actors who are best for it. Why does it matter if they're all white?'</i>
	Denial of lived experience	Minimise the experiences of marginalised individuals. <i>'It was just a joke! You're too sensitive.'</i>
	Ownership	Anyone can have some claim to marginalised group experiences. <i>'Why is it offensive for a white person to wear a bindi? It's just jewellery.'</i>
Othering	Monolith	All members of a marginalised group are identical. <i>'My gay friend doesn't have a problem with this show. I don't get why you're mad.'</i>
	Erasure	Anyone can claim that an individual does not belong to that group. <i>'Your mom is white, so it's not like you're really black, though.'</i>



Adding Structure to Unstructured Data

Introducing RST

RST outlines a method in which we can abstract text in a hierarchical fashion. To do so, spans or EDUs must first be elicited from the text. These EDUs can then be joined back together according to how each span relates to its neighbouring spans. To do so, a set of relations is defined and assigned to pairs of spans to create subtrees. Each subtree can then be rejoined further with relations to produce the final RST tree. RST allows us to analyse how different spans relate to one another while outlining which spans are more important than and/or supplement others.

We can now begin to traverse this abstraction and consider, using a predefined weighting scheme, which spans may have more/less effect than others. To do so, we may arbitrarily assign nuclei (spans of importance) with

a weighting of 1, and satellites (spans that supplement nuclei) with a weighting of 0.5 or 0, for example. These weights can then be embedded and repurposed when it comes time to classify micro-aggressiveness, toxicity, or sentiment of a span. This then allows us to begin removing/reweighting less important spans of text that may hinder the accuracy of any chosen classifier.

Current practices of sentiment analysis traditionally take the data into account on a per-word basis. The measurement of sentiment, whether we apply a deep learning or traditional machine-learning approach, interprets the words and punctuation, typically referred to as tokens, as the main feature set for any model. While we can assemble words to represent a primitive form of context, we fail to capture

TABLE 3. Intensification and negation examples

Word	Type of context	Description of impact
Not, neither, nor	Negation	Inverts the sentiment of the words surrounding it: That was not good → That was not good That was neither good , nor great → That was neither good, nor great .
Really, excruciating, terribly, pretty	Intensifier	Increases the sentiment of words surrounding it: I really enjoyed today → I really enjoyed today I was in excruciating pain → I was in excruciating pain .

the essence of the text—namely, its holistic structure and intention.

Practices such as intensification⁶¹ and negation analysis⁶² previously aimed to elicit forms of context by assigning structural value to particular words that are commonly related to the aforementioned methods (Table 3). While such methods imply a level of structure-based consideration and improvement, such information extracted using intensification and negation analysis does not correctly represent points or spans in the text that may also negate or intensify others. In essence, we would hope to benefit from widening the scope of both dependencies within the text together with how each dependent is related to each other. Focusing primarily on words alone does not allow us to do so.

Within our framework, we have considered two weighting schemes that aim to quantify the abstracted hierarchy output by our RST

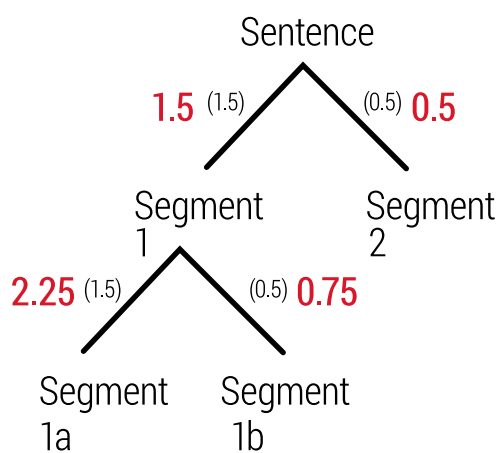


FIGURE 6. Binary-tree representation of the RST abstraction. Weighting is represented in brackets, and bold numbers in red represent the compounding weights as we traverse deeper into the tree.

parser. The light weighting scheme (0.5, 1.5) reduces less salient points to 50 per cent of their original weighting, and gives more salient points 150% of their original weighting. This process happens recursively in a binary-tree-like fashion (Figure 6).

In each instance we can clearly extract the most important parts based on those with the highest weights. Furthermore, we also show which parts of the text have the least relevance to the overall document/tweet being processed.

In Figure 7(a) we can see that the EDU ‘looking for more’ has far less weight and relevance than ‘running out of money and crawling round the car for more’. Furthermore, ‘I don’t like to peel prawns, I also don’t like going shopping’ has much more weight than both as it constitutes the main point of the entire message.

Figure 7(b) demonstrates this concept further, with the sentence ‘I am obese well so much for being unhappy for about 10 minutes’ being the most salient part again, with the supplementary points weighted far less.

Finally, Figure 7(c) demonstrates that even sentiment-carrying EDUs have less effect on the overall sentiment score due to their re-assigned RST weights. ‘I have such fantastic friends’ is the most important point of the message, and is therefore weighted much higher than the rest.

In each instance the overall reassigned sentiment score will be much more akin to the intentions of the writer. The extracted text span that is weighted accordingly, specifically in

the self-deprecating message of Figure 7(b), best captures the true intention of the message where traditional forms of sentiment analysis would fall short.

