A Review of Methods for Analyzing Online Media Texts Using Large Language Models

Pregled metoda za analizu *online* medijskih tekstova primjenom velikih jezičnih modela

Ana Meštrović
University of Rijeka,
Faculty of Informatics and Digital Technologies
R. Matejčić 2, 51000 Rijeka
amestrovic@uniri.hr

Abstract. *This paper explores the potential applications of generative artificial intelligence in tasks related to media text analysis. It highlights the capabilities of large language models, which, with their advanced pre-training and context awareness, support numerous NLP tasks essential for media text analysis, such as keyword extraction, topic modeling, named entity recognition, sentiment analysis, and disinformation detection. The study also delves into techniques such as prompt engineering, retrieval-augmented generation, and fine-tuning, which further enhance the performance of large language models. In addition, this paper provides an overview of a set of NLP-based methods for media text analysis, enhanced with generative AI techniques. By integrating these advanced technologies, it is possible to develop a powerful system for the comprehensive media texts analysis.*

Keywords. Media Texts Analysis, Natual Language Processing, Generative Artificial Intelligence

Sažetak. *Ovaj rad istražuje potencijalne primjene generativne umjetne inteligencije u zadacima vezanim uz analizu medijskih tekstova. Ističe mogućnosti velikih jezičnih modela koji, zahvaljujući svom naprednom pred-treniranju i znanju o kontekstu, podržavaju brojne NLP zadatke bitne za analizu medijskih tekstova, kao što su ekstrakcija ključnih riječi, modeliranje tema, prepoznavanje imenovanih entiteta, analiza sentimenta i otkrivanje dezinformacija. Studija se također bavi tehnikama kao što su inženjering upita, RAG tehnike i fino podešavanje, koje dodatno poboljšavaju performanse velikih jezičnih modela. Pored toga, rad pruža pregled skupa metoda za analizu medijskih tekstova temeljenih na NLP-u, poboljšanih generativnim AI tehnikama. Integracijom naprednih tehnologija moguće je razviti moćan sustav za sveobuhvatnu analizu medijskih tekstova.*

Ključne riječi. Analiza medijskih tekstova, obrada prirodnog jezika, generativna umjetna inteligencija

1 Introduction

Online media text analysis is important aspect of media monitoring that involves the systematic collection, analysis, and interpretation of media content from online platforms. This process plays a crucial role in understanding public discourse, tracking the dissemination of information, and identifying emerging trends. Media text analysis allows stakeholders to extract relevant insights from the vast amounts of texts published on online platforms daily. The ability to efficiently and accurately analyze media texts has become increasingly important in today's digital age, where information is abundant and constantly evolving. The importance of media text analysis and media

1 Uvod

Analiza medijskih tekstova važan je aspekt praćenja medija koji uključuje sustavno prikupljanje, analizu i tumačenje medijskog sadržaja s online platformi. Ovaj proces ima važnu ulogu u razumijevanju javnog diskursa, praćenju širenja informacija i prepoznavanju novih trendova. Analiza medijskih tekstova omogućava dionicima da izvuku relevantne uvide iz velike količine tekstova objavljenih na online platformama svakodnevno. Sposobnost učinkovitog analiziranja medijskih tekstova postaje sve važnija u današnje digitalno doba, gdje su informacije obilne i neprestano se mijenjaju. Važnost analize medijskih tekstova i

monitoring becomes even more pronounced in the context of crises such as war, global warming, pandemics, etc. In such scenarios, timely and accurate information is critical for decision-making, public safety, and strategic communication. Media texts analyis during crises helps authorities and organizations track the spread of information, identify misinformation or disinformation, and gauge public sentiment. For instance, during the COVID-19 pandemic, monitoring media texts enabled health authorities to understand public concerns, combat misinformation, and effectively disseminate health guidelines (Beliga et al., 2022.). Similarly, in times of conflict, media analysis helps in understanding the narrative, tracking propaganda, and assessing the impact of information campaigns on public opinion (Sufi, 2023.).

Natural Language Processing (NLP) has revolutionized the way media texts are analyzed, offering advanced techniques to process and understand large volumes of text efficiently. Analyzing media content provides insights into relevant topics and entities (such as individuals, institutions, locations, and events) present in mass media, allowing users to quickly and accurately identify information pertinent to their interests. Methods such as keyword extraction, topic modeling, named entity recognition (NER), sentiment analysis, text classification, and disinformation detection are employed. These techniques enable information extraction and the prediction of information spread, all driven by deep learning and large language models.

Large Language Models (LLMs), such as BERT (Devlin et al., 2018) and GPT (Radford et al., 2019, Achiam et al., 2023), and their successors, have demonstrated remarkable capabilities in various NLP tasks, including sentiment analysis, NER, and topic modeling. These models leverage extensive pre-training on diverse text corpora, enabling them to capture intricate linguistic patterns and contextual nuances. By applying LLMs to media text analysis, researchers and practitioners can achieve more accurate and insightful interpretations of media content, ultimately enhancing the effectiveness of media monitoring. This review aims to explore the various methods for analyzing media texts using LLMs, highlighting their potential and application in improving media texts analysis tasks.

Many research studies have already analyzed the potential of NLP in media texts anaylsis and media monitoring task (Farzindar et al., 2015; Germann et al., 2018). However, there is still a lack of research papers that study specific NLP tasks related to media texts analysis in the context of generative artificial intelligence (GenAI). Thus, the main goal of this study is to provide an overview of the set of GenAI-based methods that can be used for a comprehensive media texts analysis. Specifically, the main aim of this study is to show how LLMs can improve tasks of

praćenja medija postaje još izraženija u kontekstu kriza poput ratova, globalnog zagrijavanja, pandemija, itd. U takvim situacijama, pravovremene i točne informacije su ključne za donošenje odluka, javnu sigurnost i stratešku komunikaciju. Analza medijskih tekstova tijekom kriza pomaže vlastima i organizacijama da prate širenje informacija, identificiraju dezinformacije te procijene javno mnijenje. Npr., tijekom pandemije COVID-19, praćenje medijskih tekstova omogućilo je odgovornim osobama da razumiju zabrinutost javnosti, suzbiju dezinformacije i učinkovito prenesu zdravstvene smjernice (Beliga i sur., 2022.). Slično, u razdoblju sukoba, analiza medija pomaže u razumijevanju narativa, praćenju propagande i procjeni utjecaja informacijskih kampanja na javno mnijenje (Sufi, 2023.).

Obrada prirodnog jezika (engl. Natural Language Processing, NLP) unaprijedila je način na koji se analiziraju medijski tekstovi, nudeći napredne tehnike za učinkovitu obradu i razumijevanje velike količine tekstova. Analiza medijskog sadržaja pruža uvid u relevantne teme i entitete (kao što su pojedinci, institucije, lokacije i događaji) prisutne u masovnim medijima, omogućujući korisnicima brzo i točno prepoznavanje važnih informacija. Primjenjuju se metode poput ekstrakcije ključnih riječi, modeliranja tema, prepoznavanja imenovanih entiteta, analize sentimenta, klasifikacije teksta i detekcija dezinformacija. Ove tehnike omogućuju ekstrakciju informacija i predviđanje širenja informacija, a sve to dodatno mogu unaprijediti duboko učenje i veliki jezični modeli.

Veliki jezični modeli (engl. Large Language Models, LLMs), poput BERT-a (Devlin et al., 2018) i GPT-a (Radford et al., 2019, Achiam et al., 2023), te njihovih nasljednika, pokazali su izvanredne sposobnosti u raznim NLP zadacima, uključujući analizu sentimenta, prepoznavanje imenovanih entiteta i modeliranje tema. Ovi modeli koriste opsežan pred-trening na raznovrsnim tekstualnim korpusima, što im omogućuje obuhvaćanje složenih jezičnih obrazaca i kontekstualnih nijansi. Primjenom LLM-ova u analizi medijskih tekstova, istraživači mogu postići točnije i dublje interpretacije medijskog sadržaja, čime se u poboljšava učinkovitost praćenja medija. Vilj rada je istražiti različite metode za analizu medijskih tekstova korištenjem LLM-ova, ističući njihov potencijal za analizu medijskih tekstova.

