Cross-lingual sentiment analysis of official EU Slavic languages

Pandurang Thakkar, Gaurish

University of Zagreb

Faculty of Humanities and Social Sciences

Gaurish Pandurang Thakkar

CROSS-LINGUAL SENTIMENT ANALYSIS OF OFFICIAL EU SLAVIC LANGUAGES

DOCTORAL DISSERTATION

Zagreb, 2022

University of Zagreb

Faculty of Humanities and Social Sciences

Gaurish Pandurang Thakkar

CROSS-LINGUAL SENTIMENT ANALYSIS OF OFFICIAL EU SLAVIC LANGUAGES

DOCTORAL DISSERTATION

Supervisor:
Professor Nives Mikelić Preradović

Zagreb, 2022

Sveučilište u Zagrebu

Filozofski fakultet u Zagrebu

Gaurish Pandurang Thakkar

PREKOJEZIČNA ANALIZA MNIJENJA U SLAVENSKIM JEZICIMA SLUŽBENIMA U EU-U

DOKTORSKI RAD

Mentor:
Prof. dr. sc. Nives Mikelić Preradović

Zagreb, 2022

ABOUT THE MENTOR

Professor Nives Mikelić Preradović

Nives Mikelić Preradović is a Professor at the Department of Information Sciences, Faculty of Humanities and Social Sciences, University of Zagreb. She obtained her MA in Croatian language and literature and Information sciences at the University of Zagreb and her MPhil in Natural Language Processing at Cambridge University, UK. She obtained her PhD in 2008 at the University of Zagreb. In 2006 she spent a semester doing research and service-learning at George Washington University, USA. She is Head of the Chair for Knowledge Organization at the Department of Information Sciences.

She was the first to introduce service learning in Croatia in 2006 and has mentored and administrated over 100 SL projects with the local community since then. In 2013, she received a U.S. Department of State grant for the project “Strengthening Self-employment Capabilities and Capacities through International Service-Learning Projects”.

She was the national contact person for the KA2 project Europe Engage - Developing a Culture of Civic Engagement through Service-Learning within Higher Education in Europe (2015-2017) and the Knowledge Alliances project Rural 3.0: Service Learning for the Rural Development (https://rural.ffzg.unizg.hr/). She is currently the national contact person for two KA2 projects: SLIDE - “Service-Learning as a pedagogy to promote Inclusion, Diversity, and Digital Empowerment” and eSL4EU - “e-Service Learning for more digital and inclusive EU Higher Education systems”.

Her general research interests include service-learning, natural language processing and computer-assisted language learning. She published 2 books, 22 book chapters and about 70 research papers in international journals and conference proceedings. She has mentored over 46 final, graduate and PhD theses.

ACKNOWLEDGEMENTS

It is incredible how one’s life path can take on a form all on its own. One such occurrence resulted in my relocation to Croatia. This journey has been nothing short of spectacular and has changed my life dramatically for the better.

I was extremely fortunate to be offered an Early-Stage Researcher (and PhD) position in the CLEOPATRA Marie Curie project. Writing a doctoral dissertation is a lengthy and difficult process that requires a great deal of focus, dedication, hard work, and support. I was fortunate to have the unwavering support of everyone in my life. First and foremost, I would like to thank my mentor, Nives Mikelić Preradović, for accepting me as a student despite my lack of experience in the field of sentiment analysis. She has been with me since the beginning of this journey. She's been nothing less than a mother who loves and raises a child. Her consistent advice and insightful feedback in both my professional and personal life have had a significant impact on me. Thank you for your patience and for putting up with a slacker like me, Nives.

I would like to thank Prof. Marko Tadić, who first allowed me to work on the CLEOPATRA project and guided me in all aspects of my life. I consider myself fortunate to have him in my life as a father figure and a great boss.