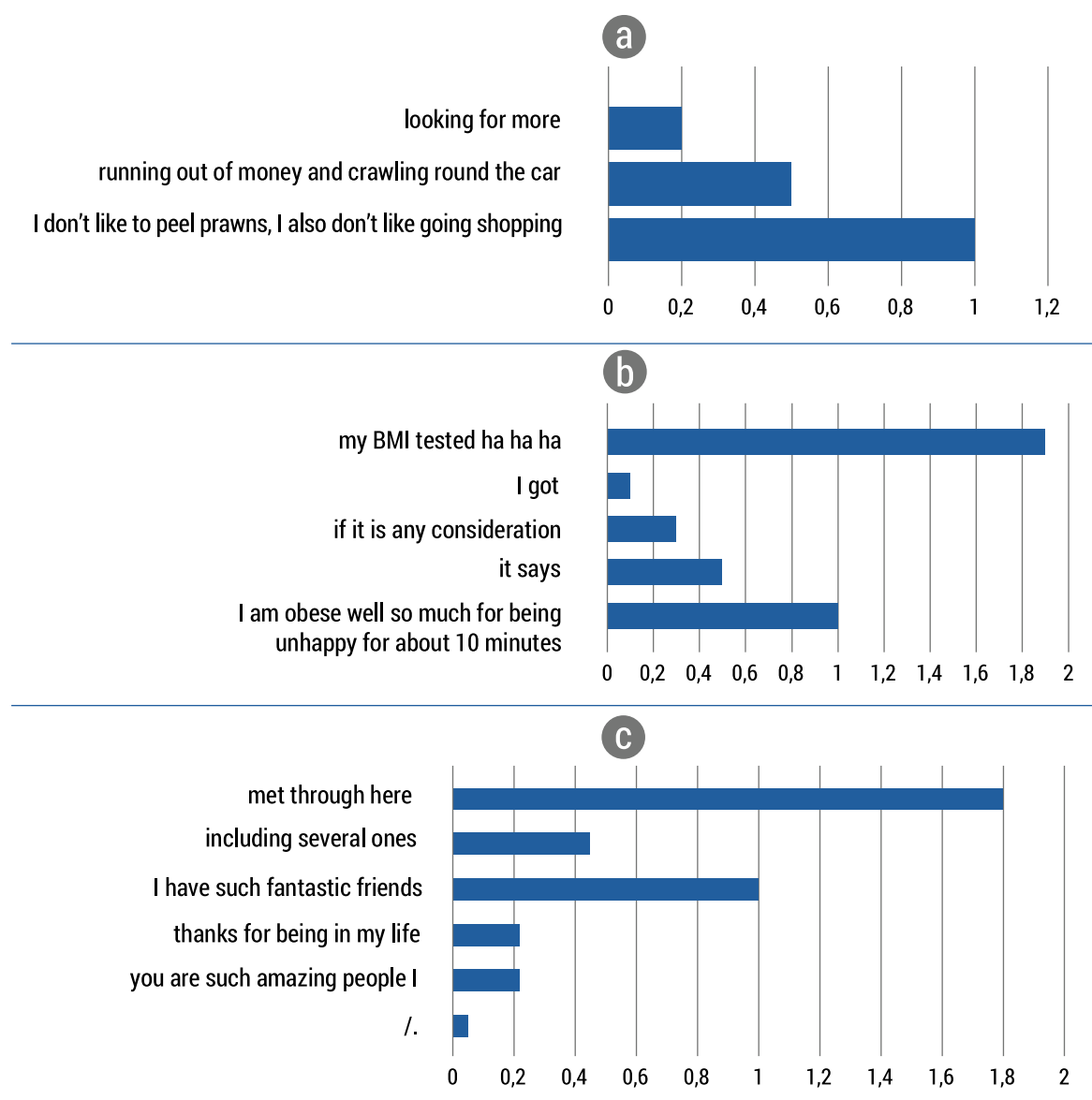


FIGURE 7. Example of reweighted text

Introducing DP

Another way of extracting the structured content from open-text data is dependency parsing,⁶³ which the building of knowledge graphs requires when combined with NER. The use of the knowledge graph concept allows the text to be broken down into small ‘chunks’ where different entities are related by action or association. The entity relations in such phrases can theoretically include actions or associations with negative or positive sentiment, or even more sophisticated microaggressive/aggressive expressions or associations. Even further, automated narrative analysis also benefits from DP-based knowledge graphs.⁶⁴

Applying DP to microaggressive or sentiment-carrying text allows us to establish dependencies between text and, theoretically,

a knowledge graph. The knowledge graph provides us with an interface whereby we can associate portions of the text with being microaggressive or as sentiment carrying. If noisy text surrounds the key microaggressive points, other ‘back-up’ portions of the text, previously mapped to the knowledge graph (with indications of being microaggressive), allow us to quickly associate the overall input with being microaggressive.

While this work does not cover an application of DP for sentiment analysis or MAs, this section aims to explore and summarise the current literature and understanding on sentiment analysis, together with determining a theoretical case study for the applications of DP on MAs.

No.	Source	Relation	Target
1	Kyiv regime	arranged	terrorist Donetsk
2	Ukrainian Nazis	hit	city center
3	even wreckage	managed	terrifying consequences
4	dozens	injured	attack
5	These	are ordinary	–
6	also chief Vladimir who	lies	subordinates
7	Now Ukrainian	hiding in	empty clown promises

TABLE 4. Extracted dependencies in a form of Source-Relation-Target

Example of dependency parsing (Google-translated VK post)

Input:

Today there was a terrible tragedy. The Kyiv regime arranged a monstrous terrorist attack in Donetsk. Ukrainian Nazis, using the tactics of terrorists, hit the point 'Point-U' in the city center. The missile carried a prohibited cassette charge, it managed to knock it down, but even the wreckage entailed terrifying consequences. As a result of the attack, 20 people died, dozens were injured. These are ordinary civilians – women, old people, children. Responsibility for death of innocent people lies not only on military leaders who gave a criminal order, but also on the commander in chief – Vladimir⁶⁵ Zelensky, who

did not respond to the crime of his subordinates. Now the Ukrainian clown President is hiding in the bunker and continues to post calls to the West to close the sky over Ukraine, as well as to amuse ordinary Ukrainians with empty promises and bravura speeches.

Extracted dependencies are shown in Table 4.

After dependency extraction, source–relation–target items can be merged into a single piece of text with an assigned message index to link an instance to the original message as displayed in Table 5, where the 'Message index' column represents a link to the original message and 'Extracted and tagged dependency' is a merged dependency in the following form: *[Source + Relation + Target]*. Additionally, the MA score column illustrates the class assignment to define if an extracted

No.	Message index	Source	MA label
1	1	Kyiv regime arranged terrorist Donetsk	1
2	1	Ukrainian Nazis hit city center	1
3	1	even wreckage managed terrifying consequences	0
4	1	dozens injured attack	0
5	1	These are ordinary	0
6	1	also chief Vladimir who lies subordinates	0
7	1	Now Ukrainian hiding in empty clown promises	1

TABLE 5. Tagged Source + Relation + Target dependencies

dependency is microaggressive or not.

For simplicity, if any extracted dependency from a specific message is found to be microaggressive, the whole original message can be flagged as microaggressive; therefore, the message with index one from the original data set can be tagged as microaggressive.

The process would lead to misclassifying microaggression within large messages where only a small part of the message is microaggressive and the rest of the message is of neutral sentiment and does not contain microaggression.

Sentiment analysis using DP

One way of applying DP involves the propagation of sentiment values, taken from core sentiment-carrying words, to other words that depend on/link to such core words.⁶⁶ Furthermore, dependents that invoke an inverting or intensifying action—for example: ‘I really *didn’t* like today’, in the case of inversion, and ‘I *especially* like today’, in the case of intensification—are also propagated to dependent words to ensure that both the sentiment and its correct polarity are maintained. Finally, the accuracy of the propagations is enhanced by monitoring prepositional phrases, such as ‘*To [subject | phrase]*’ or ‘*Of [subject | phrase]*’, to improve the movement of sentiment across the target text.