Mnoge istraživačke studije već su analizirale potencijal NLP-a u zadacima analize medijskih tekstova i praćenja medija općenito (Farzindar et al., 2015; Germann et al., 2018). Međutim, još uvijek nedostaje radova koji proučavaju specifične NLP zadatke vezane uz analizu medijskih tekstova u kontekstu generativne umjetne inteligencije. Stoga je glavni cilj ove studije pružiti pregled skupa metoda temeljenih na GenAI koje se mogu koristiti

media texts analysis. This paper discusses the potential of transformer-based models, such as BERT and GPT, for media texts analysis. Furthermore, the study analyzes techniques that can improve these models, such as prompt engineering, retrieval-augmented generation (RAG), and fine-tuning.

The deep understanding of language allows LLMs to perform more accurately and efficiently across various media texts anaylsis tasks. For instance, in keyword extraction and topic modeling, LLMs can identify and cluster relevant terms with greater precision, providing a more comprehensive understanding of the topics and trends present in media content. NER and sentiment classification also benefit from LLMs' ability to understand context and disambiguate entities and sentiments within diverse and unstructured media texts. The incorporation of retrieval-augmented generation further enhances the capabilities of LLMs for media texts analysis. RAG combines the strengths of retrieval-based and generation-based models, allowing the system to access external knowledge sources and generate contextually accurate and up-to-date analyses. This approach is particularly beneficial for tasks like disinformation detection, where accessing current and domain-specific information is crucial for identifying false or misleading content. By retrieving relevant information from external databases and knowledge graphs, RAG can provide enriched and context-aware insights that classical NLP methods might miss. Fine-tuning LLMs on specific datasets developed for media texts analysis can significantly improve their performance by adapting the models to the particularities of the media content being analyzed. Fine-tuning allows the models to learn the specific linguistic features, terminologies, and contextual nuances relevant to media texts, thereby increasing the accuracy and relevance of the analyses. Techniques such as prompt engineering can also optimize the interaction with LLMs, ensuring that the models generate more precise and relevant outputs tailored to the specific needs of media texts analysis.

The rest of the paper is organized as follows. The next section covers the motivation and background. The third section details methods based on large language models that can be utilized for efficient media text analysis. The final section presents the discussion and concluding remarks.

2 Motivation and Background

The methods and techniques integrated into media texts analysis and media monitoring systems should enable the automatic extraction of relevant information from diverse media sources, offering a comprehensive overview of content, relationships, public opinion or sentiment, and trends related to the topics and entities of interest. Developing methods to extract relevant information from a large volume of

za sveobuhvatnu analizu medijskih tekstova. Konkretno, glavni cilj ove studije je pokazati kako LLM-ovi mogu unaprijediti analizu medijskih tekstova. Ovaj rad raspravlja o mogućnostima modela temeljenih na transformerima, kao što su BERT i GPT za analizu medijskih tekstova. Nadalje, studija analizira tehnike koje mogu poboljšati te modele, kao što su inženjering upita, generiranje s poboljšanjem pretrage (engl. Retrieval-Augmented Generation RAG) i fino podešavanje.

Duboko razumijevanje jezika omogućuje LLM-ovima da preciznije i učinkovitije obavljaju različite zadatke analize medijskih tekstova. Na primjer, u ekstrakciji ključnih riječi i modeliranju tema, LLM-ovi mogu identificirati i grupirati relevantne pojmove s većom preciznošću, pružajući sveobuhvatnije razumijevanje tema i trendova prisutnih u medijskom sadržaju. Prepoznavanje imenovanih entiteta i klasifikacija sentimenta također imaju koristi od sposobnosti LLM-ova da razumiju kontekst i razjasne entitete i sentimente unutar raznolikih i nestrukturiranih medijskih tekstova.

RAG dodatno poboljšava sposobnosti LLM-ova u analizi medijskih tekstova. RAG kombinira prednosti modela temeljenih na pretraživanju i modela temeljenih na generiranju, omogućujući sustavu pristup vanjskim izvorima znanja i generiranje kontekstualno točnih i ažuriranih analiza. Ovaj pristup posebno je koristan za zadatke poput otkrivanja dezinformacija, gdje je pristup trenutnim i specifičnim informacijama ključan za prepoznavanje lažnog sadržaja. Prikupljanjem relevantnih informacija iz vanjskih baza podataka i grafova znanja, RAG može pružiti obogaćene i kontekstualno svjesne uvide koje klasične NLP metode možda propuštaju.

Fino podešavanje LLM-ova na skupovima podataka (korpusima) generiranim specifično za zadatke analize medijskih tekstova može poboljšati njihovu izvedbu prilagođavanjem modela specifičnostima analiziranog medijskog sadržaja. Fino podešavanje omogućuje modelima da nauče specifične jezične značajke, terminologiju i kontekstualne nijanse relevantne za medijske tekstove, čime se povećava točnost analiza. Tehnike poput inženjeringa upita također mogu optimizirati interakciju s LLM-ovima, osiguravajući da modeli generiraju preciznije rezultate prilagođene specifičnim potrebama za analizu medijskih tekstova.

Ostatak rada je organiziran na sljedeći način. Iduća sekcija pokriva motivaciju i pozadinu istraživanja. Treće dio detaljno opisuje metode temeljene na LLM-ovima koje se mogu koristiti za učinkovitu analizu medijskih tekstova. Završni dio predstavlja raspravu i zaključne primjedbe.

unstructured data from multiple sources in various formats or free text is a challenging task. Additional challenges in exploring media data include constantly changing trends, the vast number of different topics and entities discussed, sensationalism, biased texts, misinformation, and more.

Investigation into neural network architectures has resulted in the emergence of a wide range of neural text representation models (Babić et al., 2020). Particularly, transformer-based models like BERT or GPT have brought about significant transformations in the realm of natural language processing. These models, which are pre-trained on extensive text datasets and recognized as large language models, capture linguistic structures and semantic relationships, enabling them to surpass conventional language models in various NLP assignments. The influence of LLMs goes beyond enhanced accuracy; they facilitate the creation of more sophisticated and user-friendly language-oriented applications. LLMs employ deep learning architectures with billions of parameters, enabling them to comprehend intricate patterns and contextual subtleties in text. Consequently, LLMs have demonstrated improved performance in tasks such as text categorization (Sun et al., 2019; Balkus and Yan, 2022), sentiment analysis (Babić et al., 2021; Beliga et al., 2021), identification of paraphrases (Vrbanec & Meštrović, 2023), machine translation (Zhu et al., 2020; Yang et al., 2020), question-answering (Wang et al., 2019), prediction of information dissemination (Meštrović et al., 2022), and so forth.

In the realm of media text analysis and media monitoring tasks, numerous studies have applied methods and techniques based on deep learning. For example, Bhoi et al. (2020) propose a deep learning-based social media text analysis framework for disaster resource management. There have been research directions aimed at enhancing deep learning models with additional knowledge-based techniques. For instance, Swati et al. (2023) propose a framework for predicting political bias in news headlines. Their framework employs a neural knowledge model and knowledge graphs to simplify, interpret, and explain events that are not explicitly stated in the headlines.