I would also like to express my special appreciation and thank the members of the committee, Prof. Jan Šnajder, and Prof. Sanja Seljan for engaging with my thesis diligently and enthusiastically, and providing valuable suggestions: in particular, they helped me clarify my results and their statistical significance.

I would like to thank Diego Alves for listening to my ideas and participating in our simple but amusing conversations about feelings and other aspects of language. I am grateful for the opportunity to discuss my ideas with everyone at the Institute of Linguistics, especially Matea Filko, Vanja Štefanec, Ivana Simeon, and Daša Farkaš. I would like to thank all of my friends, previous employers, and teachers who encouraged me to take on this adventure.

Finally, I am grateful to my family, especially my mother and father, who have shown me unwavering unconditional love and support since the beginning. This work is in your honour.

We would like to express our deepest gratitude to the native speakers of South Slavic languages for their valuable feedback for verifying and correcting the negation dataset.

- Dilyana Trapcheva - Bulgarian,

- Alexandr Rosen - Czech,
- Matea Filko - Croatian,
- Elżbieta Kaczmarska and Ivana Škaričić - Polish,
- Alzbeta Brozmanova Gregorova - Slovak,
- Daša Farkaš – Slovenian.

We want to express our sincere appreciation to the following individuals (native and non-native Croatian language speakers) for their assistance in re-annotating a subset of the Croatian Pauza dataset.

- Dina Drapić
- Dunja Dejanović
- Ivana Salnović
- Diego Alves

The work presented in this thesis has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement no. 812997 and under the name CLEOPATRA (Cross-lingual Event-centric Open Analytics Research Academy).

SUMMARY

"Opinion analysis, also called opinion recognition, is a field of research that analyses people's opinions, attitudes, evaluations, and feelings toward entities and their attributes expressed in written text." (B. Liu, 2020). An individual's decision-making process is influenced by his opinion, and his decisions in turn influence the opinions and attitudes of other people who participate in decision-making. The action of an individual is usually the result of processing information (objective or subjective) gathered through interactions with the environment. This information leads to the formation of opinions and attitudes. Every piece of information we adopt leads us to build attitudes about the phenomena that surround us. Today, the primary sources that can lead to the formation of attitudes are social networks and other internet communication, TV, books, and newspapers. These interactions lead to the sharing of opinions and attitudes with others. Since people like to share their opinions and views, with the emergence of social networks, there has been an exponential growth of user content on Twitter, Facebook and Instagram, along with websites, blogs and forums, which enable the exchange of opinions.

This area has received a lot of attention recently, and it ranges from product reviews to news analysis. The prevailing attitude or opinion about a news item is an important feature in the analysis of different biases (Mejova et al., 2014), in the dissemination of textual information (el Ali et al., 2018) and in the detection of controversy in online news (Garimella et al., 2018). Special attention is paid to news media reporting on an event.

The goal of the doctoral dissertation is to create a model of opinion analysis in languages with limited supervised data sources. Different labelled datasets of languages with rich data sources (English and Russian) were used to create the model. Since the official Slavic languages of the European Union (Bulgarian, Czech, Croatian, Polish, Slovenian, Slovak) in principle have limited supervised data sources, that language family was primary for research. Classification of opinions was carried out at the sentence level and at the document level. Different possibilities of using parallel datasets, paraphrases and labelled datasets of languages with rich computational data sources are explored. The experiments were conducted using datasets for languages from the same language family, since they are typologically related language pairs. Success was compared with a typologically unrelated language - English. The performance of all approaches was measured on the corresponding

datasets. The methods presented in this doctoral dissertation advance the achievements in the field of opinion analysis in the framework of cross-linguistic approaches.

The dissertation is divided into three parts. In the first part of the dissertation, we sought to answer the following research question: how to choose a good language model for cross-linguistic opinion analysis?

Hypothesis H1 was also put forward: Linguistic diagnostic classifiers such as those for detecting negations and paraphrases achieve high accuracy in testing the existing models of opinion transmission.