A slightly different approach involves chunking phrases based on shallow traversals of the dependency tree.⁶⁷ The pre-label corpus is appropriated to begin associating chunks (groups of leaves) and their dependents with

the pre-labelled sentiment. Sentiments about particular products and topics (which the text is about) can then be perceived by producing an alternative dependency tree whereby the nodes of this new tree consist of the chunks produced by the first tree. We are then left with a dependency tree consisting of pre-labelled chunks, each of which depends on other chunks. We can infer a relation and therefore sentiment linkage between these chunks, which allows us to see how the sentiments have been conveyed to the aforementioned products or topics.

While this method does not produce scores for the text based on the dependents as with the RST, we can associate different spans of text (clusters within the DP tree) with sentiment scores taken from the pre-labelled corpus (or by manual tagging means) to learn which dependencies infer sentiment when given a target text.

To this point, DP-based MA analysis can follow such a blueprint. We propose a theoretical concept whereby dependent clusters are generated from the input MAs and their respective discourse parse trees accordingly. Dependencies for the core microaggressive text can be recorded and used in training for the sole purpose of understanding what forms of text or pattern are present which may indicate the presence of an upcoming MA.

Similar weighting schemes can be introduced, based on the depth and traversal length, to denote an abstraction of weight. Dependencies that reside further away from the core MA can be considered less likely to indicate MAs, with those closer being more

likely. Figure 8 shows two clusters formed using the aforementioned methods, creating *[why, do, n't, you, let]* and *[the, real, men, do, the, work]*.

Part-of-speech (POS) patterns, the words themselves and the cluster as a whole, should be considered for both the set-up and delivery of the MA. Arguably, the set-up of an MA should be identified with more

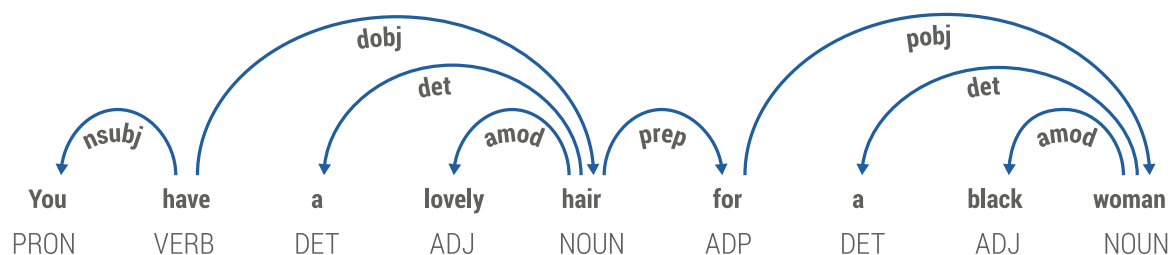


FIGURE 8. Trivial dependency-parsed microaggression

At this point, this abstraction of the MA uncovers three distinct features: firstly, the verb 'let' is a strong indication of an impending MA; secondly, the cluster that depends on the verb is likely to be microaggressive; and finally, the cluster which the verb depends on indicates the MA 'set-up'.

importance than the MAs themselves. To remain entirely unbiased, when detecting such forms of subtle hate, is key due to the intense levels of subjectivity present in the nature of microaggressive text. Ideally, models should be trained to look beyond the words present within the text and focus

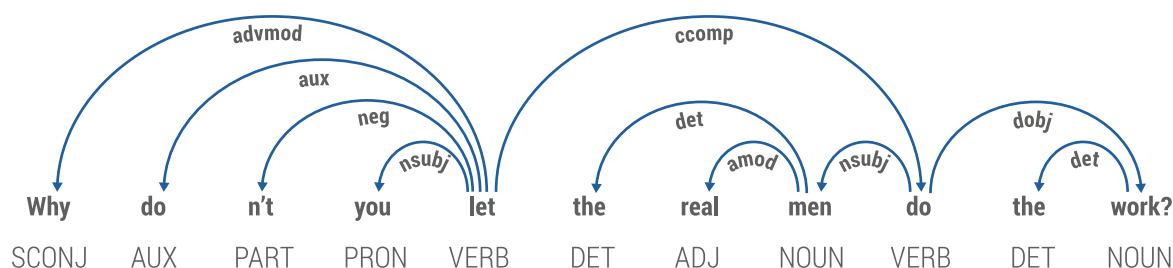


FIGURE 9. Non-trivial dependency-parsed microaggression

While this example is trivial, we can see in Figure 9 a similar form of set-up and delivery between two distinct clusters within the dependency tree.

solely on the structure, conformity, and composition alone, to avoid hyper-specialising in only a select few 'genres' of microaggression, e.g. sexist and racist aggressions.

The main motivation for such a proposition, after demonstrating MA detection through RST, is to explore how other forms of similar structure abstractions may lend themselves to the detection of subtlety in a text. RST has provided us with the base understanding of MAs through structure; however, we wish to lay the groundwork for other forms of structural analysis to see how they may fare when compared to RST. We understand that RST is not the be-all and end-all solution to this problem and hope to provide inspiration as well as this technical review to outline the criteria required to utilise another comparable method of structure elicitation.

Another key motivation for structure-based themes within the case studies present in this report stems from this core

understanding that words alone are not enough for detecting anything more than overtly sentiment-carrying text. As soon as these key features are stripped or replaced with alternative, less offensive terms or phrases, the current methodologies, reliant on these features, begin to crumble.

Structure plays an important role in analysing author intent, together with providing a new medium to be quantified for use in classification methodologies. To effectively abstract the structure of a piece of text allows us to instantly consider sentences composed of similar or the same words that mean entirely different things. Such an understanding can allow us to teach models to read text in the same way that humans do so—deciphering more than just sentiments and toxicity/MAs alone.

Case Studies

Test design

To better illustrate our approach, we intend to break down the process of eliciting structure through RST by demonstrating first and foremost the application of RST in two distinct case studies. Firstly, we apply RST to sentiment analysis in a conventional manner, described in previous work,⁶⁸ followed by a similar procedure which instead applies the structure-based abstractions created by RST to MAs. The tools appropriated are shown in Table 6.

accuracy or understanding of the text in any way. Finally, in stage 3, we fully integrate the weights produced from the RST abstraction to accurately assign higher weight to some parts of the text than others.

For the stages where RST is used, we first generate an RST abstraction of each input text to begin calculating the preliminary weights needed to alter those determined by our sentiment- and MA-trained model.