Furthermore, there have been numerous attempts to analyze themes in the media (Korenčić et al., 2018). An important direction in this domain is the analysis of trends and dynamics of themes. To identify changes in themes and content of posts on social media, Zhong & Schweidel (2020) constructed a thematic model with multiple latent change points. This allows the identification of temporal patterns in themes, including the prevalence of themes that remain constant over time, temporary changes in theme prevalence, and themes that continue to change over time. This is particularly important in the context of social media content analysis, as content such as discussions changes rapidly, necessitating further monitoring of relevant information (Srikanth et al.,

2 Motivacija i pozadina

Metode i tehnike integrirane u sustave za analizu medijskih tekstova i praćenje medija trebale bi omogućiti automatsku ekstrakciju relevantnih informacija iz različitih medijskih izvora, nudeći sveobuhvatan pregled sadržaja, odnosa, javnog mišljenja ili sentimenta te trendova povezanih s temama i entitetima od interesa. Razvijanje metoda za ekstrakciju relevantnih informacija iz velike količine nestrukturiranih podataka iz više izvora u različitim formatima ili slobodnom tekstu je izazovan zadatak. Dodatni izazovi u istraživanju medijskih podataka uključuju stalno mijenjanje trendova, veliki broj različitih tema i entiteta koji se raspravljaju, senzacionalizam, pristrane tekstove, dezinformacije i slično.

Istraživanje arhitektura neuronskih mreža rezultiralo je pojavom širokog spektra modela za reprezentaciju teksta temeljenih na neuronskim mrežama (Babić i sur., 2020). Posebno transformeri, poput BERT-a ili GPT-a, donijeli su značajne promjene u području obrade prirodnog jezika (engl. Natural Language Processing, NLP). Ovi modeli, koji su unaprijed trenirani na opsežnim tekstualnim skupovima podataka obuhvaćaju jezične strukture i semantičke odnose, omogućujući im da budu bolji od konvencionalnih pristupima u raznim NLP zadacima. Utjecaj LLM-ova nadilazi postojeće pristupe; oni olakšavaju stvaranje sofisticiranijih aplikacija usmjerenih na jezik. LLM-ovi koriste duboke neuronske arhitekture s milijardama parametara, omogućujući im razumijevanje složenih obrazaca i kontekstualnih suptilnosti u tekstu. Posljedično, LLM-ovi su pokazali poboljšane performanse u zadacima poput klasifiacije teksta (Sun et al., 2019; Balkus i Yan, 2022), analize sentimenta (Babić i sur., 2021; Beliga i sur., 2021), identifikacije parafraziranja (Vrbanec i Meštrović, 2023), strojnog prevođenja (Zhu i sur., 2020; Yang i sur., 2020), odgovaranja na pitanja (Wang i sur., 2019), predviđanja širenja informacija (Meštrović i sur., 2022) i tako dalje.

U području analize medijskih tekstova i zadataka praćenja medija, brojne studije primijenile su metode i tehnike temeljene na dubokom učenju. Na primjer, Bhoi i sur. (2020) predlažu okvir za analizu tekstova sa društvenih medija temeljen na dubokom učenju za upravljanje resursima u slučaju katastrofa. Postoje istraživački pravci usmjereni na poboljšanje modela dubokog učenja dodatnim tehnikama temeljenim na znanju. Primjerice, Swati i sur. (2023) predlažu okvir za predviđanje političke pristranosti u naslovima vijesti. Njihov okvir koristi neuronski model znanja i grafove znanja za pojednostavljenje, interpretaciju i objašnjavanje događaja koji nisu eksplicitno navedeni u naslovima.

2021). Besides new approaches, which include the application of deep neural networks, a new direction in NLP research involves embedding background knowledge into learning models. For example, Koloski et al. (2022) recently proposed a new approach for detecting fake news by combining contextual representation with knowledge graphs as a heterogeneous representation of text (Koloski et al., 2022).

Morover, it is important to develop methods and techniques for media txts analysis in low-resourced and medium-resourced languages. In the case of the Croatian language, there have been several studies dedicated to media text analysis and media monitoring. The recent successful application of NLP-based methods for media monitoring has been described in many studies. For example, in Barić et al. (2023), the authors present a sentiment and tone analysis of Croatian news headlines. Sentiment analysis can be combined with topic modeling, as demonstrated in the research analyzing COVID-19-related news on Croatian internet portals (Buhin Pandur et al., 2021). Furthermore, using transfer learning techniques, where models trained on high-resource languages are adapted to low-resource languages, can be a viable strategy. For instance, cross-lingual embeddings and multilingual models such as FinEst BERT and CroSloEngual BERT (Ulčar & Robnik-Šikonja, 2020) can facilitate the transfer of knowledge from well-resourced languages to less-resourced ones, improving the performance of media texts analysis. Additionally, the use of generative AI techniques can be particularly beneficial for low-resourced languages. By incorporating external knowledge sources and fine-tuning on available datasets, generative AI can produce high-quality analyses even in languages with limited resources. The adaptability of these models makes them ideal for handling the diverse linguistic structures and vocabularies of less-resourced languages, thereby enhancing the accuracy and effectiveness of media texts analysis across different linguistic contexts.

Most existing research studies focus on specific domains and specific tasks related to media text analysis (e.g., sentiment analysis, disinformation detection, or topic modeling). Therefore, there is a need for studies on systems that integrate all possible NLP solutions for media text analysis into one comprehensive system. This study aims to explore cutting-edge NLP approaches and techniques based on LLMs and to propose a set of methods for media text analysis.

Osim toga, bilo je brojnih pokušaja analize tema u medijima (Korenčić i sur., 2018). Važan smjer u toj domeni je analiza trendova i dinamike tema. Kako bi identificirali promjene u temama i sadržaju objava na društvenim medijima, Zhong i Schweidel (2020) izgradili su tematski model s više latentnih točaka promjene. Ovo omogućava identifikaciju vremenskih obrazaca u temama, uključujući zastupljenost tema koje ostaju konstantne tijekom vremena, privremene promjene u zastupljenosti tema te zastupljenost tema koje se nastavljaju mijenjati tijekom vremena. Ovo je posebno važno u kontekstu analize sadržaja društvenih medija, budući da se sadržaji poput rasprava brzo mijenjaju te je potrebno dodatno praćenje relevantnih informacija (Srikanth i sur., 2021). Osim novih pristupa, koji uključuju primjenu dubokih neuronskih mreža, novi smjer istraživanja u domeni NLP-a uključuje ugrađivanje pozadinskog znanja u modele učenja. Tako, primjerice, Koloski i sur. (2022) nedavno predlažu novi pristup za otkrivanje lažnih vijesti kombiniranjem kontekstualne reprezentacije s grafovima znanja kao heterogene reprezentacije teksta.

Također, važno je razvijati metode i tehnike za analizu medijskih tekstova u jezicima s malo resursa i jezicima sa srednje dostupnim resursima. U slučaju hrvatskog jezika, postoji više studija posvećenih analizi medijskih tekstova i praćenju medija. Nedavna uspješna primjena metoda temeljenih na NLP-u za praćenje medija opisana je u mnogim studijama. Na primjer, Barić i sur. (2023) predstavljaju analizu sentimenta i tona naslova vijesti na hrvatskom jeziku. Analiza sentimenta može se kombinirati s modeliranjem tema, kao što je prikazano u istraživanju analize vijesti vezanih za COVID-19 na hrvatskim internet portalima (Buhin Pandur i sur., 2021). Nadalje, korištenje tehnika prijenosnog učenja, gdje se modeli trenirani na jezicima s puno resursa prilagođavaju jezicima s malo resursa, može biti održiva strategija. Na primjer, višejezični modeli kao što su FinEst BERT i CroSloEngual BERT (Ulčar i Robnik-Šikonja, 2020) mogu olakšati prijenos znanja s jezika s puno resursa na jezike s malo resursa, poboljšavajući performanse sustava za analizu medijskih tekstova. Dodatno, uporaba generativnih AI tehnika može biti posebno korisna za jezike s malo resursa. Integriranjem vanjskih izvora znanja i finim podešavanjem dostupnih skupova podataka, generativna AI može proizvesti visokokvalitetne analize čak i u jezicima s ograničenim resursima. Prilagodljivost ovih modela čini ih idealnima za obradu raznolikih jezičnih struktura i vokabulara jezika s malo resursa, čime se povećava točnost i učinkovitost analize medijskih tekstova u različitim jezičnim kontekstima.