The selection of a good candidate for the model was performed using diagnostic classifiers and relied on the fact that models that can successfully detect negation and paraphrases are better at cross-linguistic opinion classification. To obtain an answer to the first research question, available multilingual linguistic models were chosen, for which two sources were needed: a set of sentences with negation and those without it, and a set of paraphrases for all languages in the research. For this purpose, a bilingual corpus of Tapoca paraphrases was used. This corpus is a collection of paraphrases in 75 languages. In the corpus, English is the source language, and all paraphrases in other languages are linked by a unique group identifier. Indicators of negation were collected from the literature on negation. All sentences corresponding to explicit indicators of negation in the corpus in English have already been extracted in the first step. A corpus of sentences with and without negation was created using the previously collected corpus. Using English sentences with and without negation as sources, corresponding pairs were generated for the other languages in the study. Paraphrases were used directly from the Tapoca dataset. Datasets were used to apply different models. The models were trained for cross-linguistic opinion analysis. The measurement of the success of opinion analysis and the cosine similarity of sentences with negation or paraphrases were compared using correlation values.

In the first part of the dissertation, we gave an overall overview of the procedure of diagnostic classification of trained language models for cross-linguistic opinion analysis. In the first phase, we tested different trained language models on simple tasks in order to check their correlation with the tasks of opinion analysis and multi-task learning. We concluded that negation has a moderate correlation with opinion comprehension in cross-linguistic opinion analysis. This led us to the conclusion that simple negation can be used to select a well-trained language model for the further task of opinion analysis. The XLM-Roberta-base model achieved better performance compared to other models.

In the second part of the dissertation, we sought to answer the following research question: what is the effect of linguistic similarity and available computer data sources in multilingual linguistic models?

Hypotheses were also put forward:

• H2: Cross-linguistic transfer is more successful in typologically similar than in typologically different languages.

• H3: A large set of labelled data from a more distant language family can help overcome typological differences, unlike a small set from a closer language family.

Languages belonging to the same language family share a subset of vocabulary and common features. Therefore, the initial assumption was that the computer data sources of those languages are more suitable for cross-linguistic transfer of knowledge. In order to obtain an answer to the second research question, i.e. to study the influence of typology on the performance of cross-lingual opinion analysis, models were trained on datasets of English and Slavic languages. Data from languages with rich data sources (English and Russian) and from the same language family were used. Results were calculated and compared with the previous results. The best hub language for the transmission of opinions has been identified. The interior of the trained model was also examined to understand the strengths and weaknesses of the model. The model that was identified in the first phase of this research as the best for feature coding was used.

In this part of the dissertation, we proposed a deep learning framework for using the existing markup of languages with rich computational resources on languages with limited computational resources. We have conducted numerous experiments on languages that belong to the same language family. We studied how well opinion classification ability can be transferred by adding data from the same language family compared to a distant language family. We have proven that our framework improves upon simple fine-tuning, considering multiple large training datasets. We concluded that the best method is to jointly train the opinion analysis system to alleviate the problem of limited resources in the target languages.

We found that the transfer of opinion knowledge improves in the case of the same language families, i.e. the closer the language, the easier the transfer of opinion if we have a large dataset. We also found that having a large dataset from a distant family when training the model can give better performance than using smaller datasets from similar languages. Quantitative experiments show that adding a large amount of data from a similar language and a language from a distant language family is beneficial for cross-language transfer of opinions.

The third part of the dissertation sought to answer the following research question: can data augmentation be effectively used for sentiment analysis in resource-poor languages?

Hypothesis H4 was also put forward: data augmentation achieves accuracy comparable to the accuracy of supervised approaches in typologically similar languages.