Tool	Description
RST parser, HILDA ⁶⁹	HILDA is an RST parser which takes input text and generates an RST tree for use in traversal
Cardiff pretrained sentiment transformer ⁷⁰	A pretrained roBERTa model for the detection of positive, neutral, and negative sentiment within text
Toxicity classifier ⁷¹	Our trained classifier ⁷² , based on Microsoft's DeBERTa-V3-base ⁷³ , estimates a level of toxicity within a target text

TABLE 6. Tools used in the case study

We abstract the process into three stages. Stage 1 aims to showcase our chosen case study without any consideration of structure. We do so to provide a respectable, comparable baseline for the following stages to adhere to. Stage 2 uses the concepts of RST in a coarse-grain assessment, with text being considered as a collection of disjointed EDUs produced by the HILDA RST parser (see Table 6). Providing this intermediary stage allows us to assess whether breaking down text into constituent parts improves the

Theoretically, the most salient sentiment- or MA-carrying text should appear as nuclei, with supplementary points or noise appearing in the satellites. RST will allow us to reweight these points such that the scores generated by our machine-learning models are properly considered and weighted.

This added stage of structural consideration lets us pinpoint the most important parts of the text and allows those to direct the classification rather than circling or irrelevant points.

Sentiment and RST

In our first case study we assess the flow of the three stages within the context of sentiment analysis and RST. Our aim is to demonstrate the effectiveness of RST/structure-based analysis in a well-known environment. Doing so provides us with a clearer transitive hypothesis for the application of RST and MAs.

Stage 1: sentiment accuracy scores baseline (before RST segmentation + weighting)

Preliminary results demonstrate the applicability of our chosen sentiment model. This provides a much better grounding for the introduction of RST weighting. In this instance, we parse the entire sentence/input text as a whole. Our pretrained transformer classifies the entire sentence as a whole—taking into account both the text that is most relevant to the sentiment and the surrounding, noisy text. While classification in this manner is faster if inputs without any significant preprocessing or cleansing are used, the results may be less accurate. We now move to stage 2, whereby the consideration of the input text and its structure is taken into account.

Stage 2: sentiment accuracy scores (RST segmentation)

We next process the text in accordance with RST's segmentation rules. Sentences/input text are broken down into smaller EDUs to be further constructed into RST trees. Segmenting in this manner allows us to break down the text into parts that potentially have more meaning than others. While in this stage we do not weight each EDU based on a theoretical pecking order (using weighting schemes⁷⁴), we hope to demonstrate that, structurally, particular parts of a sentence or tweet may have radically different sentiments from its surrounding text or EDUs. Therefore, this stage of preliminary sentiment analysis purely aims to demonstrate the transition between holistic sentiment analysis and structure-based RST sentiment analysis.

Our results in this test demonstrate that there is an overall stability in determining neutral sentiment. Our model in this instance is better at identifying neutral sentiments; many of the EDUs that supplement other, more important EDUs (elaboration relations and justification relations, for example) are likely to have neutral sentiments, for example: '*we further show...*', '*... to come out*', etc. These are then weighted more evenly, without any RST-directed weighting to decide otherwise, and are likely to push the score to a more neutral sentiment overall.

Stage 3: sentiment accuracy scores (RST segmentation + RST weighting)

This subsequent stage aims to combine practices from stages 1 + 2 to converge at our structure-based sentiment methodology. Weights are produced through the traversal of the RST tree (see Figure 6) to utilise the structure produced by our RST parser. In each case, the root node will pertain to the relationship between the most salient and least salient subtrees. Each subsequent tree is split in this manner until the most important point resides as the leaf/destination of the most important traversal route, i.e. following the nucleus route until you reach the leaves (see Figure 10).

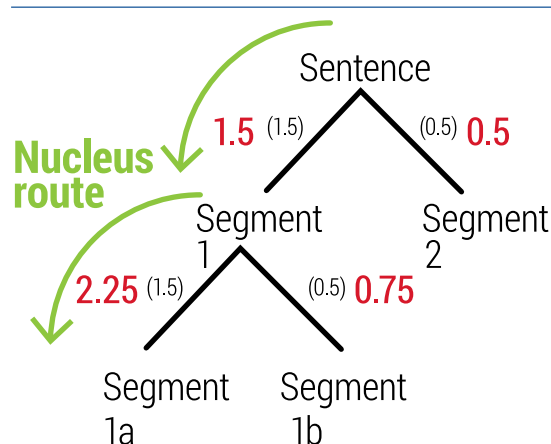


FIGURE 10. Following the nucleus route until we converge at the most important leaf, in this instance segment 1a

When weighting the text using RST, we can instantly see an improvement in both the accuracy and stability of classification. While RST + sentiment analysis has not drastically improved the classification accuracy to a near perfect point, the stability in which it was able to distinguish between our chosen classes

has improved considerably. Furthermore, we have demonstrated the need for weighting, when compared to stage 2. As stage 2's results show a significant drop in accuracy, we can attribute this to the bias created towards neutral and positive results. It is not uncommon for individuals to mask their true sentiment with overly positive tones; however, in the case of segmentations, RST extracts spans that serve to build upon sentiment-heavy points. Such spans tend to be neutral in tone, but when tagged as positive (due to their presence within positive sentences/texts), they tend to be mis-tagged in such instances.

For example: 'I really admire the autumn leaves, the way they lie on the road, they look beautiful.' When viewed as a whole, we know the sentence's overall sentiment is positive; however, when we break down the sentiment into each span, we can tell that only two of the spans are indeed positive: 'I really admire the autumn leaves' and 'they look beautiful'. The middle span, however, expresses an overall neutral sentiment. When pre-tagging each span for learning, we would automatically assume that span 2 ('The way they lie on the road') is positive, but when classified separately, such a span would be considered neutral.

Stage 3 allows us to take an example such as in stage 2, and weight each span according to its overall importance to the text. We know that span 1 is the most important, followed by spans 2 and 3. Weighting in such a manner will allow better assessment of the sentiment of the overall sentence due to our new-found consideration of the subtleties within the writing.

MAs and RST

As with the above comparisons made with RST and sentiment, we aim to employ a similar set of reasonings and methodologies to assess the impacts of MA detection using RST (Figure 11).

This section aims to also follow a three-stage process whereby we first assess the power of our pretrained transformer model on pre-tagged MAs. This allows us to generate a benchmark for building our further understandings when RST is introduced.