Većina postojećih istraživačkih studija fokusira se na specifične domene i specifične zadatke vezane uz analizu medijskih tekstova (npr. analiza

3 Methods and Techniques for Media Texts Analysis

3.1 LLMs and Media Texts Analysis

Large language models, such as BERT, GPT-3, and their successors, have already demonstrated potential in the field of NLP (Balkus & Yan, 2022). These models leverage extensive pre-training on vast text corpora, allowing them to capture intricate linguistic patterns and semantic nuances that are essential for understanding media texts. One of the primary ways LLMs are employed in media texts analysis tasks is through their capability to perform various NLP functions with high accuracy and efficiency.

For keyword extraction, LLMs can identify and extract relevant keywords from a body of text, facilitating the quick identification of key themes and topics within large datasets. This capability is particularly useful in monitoring media coverage on specific issues or events, enabling stakeholders to stay informed about the most pertinent aspects without sifting through vast amounts of information. In topic modeling, LLMs can analyze large volumes of text to uncover underlying themes and trends. By clustering related terms and phrases, these models can generate comprehensive summaries of the main topics being discussed in the media. This is invaluable for organizations seeking to understand public discourse and sentiment around particular subjects or to track the evolution of media narratives over time. NER is another task where LLMs can be useful. By accurately identifying and categorizing entities such as people, organizations, locations, and events mentioned in media texts, LLMs facilitate the organization and retrieval of information. This enables users to quickly pinpoint specific entities of interest and understand their relevance within the context of the media coverage.

Large language models (LLMs) are highly effective for media text classification due to their ability to process and analyze large volumes of text data with remarkable accuracy and efficiency. Within the media texts analysis task, text classification involves categorizing articles, posts, and other media content into relevant categories such as topics, sentiment, or relevance to specific events or entities. LLMs, like BERT and GPT-3, excel in this task by leveraging their deep understanding of language semantics and context, which they acquire through extensive pre-training on diverse text corpora. This enables them to discern subtle differences in language use, such as distinguishing between positive and negative sentiments or identifying nuanced topics within complex text. Furthermore, LLMs can handle the variability and richness of media content, including slang, idioms, and varying writing styles, ensuring accurate classification across different

sentimenta, otkrivanje dezinformacija ili modeliranje tema). Stoga postoji potreba za studijama koje istražuju sustave koji integriraju sva moguća NLP rješenja za analizu medijskih tekstova u jedan sveobuhvatan sustav. Ova studija ima za cilj istražiti najnovije NLP pristupe i tehnike temeljene na LLM-ovima te predložiti skup metoda za analizu medijskih tekstova.

3 Metode i tehnike za analizu medijskih tekstova

3.1 LLM-ovi i analiza medijskih tekstova

Veliki jezični modeli, kao što su BERT, GPT-3 i njihovi nasljednici, već su pokazali potencijal u području obrade prirodnog jezika (Balkus i Yan, 2022). Ovi modeli koriste opsežno pre-treniranje na ogromnim tekstualnim korpusima, što im omogućuje obuhvaćanje složenih jezičnih obrazaca i semantičkih nijansi koje su ključne za razumijevanje medijskih tekstova. Jedan od glavnih načina na koji se LLM-ovi koriste u zadacima analize medijskih tekstova je njihova sposobnost obavljanja raznih NLP zadataka s visokim stupnjem točnosti i učinkovitosti.

Za ekstrakciju ključnih riječi, LLM-ovi mogu identificirati i izdvojiti relevantne ključne riječi iz teksta, olakšavajući brzo prepoznavanje ključnih tema i pitanja unutar velikih skupova podataka. Ova sposobnost posebno je korisna u praćenju na koje načine mediji pokrivaju određena pitanja ili događaje, omogućujući dionicima da budu informirani o najvažnijim aspektima bez potrebe za pretraživanjem ogromnih količina informacija. U modeliranju tema, LLM-ovi mogu analizirati velike količine teksta kako bi otkrili temeljne teme i trendove. Grupiranjem povezanih pojmova i fraza, ovi modeli mogu generirati sveobuhvatne sažetke glavnih tema koje se raspravljaju u medijima. To je vrlo važno za organizacije koje žele razumjeti javni diskurs i sentiment o određenim temama ili pratiti promjene medijskih narativa tijekom vremena.

Prepoznavanje imenovanih entiteta još je jedan zadatak u kojem LLM-ovi mogu biti korisni. Točnim prepoznavanjem i kategorizacijom entiteta kao što su ljudi, organizacije, lokacije i događaji spomenuti u medijskim tekstovima, LLM-ovi olakšavaju organizaciju i dohvat informacija. To korisnicima omogućuje brzo prepoznavanje specifičnih entiteta od interesa i razumijevanje njihove relevantnosti u kontekstu medija.

LLM-ovi vrlo su učinkoviti za klasifikaciju medijskih tekstova zbog svoje sposobnosti obrade i analize velikih količina tekstualnih podataka s visiokom točnošću i učinkovitošću. U okviru analize medijskih tekstova, klasifikacija teksta

media sources. Their capability to learn and adapt to new data through fine-tuning makes them particularly adept at evolving with the dynamic nature of media narratives. By integrating LLMs into media monitoring systems that enables automatic text analysis, organizations can achieve more precise and insightful classifications, facilitating better tracking of trends, public opinion, and emerging issues, ultimately enabling more informed decision-making and strategic communication.

Sentiment classification is enhanced by the nuanced understanding that LLMs have of language. These models can determine the sentiment expressed in media texts—whether positive, negative, or neutral—with high precision. This allows for a more granular analysis of public opinion and the emotional tone of media coverage, which is crucial for reputation management and crisis communication. Disinformation detection benefits greatly from the advanced pattern recognition capabilities of LLMs. By analyzing text for inconsistencies, factual inaccuracies, and other indicators of false information, LLMs can help identify and flag potential disinformation in media content. This is particularly important in the current digital landscape, where the spread of false information can have significant consequences.

Additionally, the flexibility of LLMs allows for the incorporation of advanced techniques such as prompt engineering, retrieval-augmented generation, and fine-tuning. These techniques can further enhance the performance of LLMs in media texts analysis tasks by adapting them to specific contexts and improving their ability to generate accurate and contextually relevant outputs.

3.2 Enhancing Media Texts Analysis with Retrieval-Augmented Generation Techniques

RAG is a versatile fine-tuning technique that combines pre-trained parametric and non-parametric memory for language generation (Lewis et al., 2020). This method has demonstrated significant promise by integrating knowledge from external databases, leading to enhanced accuracy and credibility of models, especially for tasks that require substantial domain knowledge (Gao et al., 2023). It also facilitates continuous knowledge updates and the incorporation of domain-specific information.