Augmentation techniques aim to increase the size of the training corpus in various ways. However, their application in language processing with limited data resources remains unexplored. In this section, we presented our results based on three data augmentation techniques. We experimented with WordNet and two transformer-based techniques for all languages in the study. We have proposed an additional technique that does not require the use of additional language processing tools. Furthermore, we tested different amounts of augmented data and presented opinion classification results for four Slavic languages with limited resources. Although data augmentation allows us to artificially expand the instance space for deep learning methods, using a transformer-based text encoder does not lead to a large performance improvement for the most part.

At the end, an error analysis was performed and various phenomena encountered during the evaluation process were pointed out.

SAŽETAK

Analiza mnijenja, koja se naziva i prepoznavanjem mnijenja, područje je istraživanja koje analizira mnijenja, stavove, ocjene i osjećaje ljudi prema entitetima i njihovim atributima izraženima u pisanom tekstu." (B. Liu, 2020). Na proces odlučivanja pojedinca utječu njegovo mnijenje, a njegove odluke zauzvrat utječu na mnijenje i stavove drugih osoba koje sudjeluju u donošenju odluka. Djelovanje je pojedinca obično posljedica obrade informacija (objektivnih ili subjektivnih) prikupljenih kroz interakcije s okolinom. Te informacije dovode do formiranja mnijenja i stavova. Svaka informacija koju usvojimo vodi nas u izgradnju stavova o pojavama koje nas okružuju. Danas su primarni izvori, koji mogu dovesti do formiranja stavova društvene mreže i ostale internetske komunikacije, TV, knjige, novine. Te interakcije dovode do dijeljenja mnijenja i stavova s drugima. Budući da ljudi vole dijeliti svoja mnijenja i stavove, s nastankom društvenih mreža došlo je do eksponencijalnoga rasta korisničkoga sadržaja na Twitteru, Facebooku i Instagramu, uz internetske stranice, blogove i forume, koji omogućuju razmjenu mnijenja.

Ovo je područje u posljednje vrijeme dobilo veliku pozornost, a proteže se od ocjena o proizvodima do analize vijesti. Prevladavajući stav ili mnijenje o nekoj vijesti važna je značajka pri analizi različitih pristranosti (Mejova et al., 2014), u širenju tekstnih informacija (el Ali et al., 2018) i otkrivanju prijepora u internetskim vijestima (Garimella et al., 2018). Posebna se pozornost posvećuje izvješćivanju novinskih medija o nekome događaju.

Jezici s bogatim računalnim podatkovnim izvorima, kao što je engleski, imaju veliku količinu podataka u obliku "zlatnoga standarda" za različite zadatke analize mnijenja. Međutim, tako nije s drugim jezicima koji ili nemaju nikakvih označenih podatkovnih skupova ili imaju vrlo male korpuse s označenim pozitivnim ili negativnim mnijenjem. Stvaranje takvih podatkovnih izvora skupo je i dugotrajno, a za dosljedno i sustavno označavanje mnijenja prethodno je potrebna izrada jasnih smjernica za označavanje kao i izobrazba označavatelja. Razvoj jezičnih modela za velike jezike, kao što je npr. mBERT temeljen na Transformer arhitekturi, pokazao se uspješnim načinom iskorištavanja postojećih računalnih podatkovnih izvora za zadatke kao što je analiza mnijenja, te se tako otvara mogućnost za istraživanje prekojezičnih tehnika za analizu mnijenja i u jezicima za koje postoji malo računalnih izvora podataka.

Cilj doktorske disertacije bio je stvaranje modela analize mnijenja u jezicima s ograničenim računalnim podatkovnim izvorima. Za izradu modela koristili su se različiti označeni skupovi podataka jezika s bogatim računalnim podatkovnim izvorima (engleskog i ruskog). Budući da službeni slavenski jezici Europske unije (bugarski, češki, hrvatski, poljski, slovenski, slovački) u načelu posjeduju ograničene računalne podatkovne izvore, ta je jezična porodica bila primarna za istraživanje. Klasifikacija mnijenja se provela na rečeničnoj razini i na razini dokumenta. Istražene su različite mogućnosti korištenja paralelnih skupova podataka, parafraza i označenih skupova podataka jezika s bogatim računalnim podatkovnim izvorima. Eksperimenti su se provodili koristeći skupove podataka za jezike iz iste jezične porodice, budući da se radi o tipološki srodnim jezičnim parovima. Uspješnost se usporedila s tipološki nepovezanim jezikom - engleskim. Provedeno je mjerenje uspješnosti svih pristupa na pripadajućim skupovima podataka. Metode predstavljene u ovoj doktorskoj disertaciji unaprjeđuju dostignuća u području analize mnijenja u okviru prekojezičnih pristupa.