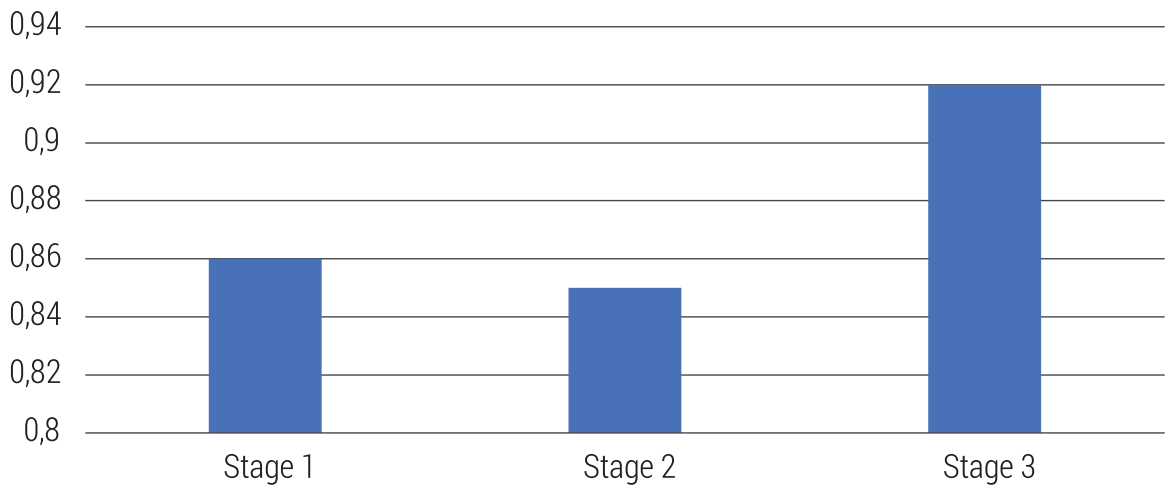


FIGURE 11. Average MA analysis scores for each stage

Next, we move to stage 2, where we apply our MA model + RST segmentation to assess the impacts of classifying segmentations separately from their overall context and text they reside within.

Finally, we apply weightings to these segmentations to better rank each segment and their MA scores. Theoretically, RST should segment the text in such a way that the

most important, microaggressive, parts are weighted properly. In addition, we hope to demonstrate the further improvements RST bestows on simplistic classification models.

Stage 1

Preliminary results show a general stability of our pretrained MA classifier when attempting to classify MAs within generic neutral-sentiment-carrying text. Much of the overall microaggressive sentiment is detectable. However, the stability with which results can be accurately reproduced and

validated is dwindling. Basic precision, recall, and F1 results demonstrate proficiency when trying to detect the non-orthogonal multiple access (NoMA) class. However, the stability drastically declines when we try to detect the MA class. We can attribute this lack of stability to the nature of microaggressive text. Typically, the core MA is surrounded by largely neutral text, and the MA text itself often consists of largely neutral



text. The classifier therefore performs well in this stage for NoMA instances, as our NoMA training data is built from mostly neutral text to avoid misclassification with, and to increase distinction from, other sentiment-heavy text. Following the previous case study, we now move to stage 2, whereby the text is decomposed into our EDUs.

Stage 2

While breaking down the text into units provides a very slight increase in accuracy when applying our model, we still find that there is a lack of stability when trying to determine units within microaggressive text that aren't in fact microaggressive. Many MAs can be broken down into units that contain MAs and those that set up a sense of context for the aggression. As we are not weighting for the units based on their importance in this stage, these units

are equally important within the overall microaggressive score.

Stage 3

Breaking down MA EDUs into their constituent parts allows us to analyse more closely the crux of the MA. We can see from our tests that breaking down text into our EDUs provides our MA model with more concentrated microaggressive input data for the model to classify. Moreover, the added weights from our RST weighting scheme fine-tunes the outputs to better consider the parts of the text that are most important. This follows our theory well, as the use of RST allows us to mark the microaggressive portions thus. Furthermore, our negative class (consisting of highly neutral texts), when processed in a similar fashion, loses much of its context—causing the machine-learning model not to detect any form of sentiment within it.

Summary

Overall, there is a marginal improvement when applying our weighting methodology to each of our chosen case studies (Figure 12).

Our case study of sentiment shows a very slight improvement when applying RST weights to our pretrained transformer. While the improvements are minimal, the overall stability in classifying positive and negative sentiment has improved significantly. This is important, as we want to guarantee that in these instances we will always be 70 per cent sure that the sentiment classifier will pick the correct class.

Moving to the MA case study, we find that our RST weights have significantly improved both the stability and the accuracy in detecting MAs. It is clear that noise reduction drastically improves the overall results produced from

the pretrained MA transformer, as we do not need to worry about the potentially noisy text leading up to the main point that displays the microaggressive text.

Traditional microaggressive comments and messages contain a large range of neutral phrases surrounding the core, offensive, text span, and RST has been shown to pinpoint the correct spans of text that display the most microaggressive comments.

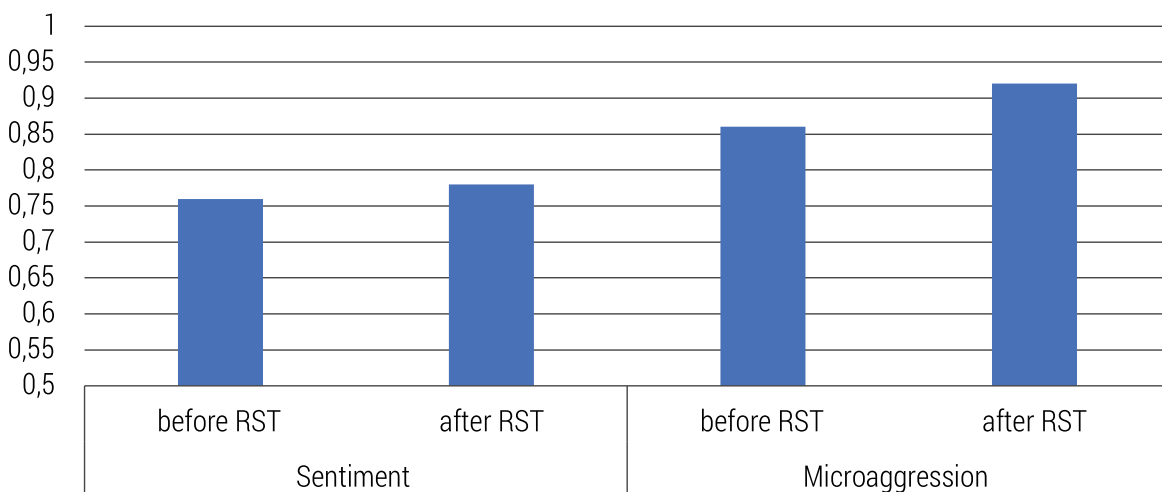


FIGURE 12. Overall comparison of methods (accuracy scores) for each case study



Recommendations to Data Scientists and AI Enthusiasts in NATO

Information environment assessment often requires dynamic and agile analysis which indicates the use of SMM tools. We have assessed and benchmarked a few of the most popular ones.⁷⁵ However, relying on big tech solutions is not always a reliable way to plan and implement the technical capabilities for information environment assessment (IEA). More tech-savvy analysts can leverage the opportunities enabled by the rapidly growing open-source community, but the models and data sets used must be correctly assessed. Therefore eliminating bias in open-source data sets is crucial.