In the context of media txts analysis, the RAG technique can provide a powerful tool for improving various tasks. By merging the strengths of retrieval-based and generation-based models, RAG enhances the accuracy and contextual relevance of media analysis. This hybrid approach leverages external knowledge sources such as knowledge graphs, background knowledge databases, and real-time information feeds to provide contextually rich and precise analyses. RAG operates by first retrieving

uključuje kategorizaciju članaka, objava i drugih medijskih sadržaja u relevantne kategorije kao što su teme, sentiment ili relevantnost za određene događaje ili entitete. LLM-ovi, poput BERT-a i GPT-3, pokazali su se dobri u ovom zadatku iskorištavanjem svog dubokog razumijevanja semantike i konteksta, koje stječu kroz opsežan pred-trening na raznolikim tekstualnim korpusima. To im omogućuje prepoznavanje razlika u korištenju jezika, pozitivnog, negativnog i neutralnog sentimenata, prepoznavanje nijansiranih tema unutar složenih tekstova i sl. Nadalje, LLM-ovi mogu rukovati varijabilnošću i bogatstvom medijskog sadržaja, uključujući žargon, idiome i različite stilove pisanja, osiguravajući točnu klasifikaciju kroz različite medijske izvore. Njihova sposobnost učenja i prilagodbe novim podacima kroz fino podešavanje čini ih posebno vještima u evoluciji s dinamičnom prirodom medijskih narativa. Integracijom LLM-ova u sustave praćenja medija, organizacije mogu postići preciznije i dublje klasifikacije, olakšavajući bolje praćenje trendova, javnog mišljenja i novih pitanja, što u konačnici omogućuje informiranije donošenje odluka i stratešku komunikaciju.

Klasifikacija sentimenta poboljšana je nijansiranim razumijevanjem jezika koje posjeduju LLM-ovi. Ovi modeli mogu s velikom preciznošću odrediti sentiment izražen u medijskim tekstovima - bilo pozitivan, negativan ili neutralan. To omogućuje detaljniju analizu javnog mišljenja i emocionalnog tona medijskog teksta, što je ključno za upravljanje reputacijom i komunikaciju u kriznim situacijama. Otkrivanje dezinformacija također ima velike koristi od naprednih sposobnosti prepoznavanja obrazaca LLM-ova. Analizom teksta za nedosljednosti, činjenične netočnosti i druge indikatore lažnih informacija, LLM-ovi mogu pomoći u identificiranju i označavanju potencijalnih dezinformacija u medijskom sadržaju. To je posebno važno u današnjem digitalnom okruženju, gdje širenje lažnih informacija može imati negativne posljedice.

Osim toga, fleksibilnost LLM-ova omogućuje primjenu naprednih tehnika poput inženjeringa upita, generiranja s poboljšanjem pretrage i finog podešavanja. Ove tehnike mogu dodatno poboljšati performanse LLM-ova u zadacima analize medijskih tekstova prilagođavanjem specifičnim kontekstima i poboljšanjem njihove sposobnosti generiranja točnih i kontekstualno relevantnih rezultata.

3.2 Unaprijeđivanje analize medijskih tekstova primjenom RAG tehnika

RAG je tehnika za optimizaciju LLM-ova koja kombinira unaprijed treniranu parametrijsku i neparametrijsku memoriju za generiranje jezika (Lewis et al., 2020). Ova tehnika je pokazala

relevant information from external sources before generating an analysis or summary of the media content. This retrieval step is crucial because it allows the model to access and incorporate up-to-date and domain-specific knowledge, enhancing the quality and accuracy of the output. By integrating this external knowledge, RAG can produce insights that are not only linguistically accurate but also contextually appropriate and informative.

One of the key benefits of RAG in media texts analysis tasks is its use of knowledge graphs. Knowledge graphs, such as Google Knowledge Graph, WordNet, or DBPedia, provide structured information about entities and their relationships. By leveraging these graphs, RAG can better understand the context and meaning of complex terms or concepts within media texts. For example, when monitoring news articles, the model can retrieve relevant background information and explanations from a knowledge graph, ensuring that the analysis remains accurate and informative. Background knowledge databases play a similar role in enhancing media texts analysis. These databases contain extensive information on various topics, which the model can use to generate more comprehensive and context-aware analyses. For instance, when monitoring reports on geopolitical events, the model can access background information about specific regions or historical contexts, allowing it to provide analyses that maintain the richness and depth of the original content.

Dictionaries and thesauri are also valuable resources in the RAG framework. They offer synonyms, definitions, and usage examples that the model can use to clarify complex terms and ensure that the analysis is both accurate and accessible. This capability is particularly useful for maintaining the readability of the content while ensuring that the insights are still precise and contextually appropriate. For example, technical jargon in economic media texts can be clarified with simpler terms, making the analysis more understandable to a broader audience.

3.3 Fine-Tuning Techniques of LLMs for Media Texts Analysis Tasks

Large language models can be fine-tuned for various tasks related to media texts analysis. Fine-tuning these models allows them to adapt specifically to the intricacies of media content, improving their performance on tasks such as sentiment analysis, named entity recognition, and disinformation detection. By leveraging their pre-trained knowledge, LLMs can be tailored to recognize and classify specific topics, identify key entities like people and organizations, and even detect changes in public sentiment over time. For instance, in sentiment analysis, a fine-tuned LLM can accurately gauge the public's reaction to a news event, providing valuable

značajan potencijal integriranjem znanja iz vanjskih baza podataka, što dovodi do poboljšane točnosti i vjerodostojnosti modela, posebno za zadatke koji zahtijevaju značajno domensko znanje (Gao et al., 2023). Također olakšava kontinuirano ažuriranje znanja i uključivanje informacija specifičnih za određenu domenu.

U kontekstu analize medijskih tekstova, tehnika RAG može pružiti moćan alat za poboljšanje raznih zadataka. Spajanjem modela temeljenih na pretraživanju i generiranju teksta, RAG povećava točnost i kontekstualnu relevantnost analize medija. Ovaj hibridni pristup koristi vanjske izvore znanja kao što su grafovi znanja, baze podataka s pozadinskim znanjem i izvori informacija u stvarnom vremenu kako bi pružio kontekstualno bogate i precizne analize. RAG funkcionira tako da prvo pretražuje relevantne informacije iz vanjskih izvora prije nego što provede analizu ili generira sažetak medijskog sadržaja. Ovaj korak pretraživanja je ključan jer omogućuje modelu pristup i uključivanje ažuriranih i specifičnih znanja, čime se poboljšava kvaliteta i točnost rezultata. Integriranjem vanjskog znanja, RAG može proizvesti uvide koji su ne samo jezično točni već i kontekstualno prikladni i informativni.

Jedna od ključnih prednosti RAG-a u analizi medijskih tekstova je njegova uporaba grafova znanja. Grafovi znanja, poput Google Knowledge grafa, WordNet-a ili DBPedije, pružaju strukturirane informacije o entitetima i njihovim odnosima. Korištenjem ovih grafova, RAG može bolje razumjeti kontekst i značenje složenih pojmova ili koncepata unutar medijskih tekstova. Na primjer, prilikom praćenja novinskih članaka, model može preuzeti relevantne pozadinske informacije i objašnjenja iz grafova znanja, osiguravajući da analiza ostane točna i informativna. Baze podataka s pozadinskim znanjem igraju sličnu ulogu u poboljšanju praćenja medija. Ove baze podataka sadrže opsežne informacije o raznim temama koje model može koristiti za generiranje sveobuhvatnijih i kontekstualno svjesnijih analiza. Na primjer, prilikom praćenja izvještaja o geopolitičkim događajima, model može pristupiti pozadinskim informacijama o specifičnim regijama ili povijesnom kontekstu, omogućujući analize koje zadržavaju bogatstvo i dubinu izvornog sadržaja.

Rječnici i tezaurusi također su vrijedni resursi u okviru RAG-a. Oni nude sinonime, definicije i primjere upotrebe koje model može koristiti za pojašnjavanje složenih pojmova i osiguravanje da analiza bude i točna i dostupna. Ova sposobnost je posebno korisna za održavanje čitljivosti sadržaja, dok se osigurava da su uvjeti još uvijek precizni i kontekstualno prikladni. Na primjer, tehnički žargon u ekonomskim medijskim tekstovima može se pojasniti jednostavnijim pojmovima, čineći analizu razumljivijom širem krugu publike.

insights for strategic communication. Similarly, in named entity recognition, these models can consistently identify and categorize entities, even in diverse and unstructured media sources.