Disertacija je podijeljena u tri dijela. U prvom dijelu disertacije tražio se odgovor na istraživačko pitanje: kako odabrati dobar jezični model za prekojezičnu analizu mnijenja?

Postavljena je i hipoteza H1: Lingvistički dijagnostički klasifikatori kao što su oni za otkrivanje negacija i parafraze postižu visoku točnost u ispitivanju postojećih modela prijenosa mnijenja.

Odabir dobrog kandidata za model izvršio se pomoću dijagnostičkih klasifikatora i oslonio na činjenicu da su modeli koji mogu uspješno detektirati negaciju i parafraze bolji u prekojezičnoj klasifikaciji mnijenja. Za dobivanje odgovora na prvo istraživačko pitanje odabrali su se dostupni višejezični lingvistički modeli, za što su bila potrebna dva izvora: skup rečenica s negacijom i onih bez nje te skup parafraza za sve jezike u istraživanju. U tu svrhu korišten je dvojezični korpus parafraza Tapoca. Ovaj korpus je zbirka parafraza na 75 jezika. U korpusu je engleski izvorišni jezik, a sve parafraze na drugim jezicima povezane su jedinstvenim grupnim identifikatorom. Pokazatelji negacije prikupili su se iz literature o negaciji. Sve rečenice koje odgovaraju eksplicitnim pokazateljima negacije u korpusu na engleskom izdvojene su već u prvom koraku. Korpus rečenica s negacijom i onih bez nje kreiran je korištenjem ranije prikupljenog korpusa. Uporabom engleskih rečenica s negacijom i onih bez nje kao izvora, generirani su odgovarajući parovi za druge jezike u istraživanju. Parafraze su korištene izravno iz skupa podataka Tapoca. Skupovi podataka koristili su se za primjenu različitih modela. Modeli su se naučili za prekojezičnu analizu mnijenja. Mjerenje uspješnosti analize mnijenja i kosinusna sličnost rečenica s negacijom odnosno parafrazama usporedili su se s pomoću korelacijskih vrijednosti.

Prekojezična analiza mnijenja ima za cilj iskoristiti postojeće resurse iz jezika s bogatim računalnim podatkovnim izvorima i poboljšati ukupnu učinkovitost klasifikacije mnijenja za jezike s ograničenim računalnim podatkovnim izvorima. Resurs iz izvornog jezika izravno utječe na performanse označavanja podataka na ciljnom jeziku. Dakle, odabirom dobrog početnog resursa možemo poboljšati konačne performanse modela. U ovom dijelu istraživanja cilj nam je bio iskoristiti negaciju kao dijagnostički klasifikator za odabir dobrog kandidata za model. U prvom koraku smo procijenili koliko učinkovito jezični model detektira negaciju. Potom smo izračunali koliko je dobra analiza mnijenja u jezičnim modelima koji dobro detektiraju negaciju. Na kraju, proveli smo višezadaćno učenje modela kako bismo obogatili najbolji model. Kvantitativni eksperimenti otkrili su da je negacija umjereni signal za ispitivanje postojećeg naučenog jezičnog modela za prekojezični prijenos mnijenja.