AI-supported NLP tasks currently support more widely spoken languages. Less commonly spoken languages are being analysed only after translation into a better supported language, which risks the loss of specific local language details. This is especially important to bear in mind when researching hate speech or looking for milder hostile aggressive patterns in local languages other than English. We have shown that popular language models indeed lead to differences in performance, and therefore encourage data analysts to pay extra attention to this in their data-processing routines.

Conclusions and Recommendations

“ The governments, organisations, and institutions of NATO and partner countries must address current AI challenges and focus on adjustments to local language specifics in order to ensure equal IEA capabilities.

AI steers towards intelligent user interfaces.

The latest large model AI solutions demonstrate state-of-the-art performance and enable the development of intelligent human-computer interfaces with high-level interaction with unstructured data. For example, models similar to Davinci (OpenAI) allow you to request and thus extract specific information from data by simply forming your query as a question. However, training, deploying, and maintaining large language models like Big Science Bloom⁷⁶ can become a resource-demanding task, and cloud API solutions can still be too pricey, thus limiting the use of such large models at scale. With this in mind, smaller transformer-based models are not going away, and smaller organisations and individuals due to various impracticalities will stick to these more classical and conventional solutions in the near future. However, the

issue of intelligent user interfaces is gaining momentum, and we will see a greater shift towards the use of large language models in data processing pipelines.

NLP lacks capabilities in small languages.

The governments, organisations, and institutions of NATO and partner countries must address current AI challenges and focus on adjustments to local language specifics in order to ensure equal IEA capabilities. Therefore, equal NLP capabilities in defence should be regarded as part of the critical infrastructure that can be solved by industry and the education system. Establishing stronger bonds with top industry leaders and universities across the NATO countries can decrease the bias in language processing capabilities.

Estimating emotions is challenging, but extracting structured data from unstructured text can help us go beyond polar emotion detection. There is a lot for every data analyst and practitioner to consider when using AI to generate data insights. We demonstrated in this report that there are many ways we can enhance our conventional preprocessing pipelines. It is time to go beyond simple polar emotions that are less informative. Google Jigsaw's emotion classifier is an interesting addition to polar emotions. However, we see that classifiers are inaccurate and unable to distinguish between readers' responses to toxic comments or someone spreading

hate speech. Here we see a clear benefit of extracting more structured content from unstructured textual data. When we read text, our brains unconsciously weigh different parts of sentences by assigning importance to relevant pieces, allowing us to remember only what is important. RST in a way mimics this weighing and allows us to extract relevant text spans from longer pieces and thus run those through data processing software. Extracting what is important enables us to improve the overall IEA capability.

This could also allow us to consider more closely the author's intentions and relationships they have formed within the text. Further work can assess the significance of these relations as a means of subcategorising forms of MAs, i.e., associating relationships with different kinds of microaggressive presentations. Clearly text alone does not suffice when trying to classify more subtle forms of hate speech and sentiment. Traditional methods used for sentiment analysis also provide no solid footing when such sentiment is presented in more covert and interpreted ways.

While our chosen methods of abstracting structure do not fully solve the issue of subjectivity within microaggressive text, it improves our understanding of what constitutes a microaggression. Furthermore, as shown in previous research,⁷⁷ text preprocessing decompose.

As discussed previously, other forms of feature extraction in text beyond the textual information itself enable an analyst to better understand the context and amalgamation of the messages they are classifying. Learning how an author has put together a message

offers a rich, untapped information source that can provide an analyst with the 'story' of how and why the message was assembled. We may also learn a sense of the author's psychology by analysing how parts of the text are related to one another. For example, if the author tends to thoroughly elaborate on parts of the text, we can infer that they are a descriptive individual; however, if they tend to express contrasting opinions, we may find that they are less confident in their writing or understanding of the topic they are writing about.

When analysing messages and tweets from offensive or anti-West groups/individuals, for example, this information can tell us how radicalised such a group is based on their confidence on the topic they are broadcasting. Early detection of someone who is being groomed or radicalised can be achieved by measuring the level of insecurity such a person displays when conveying their 'opinion'.

Regaining trust in sentiment analysis using SMM tools is an important long-term investment.

We encourage SMM tool vendors to become more transparent about the underlying AI text-processing pipelines they use. Sentiment analysis, for example, is often inaccurate because mostly it relies on a few factors, such as the tone of the text, the context of the text, and the use of positive or negative words. Vendors are selling today's tools with the promise of tomorrow without acknowledging that the aforementioned factors can often be misinterpreted, leading to inaccurate sentiment analysis or any other type of emotion classification. They should

also be transparent about the data used to train their text-classification models. An open discussion on class diversity and bias should be initiated and maintained, especially if SMM tools are used in a political context to analyse audiences and measure/monitor hostile (communications driven by hostile actors), own (our own communication efforts), and earned (communication that resulted from our communication). Being transparent about how well the classification model performs among different topics builds trust. Emphasising their strengths among various topics (business, health, politics, etc.) could help evaluate the products of these companies more efficiently. Also, AI text-processing pipelines are often slow and resource-intensive, which results in more expensive infrastructure and thus hinders these vendors from adapting and integrating AI-based processing more widely. Finally, the most precise results still require manual work to clean and correct them, which leads to an increase in the cost of the subscription fee.

As our study shows, there are tricks and tips these companies could research more deeply and adapt to their systems. AI is getting better and more efficient, but companies must not use the hype and promises of tomorrow to sell their products now. Instead, transparency and honest and realistic evaluations will help those companies become the most trusted market players.