Furthermore, fine-tuned LLMs can enhance disinformation detection by identifying subtle cues and inconsistencies that signal false information. This tailored approach not only increases the accuracy and relevance of media texts analysis but also enables organizations to respond swiftly and effectively to the evolving media landscape.

3.4. An Overview of the NLP Methods for Media Texts Analysis

Integrating NLP methods based on GenAI into a comprehensive system for media text analysis offers a robust framework for processing and understanding vast amounts of media content. By combining various NLP techniques, such as keyword extraction, NER, sentiment analysis, topic modeling, hate speech detection, and disinformation detection, a holistic approach to media texts anaylsis can be achieved. Each method plays a critical role as it is shown in Table 1. These technologies are particularly powerful when applied to the analysis of trends and changes over time, allowing stakeholders to monitor the evolution of public discourse and sentiment across different media platforms.

Table 1. Description of NLP tasks and methods for media texts analysis

NLP task	**Media texts analysis - application description**
Keyword Extraction	Extraction of key terms that highlight the central topics in media content. **Example:** Fine-tuning LLMs for KE in online media texts
Topic Modelling	Detection of main topics that are in the focus of media. **Example:** Fine-tuning LLMs for topic modeling in online media texts. This method can be combined with sentiment analysis to assess sentiment related to specific topics and can include the analysis of trends and changes over time.
NER	Identification of specific entities such as people, organizations, and

3.3 Tehnike finog podešavanja velikih jezičnih modela za analizu medijskih tekstova

Veliki jezični modeli mogu se fino podešavati za razne zadatke povezane s analizom medijskih tekstova. Fino podešavanje LLM-ova omogućuje prilagodbu specifičnostima medijskog sadržaja, poboljšavajući njihovu izvedbu u zadacima kao što su analiza sentimenta, prepoznavanje imenovanih entiteta i otkrivanje dezinformacija. Korištenjem svog pred-treninga, LLM-ovi mogu se prilagoditi za prepoznavanje i klasifikaciju specifičnih tema, identifikaciju ključnih entiteta kao što su ljudi i organizacije, pa čak i detekciju promjena u javnom sentimentu tijekom vremena. Na primjer, u analizi sentimenta, fino podešavan LLM može točno procijeniti reakciju javnosti na neki događaj u vijestima, pružajući vrijedne uvide za stratešku komunikaciju. Slično, u prepoznavanju imenovanih entiteta, ovi modeli mogu dosljedno identificirati i kategorizirati entitete, čak i u raznolikim i nestrukturiranim medijskim izvorima.

Nadalje, fino podešavani LLM-ovi mogu poboljšati otkrivanje dezinformacija identificiranjem suptilnih znakova i nedosljednosti koje ukazuju na lažne informacije. Ovaj prilagođeni pristup ne samo da povećava točnost i relevantnost analize medijskih tekstova, već također omogućuje organizacijama da brzo i učinkovito reagiraju na promjenjivost medija.

3.4. Tehnike finog podešavanja velikih jezičnih modela za analizu medijskih tekstova

Integracija NLP metoda temeljenih na GenAI u sveobuhvatan sustav za analizu medijskih tekstova nudi snažan okvir za obradu i razumijevanje velike količine medijskog sadržaja. Kombiniranjem različitih NLP tehnika, poput ekstrakcije ključnih riječi, NER, analize sentimenta, modeliranja tema, otkrivanja govora mržnje i otkrivanja dezinformacija, može se postići cjelovit pristup analizi medijskih tekstova. Svaka metoda ima važnu ulogu, što je prikazano u Tablici 1. Ove tehnologije su posebno moćne kada se primjenjuju na analizu trendova i promjena tijekom vremena, omogućujući dionicima praćenje evolucije javnog diskursa i sentimenta na različitim medijskim platformama.

Tablica 1. Opis NLP zadataka i metoda za analizu medijskih tektova

NLP zadatak	**Analiza medijskih tekstova – opis primjene**

	locations mentioned in media content. **Example:** Applying LLMs for the identification of crucial persons, locations, institutions, and other important entities appearing in online media. This method can be combined with sentiment analysis to assess sentiment related to specific entities and can include the analysis of trends and changes over time.
Sentiment Analysis	Evaluation of the emotional tone expressed in media content, determining whether it is positive, negative, or neutral. **Example:** Prompt engineering with LLMs in combination with RAG, incorporating external knowledge about sentiment.
Hate Speech Recognition	Identification of offensive or harmful language in media content that targets individuals or groups based on attributes like race, religion, or gender. **Example:** Fine-tuning LLMs for the task of media text classification into two or more classes, including one that contains hate speech.
Disinformation Detection	Detection of false or misleading information in media content that is intentionally spread to deceive or misinform the audience. **Example:** Fine-tunning LLMs in combination with RAG, incorporating external knowledge about the domain of interest.

By fine-tuning LLMs for these specific tasks, the system can efficiently handle diverse and unstructured media data, providing deeper insights and more accurate classifications. Moreover, integrating external knowledge sources through RAG

Ekstrakcija ključnih riječi	Ekstrakcija ključnih pojmova koji ističu središnje teme u medijskom sadržaju. **Primjer:** Fino podešavanje LLM-ova za ekstrakciju ključnih pojmova u online medijskim tekstovima.
Modeliranje tema	Otkrivanje glavnih tema koje su u fokusu medija. **Primjer:** Fino podešavanje LLM-ova za modeliranje tema u online medijskim tekstovima. Ova metoda se može kombinirati s analizom sentimenta kako bi se odredio sentiment povezan s određenim temama te može uključivati analizu trendova i promjena tijekom vremena.
NER	Identifikacija specifičnih entiteta kao što su ljudi, organizacije i lokacije spomenuti u medijskom sadržaju. **Primjer:** Primjena LLM-ova za identifikaciju ključnih osoba, lokacija, institucija i drugih važnih entiteta koji se pojavljuju u online medijima. Ova metoda se može kombinirati s analizom sentimenta kako bi se procijenio sentiment povezan s određenim entitetima te može uključivati analizu trendova i promjena tijekom vremena.
Analiza sentimenta	Analiza sentimenta izraženog u medijskom sadržaju, određivanje je li pozitivan, negativan ili neutralan. **Primjer:** Prompt inženjering upita s LLM-ovima u kombinaciji s RAG-om, uz uključivanje vanjskog znanja o sentimentu.
Prepoznavanje govora mržnje	Identifikacija govora mržnje u medijskom sadržaju koji cilja pojedince ili skupine na

and employing techniques like prompt engineering can further enhance the system's capability, ensuring that it remains up-to-date and contextually relevant. This comprehensive approach not only improves the accuracy and efficiency of media analysis but also enables stakeholders to respond quickly and effectively to emerging trends and issues in the media landscape.

4 Conclusion

This paper has focused on the methods for media texts analysis, particularly those based on GenAI. GenAI-based methods offer significant improvements in accuracy, efficiency, and contextual understanding over traditional NLP approaches, enabling more effective analysis and interpretation of media texts.

The main contribution of this paper is an overview of NLP methods based on GenAI and their potential in the field of media text analysis. This is the first step toward the definition and implementation of a comprehensive media monitoring system.

The integration of large language models into systems for media texts analysis and media monitoring marks a significant leap forward in the ability to analyze and understand media content. These models excel at performing complex NLP tasks with high precision and adaptability, making them indispensable tools for modern media monitoring and media texts analysis. Their ability to handle diverse and unstructured media texts, capture intricate linguistic patterns, and provide nuanced insights sets them apart from classical methods.

Furthermore, the application of RAG to media texts analysis represents a substantial advancement, enabling more accurate, informed, and contextually relevant analyses. By combining the strengths of retrieval and generation capabilities with extensive external knowledge, RAG enhances the overall quality and utility of media txts analysis efforts. This hybrid approach ensures that the models are not only current with real-time information but also enriched with domain-specific knowledge, thereby improving their effectiveness in detecting misinformation, conducting sentiment analysis, and extracting key entities and topics.

Future research in this direction will focus on developing and refining GenAI-based systems for media texts analysis. This includes enhancing the integration of external knowledge sources, improving fine-tuning techniques, and optimizing model interactions through advanced methods such as prompt engineering. By continuing to innovate and apply GenAI technologies, we can further elevate the capabilities of media texts analysis systems, ensuring they remain robust, accurate, and contextually aware in an ever-evolving digital landscape.

	temelju atributa poput rase, religije ili spola. **Primjer:** Fino podešavanje LLM-ova za zadatak klasifikacije medijskih tekstova u dvije ili više kategorija, uključujući jednu koja sadrži govor mržnje.
Detekcija dezinformacija	Otkrivanje lažnih ili obmanjujućih informacija u medijskom sadržaju koje se namjerno šire kako bi se prevarilo ili dezinformiralo publiku. **Primjer:** Fino podešavanje LLM-ova u kombinaciji s RAG-om, uz uključivanje vanjskog znanja o domeni interesa.

Finim podešavanjem LLM-ova za ove specifične zadatke, sustav može učinkovito upravljati raznolikim i nestrukturiranim medijskim podacima, pružajući dublje uvide i točnije klasifikacije. Nadalje, integracija vanjskih izvora znanja putem RAG-a i primjena tehnika poput inženjeringa upita mogu dodatno poboljšati sposobnosti sustava, osiguravajući da ostane ažuran i kontekstualno relevantan. Ovaj sveobuhvatan pristup ne samo da poboljšava točnost i učinkovitost analize medija, već također omogućuje dionicima da brzo i učinkovito reagiraju na nove trendove i pitanja u medijskom okruženju.

4 Zaključak

Ovaj istražuje metode analize medijskih tekstova, posebno one temeljene na gnerativnoj umejtnoj inteligenciji. Metode temeljene na GenAI nude značajna poboljšanja u točnosti, učinkovitosti i kontekstualnom razumijevanju u usporedbi s tradicionalnim pristupima NLP-a, omogućujući učinkovitiju analizu i interpretaciju medijskih tekstova.

Glavni doprinos ovog rada je pregled NLP metoda temeljenih na generativnoj umjetnoj inteligenciji i njihovog potencijala u području analize medijskih tekstova. Ovo je prvi korak prema definiranju i implementaciji sveobuhvatnog sustava za praćenje medija.

Integracija LLM-ova u sustave za analizu medijskih tekstova i praćenje medija predstavlja značajan iskorak u sposobnosti analize i razumijevanja medijskog sadržaja. Veliki jezični modeli izvrsno obavljaju složene NLP zadatke s visokom preciznošću i prilagodljivošću, čineći ih

Acknowledgments

This work has been supported by the UNIRI project uniri-iskusni-drustv-23-95.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

Babić, K., Martinčić-Ipšić, S., & Meštrović, A. (2020). Survey of neural text representation models. Information, 11(11), 511.

Babić, K., Petrović, M., Beliga, S., Martinčić-Ipšić, S., Matešić, M., & Meštrović, A. (2021). Characterisation of COVID-19-related tweets in the Croatian language: framework based on the Cro-CoV-cseBERT model. Applied Sciences, 11(21), 10442.

Balkus, S. V., & Yan, D. (2022). Improving short text classification with augmented data using GPT-3. Natural Language Engineering, 1-30.

Barić, A., Majer, L., Dukić, D., Grbeša-Zenzerović, M., & Šnajder, J. (2023, May). Target Two Birds With One SToNe: Entity-Level Sentiment and Tone Analysis in Croatian News Headlines. In Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023) (pp. 78-85).

Beliga, S., Martinčić-Ipšić, S., Matešić, M., Petrijevčanin Vuksanović, I., & Meštrović, A. (2021). Infoveillance of the Croatian online media during the COVID-19 pandemic: one-year longitudinal study using natural language processing. JMIR public health and surveillance, 7(12), e31540.

Beliga, S., Martinčić-Ipšić, S., Matešić, M., & Meštrović, A. (2022). Natural language processing and statistic: the first six months of the COVID-19 infodemic in Croatia. In The Covid-19 pandemic as a challenge for media and communication studies (pp. 78-92). Routledge.

Bhoi, Ashutosh, Sthita Pragyan Pujari, and Rakesh Chandra Balabantaray. "A deep learning-based social media text analysis framework for disaster resource management." Social Network Analysis and Mining 10 (2020): 1-14.

Buhin Pandur, M., Dobša, J., Beliga, S., & Meštrović, A. (2021). Topic modelling and sentiment analysis of COVID-19 related news on Croatian Internet portal. In SiKDD Conference on Data Mining and Data Warehouses. Slovenian

neophodnim alatima za moderno praćenje i analizu medija. Njihova sposobnost obrade raznolikih i nestrukturiranih medijskih tekstova, hvatanje složenih jezičnih obrazaca i pružanje nijansiranih uvida izdvaja ih od klasičnih metoda.

Nadalje, primjena RAG tehnika na sustave za analizu medijskih tekstova predstavlja značajan napredak, omogućujući točnije, informiranije i kontekstualno relevantne analize. Kombinirajući sposobnosti pretraživanja i generiranja s opsežnim vanjskim znanjem, RAG poboljšava ukupnu kvalitetu i korisnost postupaka analize medijskih tekstova. Ovaj hibridni pristup osigurava da su modeli ne samo u skladu s informacijama u stvarnom vremenu, već i obogaćeni znanjem specifičnim za određenu domenu, čime se poboljšava njihova učinkovitost u otkrivanju dezinformacija, provođenju analize sentimenta i izdvajanja ključnih entiteta i tema.

Buduća istraživanja u ovom smjeru fokusirat će se na razvoj i usavršavanje sustava za analizu medijskih tekstova temeljenih na generativnoj umjetnoj inteligenciji. To uključuje poboljšanje integracije vanjskih izvora znanja, usavršavanje tehnika finog podešavanja i optimizaciju interakcija modela kroz napredne metode poput inženjeringa upita. Nastavljajući inovirati i primjenjivati tehnologije GenAI, možemo dodatno unaprijediti sposobnosti sustava za analizu medijskoh tekstova, osiguravajući da ostanu robusni, točni i kontekstualno svjesni u stalno promjenjivoj digitalnoj okolini.

Zahvale

Ovo istraživanje i rad podržan je UNIRI projektom uniri-iskusni-drustv-23-95.

Reference

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

Babić, K., Martinčić-Ipšić, S., & Meštrović, A. (2020). Survey of neural text representation models. Information, 11(11), 511.

Babić, K., Petrović, M., Beliga, S., Martinčić-Ipšić, S., Matešić, M., & Meštrović, A. (2021). Characterisation of COVID-19-related tweets in the Croatian language: framework based on the Cro-CoV-cseBERT model. Applied Sciences, 11(21), 10442.

Balkus, S. V., & Yan, D. (2022). Improving short text classification with augmented data using GPT-3. Natural Language Engineering, 1-30.

KDD Conference on Data Mining and Data Warehouses.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Farzindar, A., Inkpen, D., & Hirst, G. (2015). Natural language processing for social media. San Rafael: Morgan & Claypool.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., ... & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997.