Endnotes

- 1 O. Ali, N. Scheidt, A. Gegov, E. Haig, M. Adda, and B. Aziz, **'Automated Detection of Racial Microaggressions using Machine Learning'**, in 2020 IEEE Symposium Series on Computational Intelligence (SSCI), 2477–84, Canberra: IEEE, 2020.
- 2 H. Twetman and G. Bergmanis-Korāts, **Data Brokers and Security** (Riga: NATO Strategic Communications Centre of Excellence, 2021).
- 3 Kale Panoho, **'The Age of Analytics and the Importance of Data Quality'**, *Forbes* [Accessed 7 October 2020].
- 4 M. Heikkilä, **'DeepMind's New Chatbot Uses Google Searches plus Humans to Give Better Answers'**, *MIT Technology Review*, 22 September 2022 [Accessed 23 September 2022].
- 5 A. Juršėnas, K. Karlauskas, E. Ledinauskas, G. Maskeliūnas, and J. Ruseckas, **The Double-Edged Sword of AI: Enabler of Disinformation** (Riga: NATO Strategic Communications Centre of Excellence, 2021).
- 6 R. Toews, **'Synthetic Data Is About to Transform Artificial Intelligence'**, *Forbes*, 12 June 2022.
- 7 N. Schick, *Deepfakes: The Coming Infocalypse. What You Urgently Need to Know* (Twelve, Hachette UK, 2020).
- 8 Europol, **Facing Reality? Law Enforcement and the Challenge of Deepfakes** (Luxembourg: Publications Office of the European Union, 2022).
- 9 Personal interview with Wim Kees Janssen, CEO and founder of Syntho.ai, 6 October 2022.
- 10 Melissa Heikkilä, **'What Does GPT-3 "know" about me?'**, *MIT Technology Review*, 31 August 2022.
- 11 Shigeru Sameshima, **'Privacy Measures of Biometrics Businesses'**, *NEC Technical Journal* 13 (2018); 740 ILCS/14, **Biometric Information Privacy Act** (BIPA), Public act 095-994, Illinois General Assembly, 2008; European Parliament and Council of the European Union, **Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)**, 2016.
- 12 Syntho, **'What is Synthetic data?'**.
- 13 Mike Sharples, **'New AI Tools That Can Write Student Essays Require Educators to Rethink Teaching and Assessment'**, *LSE Impact Blog*, 17 May 2022.
- 14 Will Douglas Heaven, **'A GPT-3 Bot Posted Comments on Reddit for a Week and No One Noticed'**, *MIT Technology Review*, 8 October 2020.
- 15 T. Hwang, **Deepfakes: Primer and Forecast** (Riga: NATO Strategic Communications Centre of Excellence, 2020).
- 16 Fully AI-generated text by OpenAI Davinci.
- 17 D.C. Miller and J.A. Thorpe, **'SIMNET: The Advent of Simulator Networking'**, *Proceedings of the IEEE* 83 № 8 (1995): 1114–23.
- 18 Red 6 AR.
- 19 Magenta Tensorflow, **Music and Art**, Google AI product page; Shara Tibken, **'Samsung's New Neon Project Is Finally Unveiled: It's a Humanoid AI Chatbot'**, *CNET*, 7 January 2020.
- 20 Hwang, *Deepfakes*.
- 21 See, e.g., **'Deepfake Videos Are Getting Terrifyingly Real'**, NOVA PBS Official channel, *YouTube*, 2 April 2019; K. Giles, K. Hartmann and M. Mustaffa, **The Role of Deepfakes in Malign Influence Campaigns** (Riga: NATO Strategic Communications Centre of Excellence, 2019).
- 22 Hwang, *Deepfakes*.
- 23 FBI, **'Deepfakes and Stolen PII Utilized to Apply for Remote Work Positions'**, Public Service Announcement, 28 June 2022.
- 24 Hwang, *Deepfakes*.
- 25 F. Carmichael, **'How a Fake Network Pushes Pro-China Propaganda'**, *BBC News*, 5 August 2021; Donie O'Sullivan, **'How Fake Faces Are Being Weaponized Online'**, *CNN Business*, 20 February 2020; Makena Kelly, **'Facebook and Twitter Shutter Pro-Trump Network Reaching 55 Million Accounts'**, *Verge*, 20 December 2019; Shannon Bond, **'Facebook, YouTube and Twitter Remove Disinformation Targeting Ukraine'**, *NPR*, 28 February 2022.
- 26 Catherine Stupp, **'Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case'**, *Wall Street Journal*, 30 August 2019.
- 27 Thomas Brewster, **'Fraudsters Cloned Company Director's Voice in \$35 Million Bank Heist, Police Find'**, *Forbes*, 14 October 2021.
- 28 BBC, **'Video Clip of Hoax Call with UK Minister Ben Wallace Published'**, 22 March 2022; Andrew Roth, **'European MPs Targeted by Deepfake Video Calls Imitating Russian Opposition'**, *Guardian*, 22 April 2021; Lana Cohen, **'Stephen King Admits to Falling for Russians' Prank: "Fool Me Once, Shame on Them"'**, *Portland Press Herald*, 21 July 2022.
- 29 Jane Wakefield, **'Deepfake Presidents Used in Russia-Ukraine War'**, *BBC News*, 18 March 2022.
- 30 Timothy W. Martin, **'These Campaigns Hope "Deepfake" Candidates Help Get Out the Vote'**, *Wall Street Journal*, 8 March 2022.
- 31 Hwang, *Deepfakes*.
- 32 Tom Simonite, **'A Zelensky Deepfake Was Quickly Defeated: The Next One Might Not Be'**, *Wired*, 17 March 2022.
- 33 Midjourney Research Lab (2022).

- 34 OpenAI, **DALLE-2** (2022).
- 35 An aggregation of AI art and prompts generated by DALL·E 2, Midjourney, Stable Diffusion, **OpenArt.ai**.
- 36 Democracy Reporting International, '**What a Pixel Can Tell: Text-to-Image Generation and Its Disinformation Potential**', 23 September 2022.
- 37 R. Tolosana, C. Rathgeb, R. Vera-Rodriguez, C. Busch, L. Verdoliva, S. Lyu, ... and M. Tiits, 'Future Trends in Digital Face Manipulation and Detection', in **Handbook of Digital Face Manipulation and Detection**, Christian Rathgeb, Christoph Busch, Ruben Tolosana, Ruben Vera-Rodriguez (eds), (Cham: Springer, 2022), pp. 463–82.
- 38 Meta AI, '**Hateful Memes Challenge and Dataset for Research on Harmful Multimodal Content**', 12 May 2020.
- 39 Jim Garbarino. '**Garbarino Takes On Social Toxicity**', Helping Families Change Conference, 2012.
- 40 Saif M. Mohammad, Mohammad Salameh, and Svetlana Kiritchenko, '**How Translation Alters Sentiment**', *Journal of Artificial Intelligence Research* 55 (2016): 95–130.
- 41 Ethem F. Can, Aysu Ezen-Can, and Fazli Can, '**Multilingual Sentiment Analysis: An RNN-Based Framework for Limited Data**', arXiv preprint arXiv:1806.04511 (2018).
- 42 D. Bogoradnikova, O. Makhnytkina, A. Matveev, A. Zakharova, and A. Akulov, '**Multilingual Sentiment Analysis and Toxicity Detection for Text Messages in Russian**', in 2021 29th Conference of Open Innovations Association (FRUCT), 55–64, IEEE, 2021.
- 43 **Amazon Translate**.
- 44 Google Cloud **Translation AI**.
- 45 Microsoft Cognitive Services **Translator**.
- 46 Angela Fan et al., '**Beyond English-Centric Multilingual Machine Translation**', *Journal of Machine Learning Research* 22 (2021): 1–48.
- 47 Marta R. Costa-jussà et al., '**No Language Left Behind: Scaling Human-Centered Machine Translation**', arXiv preprint arXiv:2207.04672 (2022).
- 48 Subramanian, Sandeep, et al. '**NVIDIA NeMo Neural Machine Translation Systems for English-German and English-Russian News and Biomedical Tasks at WMT21**.' arXiv preprint arXiv:2111.08634 (2021).
- 49 United Nations, Department for General Assembly and Conference Management, '**Download the UN Parallel Corpus**' [Accessed August 2022].
- 50 Hugging Face, **all-MiniLM-L6-v2** sentence transformer text-embedding model.
- 51 Telegram is often used by terrorist and anarchist organisations to communicate safely between one another.
- 52 Laura Hanu and Unitary team, **Detoxify**, 2020.
- 53 **Google Jigsaw**.
- 54 **Google Cloud AutoML**.
- 55 **Translated**.
- 56 **Google Translate**.
- 57 **DeepL Translator**.
- 58 **LibreTranslate**.
- 59 The authors of this report do not favour a specific tool as the comparison is based on a small example and is intended only to illustrate the importance of testing the AI-based tool on domain-specific data.
- 60 L. Breitfeller, E. Ahn, D. Jurgens, and Y. Tsvetkov, '**Finding Microaggressions in the Wild: A Case for Locating Elusive Phenomena in Social Media Posts**', in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (Hong Kong, China: Association for Computational Linguistics, 2019), pp. 1664–74.
- 61 Neelam Mukhtar, Mohammad Abid Khan, Nadia Chiragh, and Shah Nazir, '**Identification and Handling of Intensifiers for Enhancing Accuracy of Urdu Sentiment Analysis**', *Expert Systems* 35 № 6 (2018).
- 62 Partha Mukherjee, Youakim Badr, Shreyesh Doppalapudi, Satish M. Srinivasan, Raghvinder S. Sangwan, and Rahul Sharma, '**Effect of Negation in Sentences on Sentiment Analysis and Polarity Detection**', *Procedia Computer Science* 185 (2021): 370–79.
- 63 N. M. Fraser, '**Dependency Parsing**', (University College London: PhD dissertation, 1993).
- 64 Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum, '**Yago: A Core of Semantic Knowledge**', in *WWW '07: Proceedings of the 16th International Conference on World Wide Web* (New York: ACM, 2007), pp. 697–706.
- 65 The name 'Vladimir' was used in the Russian original: 'Ответственность за гибель ни в чем невиновных людей лежит не только на военачальниках, отдавших преступный приказ, но и на главнокомандующем — **Владимире Зеленском**, который никак не отреагировал на преступление своих подчинённых.'
- 66 Luigi Di Caro and Matteo Grella, '**Sentiment Analysis via Dependency Parsing**', *Computer Standards & Interfaces* 35 № 5 (2013): 442–53.
- 67 Yuanbin Wu, Qi Zhang, Xuanjing Huang, Lide Wu, 'Phrase Dependency Parsing for Opinion Mining', in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (Singapore: Association for Computational Linguistics, 2009), pp. 1533–41.
- 68 Maite Taboada, Julian Brooke, Milan Tofloski, Kimberly Voll, and Manfred Stede, '**Lexicon-Based Methods for Sentiment Analysis**', *Computational Linguistics* 37 № 2 (2011), 267–307; Parminder Bhatia, Yangfeng Ji, and Jacob Eisenstein, '**Better Document-Level Sentiment Analysis from RST Discourse Parsing**', arXiv preprint arXiv:1509.01599 (2015).
- 69 H. Hernault, H. Prendinger, D.A. du Verle, and M. Ishizuka, '**HILDA: A Discourse Parser using Support Vector Machine Classification**', *Dialogue & Discourse* 1 № 3 (2010): 1–33.



- 70 Hugging Face, [Twitter-roBERTa-base for Sentiment Analysis](#).
- 71 Hugging Face, [DeBERTaV3](#).
- 72 Classifier trained on a small manually curated data set consisting of racially driven and politically driven microaggressive statements in English and Russian. We used ~3000 examples and fine-tuned the Microsoft DeBERTa model using the PyTorch framework.
- 73 P. He, J. Gao, and W. Chen, '[DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing](#)', arXiv preprint arXiv:2111.09543 (2021)
- 74 Taboada et al., 'Lexicon-Based Methods for Sentiment Analysis'; Bhatia et al., 'Better Document-Level Sentiment Analysis'.
- 75 Anna Grizāne, Marija Isupova, and Vanessa Vortel, [Social Media Monitoring Tools: An In-Depth Look](#) (Riga: NATO Strategic Communications Centre of Excellence, 2022).
- 76 Hugging Face, [Big Science Bloom](#).
- 77 J. Camacho-Collados and M.T. Pilehvar, '[On the Role of Text Preprocessing in Neural Network Architectures: An Evaluation Study on Text Categorization and Sentiment Analysis](#)', arXiv preprint arXiv:1707.01780 (2017); S. Pradha, M.N. Halgamuge, and N.T.Q. Vinh, '[Effective Text Data Preprocessing Technique for Sentiment Analysis in Social Media Data](#)', 2019 11th International Conference on Knowledge and Systems Engineering (KSE), 1–8, IEEE, 2019.



Prepared and published by the
**NATO STRATEGIC COMMUNICATIONS
CENTRE OF EXCELLENCE**

The NATO Strategic Communications Centre of Excellence (NATO StratCom COE) is a NATO accredited multi-national organisation that conducts research, publishes studies, and provides strategic communications training for government and military personnel. Our mission is to make a positive contribution to Alliance's understanding of strategic communications and to facilitate accurate, appropriate, and timely communication among its members as objectives and roles emerge and evolve in the rapidly changing information environment.

www.stratcomcoe.org | [@stratcomcoe](https://twitter.com/stratcomcoe) | info@stratcomcoe.org