Germann, U., Liepins, R., Gosko, D., & Barzdins, G. (2018, July). SUMMA: Integrating multiple NLP technologies into an open-source platform for multilingual media monitoring. In ACL 2018 Workshop for Natural Language Processing Open Source Software (pp. 47-51). Association for Computational Linguistics.

Koloski, B, Stepišnik Perdih, T., Robnik-Šikonja, M., Pollak, S., and Škrlj, B. 2022. "Knowledge graph informed fake news classification via heterogeneous representation ensembles." Neurocomputing.

Korenčić, D., Ristov, S., & Šnajder, J. (2018). Document-based topic coherence measures for news media text. Expert systems with Applications, 114, 357-373.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459-9474.

Meštrović, A., Petrović, M., & Beliga, S. (2022). Retweet prediction based on heterogeneous data sources: the combination of text and multilayer network features. Applied Sciences, 12(21), 11216.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

Srikanth M., Liu A., Adams-Cohen N., Cao J., Alvarez R. M., and Anandkumar A.. 2021. Dynamic Social Media Monitoring for Fast-Evolving Online Discussions. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD '21). Association for Computing Machinery, New York, NY, USA, 3576–3584. https://doi.org/10.1145/3447548.3467171

Sufi, F. (2023). Social media analytics on Russia–Ukraine cyber war with natural language

Barić, A., Majer, L., Dukić, D., Grbeša-Zenzerović, M., & Šnajder, J. (2023, May). Target Two Birds With One SToNe: Entity-Level Sentiment and Tone Analysis in Croatian News Headlines. In Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023) (pp. 78-85).

Beliga, S., Martinčić-Ipšić, S., Matešić, M., Petrijevčanin Vuksanović, I., & Meštrović, A. (2021). Infoveillance of the Croatian online media during the COVID-19 pandemic: one-year longitudinal study using natural language processing. JMIR public health and surveillance, 7(12), e31540.

Beliga, S., Martinčić-Ipšić, S., Matešić, M., & Meštrović, A. (2022). Natural language processing and statistic: the first six months of the COVID-19 infodemic in Croatia. In The Covid-19 pandemic as a challenge for media and communication studies (pp. 78-92). Routledge.

Bhoi, Ashutosh, Sthita Pragyan Pujari, and Rakesh Chandra Balabantaray. "A deep learning-based social media text analysis framework for disaster resource management." Social Network Analysis and Mining 10 (2020): 1-14.

Buhin Pandur, M., Dobša, J., Beliga, S., & Meštrović, A. (2021). Topic modelling and sentiment analysis of COVID-19 related news on Croatian Internet portal. In SiKDD Conference on Data Mining and Data Warehouses. Slovenian KDD Conference on Data Mining and Data Warehouses.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Farzindar, A., Inkpen, D., & Hirst, G. (2015). Natural language processing for social media. San Rafael: Morgan & Claypool.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., ... & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997.

Germann, U., Liepins, R., Gosko, D., & Barzdins, G. (2018, July). SUMMA: Integrating multiple NLP technologies into an open-source platform for multilingual media monitoring. In ACL 2018 Workshop for Natural Language Processing Open Source Software (pp. 47-51). Association for Computational Linguistics.

Koloski, B, Stepišnik Perdih, T., Robnik-Šikonja, M., Pollak, S., and Škrlj, B. 2022. "Knowledge graph informed fake news classification via

processing: Perspectives and challenges. Information, 14(9), 485.

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune BERT for text classification?. In Chinese computational linguistics: 18th China national conference, CCL 2019, Kunming, China, October 18–20, 2019, proceedings 18 (pp. 194-206). Springer International Publishing.

Swati, S., Mladenić, D., & Grobelnik, M. (2023). An Inferential Commonsense-Driven Framework for Predicting Political Bias in News Headlines. IEEE Access.

Ulčar, M., & Robnik-Šikonja, M. (2020). FinEst BERT and CroSloEngual BERT: less is more in multilingual models. In Text, Speech, and Dialogue: 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8–11, 2020, Proceedings 23 (pp. 104-111). Springer International Publishing.

Vrbanec, T., & Meštrović, A. (2023). Comparison study of unsupervised paraphrase detection: Deep learning—The key for semantic similarity detection. Expert systems, 40(9), e13386.

Wang, Z., Ng, P., Ma, X., Nallapati, R., & Xiang, B. (2019). Multi-passage BERT: A globally normalized BERT model for open-domain question answering. arXiv preprint arXiv:1908.08167.

Yang, J., Wang, M., Zhou, H., Zhao, C., Zhang, W., Yu, Y., & Li, L. (2020, April). Towards making the most of BERT in neural machine translation. In Proceedings of the AAAI conference on artificial intelligence (Vol. 34, No. 05, pp. 9378-9385).

Zhong N., Schweidel D. A. (2020) Capturing Changes in Social Media Content: A Multiple Latent Changepoint Topic Model. Marketing Science 39(4):827-846.

Zhu, J., Xia, Y., Wu, L., He, D., Qin, T., Zhou, W., ... & Liu, T. Y. (2020). Incorporating BERT into neural machine translation. arXiv preprint arXiv:2002.06823.

heterogeneous representation ensembles." Neurocomputing.

Korenčić, D., Ristov, S., & Šnajder, J. (2018). Document-based topic coherence measures for news media text. Expert systems with Applications, 114, 357-373.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459-9474.

Meštrović, A., Petrović, M., & Beliga, S. (2022). Retweet prediction based on heterogeneous data sources: the combination of text and multilayer network features. Applied Sciences, 12(21), 11216.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

Srikanth M., Liu A., Adams-Cohen N., Cao J., Alvarez R. M., and Anandkumar A.. 2021. Dynamic Social Media Monitoring for Fast-Evolving Online Discussions. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD '21). Association for Computing Machinery, New York, NY, USA, 3576–3584. https://doi.org/10.1145/3447548.3467171

Sufi, F. (2023). Social media analytics on Russia–Ukraine cyber war with natural language processing: Perspectives and challenges. Information, 14(9), 485.

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune BERT for text classification?. In Chinese computational linguistics: 18th China national conference, CCL 2019, Kunming, China, October 18–20, 2019, proceedings 18 (pp. 194-206). Springer International Publishing.

Swati, S., Mladenić, D., & Grobelnik, M. (2023). An Inferential Commonsense-Driven Framework for Predicting Political Bias in News Headlines. IEEE Access.

Ulčar, M., & Robnik-Šikonja, M. (2020). FinEst BERT and CroSloEngual BERT: less is more in multilingual models. In Text, Speech, and Dialogue: 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8–11, 2020, Proceedings 23 (pp. 104-111). Springer International Publishing.

Vrbanec, T., & Meštrović, A. (2023). Comparison study of unsupervised paraphrase detection: Deep learning—The key for semantic

similarity detection. Expert systems, 40(9), e13386.

Wang, Z., Ng, P., Ma, X., Nallapati, R., & Xiang, B. (2019). Multi-passage BERT: A globally normalized BERT model for open-domain question answering. arXiv preprint arXiv:1908.08167.

Yang, J., Wang, M., Zhou, H., Zhao, C., Zhang, W., Yu, Y., & Li, L. (2020, April). Towards making the most of BERT in neural machine translation. In Proceedings of the AAAI conference on artificial intelligence (Vol. 34, No. 05, pp. 9378-9385).

Zhong N., Schweidel D. A. (2020) Capturing Changes in Social Media Content: A Multiple Latent Changepoint Topic Model. Marketing Science 39(4):827-846.

Zhu, J., Xia, Y., Wu, L., He, D., Qin, T., Zhou, W., ... & Liu, T. Y. (2020). Incorporating BERT into neural machine translation. arXiv preprint arXiv:2002.06823