U prvom dijelu rada dali smo cjelokupni pregled postupka dijagnostičke klasifikacije naučenih jezičnih modela za prekojezičnu analizu mnijenja. U prvoj fazi ispitali smo različite naučene jezične modele na jednostavnim zadacima kako bismo provjerili njihovu korelaciju sa zadacima analize mnijenja te višezadaćno učenje. Zaključili smo da negacija ima umjerenu korelaciju s razumijevanjem mnijenja u prekojezičnoj analizi mnijenja. To nas je dovelo do zaključka da se jednostavna negacija može koristiti za odabir dobrog naučenog jezičnog modela za daljnji zadatak analize mnijenja. XLM-Roberta-osnovni model postigao je bolje performanse u usporedbi s drugim modelima.

U drugom dijelu disertacije tražio se odgovor na istraživačko pitanje: kakav je učinak jezične sličnosti i raspoloživih računalnih podatkovnih izvora u višejezičnim lingvističkim modelima?

Postavljene su i hipoteze:

- H2: Prekojezični prijenos je uspješniji kod tipološki sličnih nego kod tipološki različitih jezika.
- H3: Veliki skup označenih podataka iz udaljenije jezične porodice može pomoći u nadvladavanju tipoloških razlika, za razliku od malog skupa iz bliže jezične porodice.

Jezici koji pripadaju istoj jezičnoj porodici dijele podskup vokabulara i zajedničkih značajki. Stoga je polazna pretpostavka bila da su računalni podatkovni izvori tih jezika prikladniji za prekojezični prijenos znanja. Da bismo dobili odgovor na drugo istraživačko pitanje, tj. proučili utjecaj tipologije na performanse prekojezične analize mnijenja, naučili su se modeli na podatkovnim skupovima engleskoga i slavenskih jezika. Korišteni su podaci jezika s bogatim računalnim podatkovnim izvorima (engleskog i ruskog) i iste jezične porodice. Izračunala se uspješnost prema tipološkoj udaljenosti jezika i usporedila s prethodnim rezultatima. Identificiran je najbolji čvorišni jezik za prijenos mnijenja. Također se ispitala unutrašnjost naučenog modela kako bi se razumjele prednosti i nedostatci modela. Korišten je model koji je u prvoj fazi ovog istraživanja identificiran kao najbolji za kodiranje značajki.

U ovom dijelu doktorskog rada predložili smo jedinstveni okvir dubokog učenja za korištenje postojećih oznaka jezika s bogatim računalnim podatkovnim izvorima na jezicima s ograničenim računalnim podatkovnim izvorima. Proveli smo brojne eksperimente na jezicima koji pripadaju istoj jezičnoj porodici. Proučavali smo koliko se dobro može prenijeti sposobnost klasifikacije mnijenja dodavanjem podataka iz iste jezične porodice u usporedbi s udaljenom jezičnom porodicom. Dokazali smo da se naš okvir poboljšava nakon jednostavnog finog podešavanja, uzimajući u obzir višestruke velike podatkovne skupove za učenje. Zaključili smo da je najbolja metoda udruženog učenja sustava za analizu mnijenja kako bi se ublažio problem ograničenih resursa u ciljnim jezicima.

Utvrdili smo da se prijenos znanja o mnijenju poboljšava u slučaju istih jezičnih porodica, tj. što je jezik bliži lakši je prijenos mnijenja ako imamo veliki podatkovni skup. Također smo otkrili da posjedovanje velikog podatakovnog skupa iz udaljene obitelji prilikom učenja modela može dati bolje performanse od uporabe manjih podatkovnih skupova iz sličnih jezika. Kvantitativni eksperimenti pokazuju da je dodavanje velike količine podataka iz sličnog jezika i jezika iz udaljene jezične porodice korisno za prekojezični prijenos mnijenja.

U trećem dijelu disertacije tražio se odgovor na istraživačko pitanje: može li se povećanje podataka učinkovito koristiti za analizu osjećaja u jezicima sa siromašnim resursima? Postavljena je i hipoteza H4: povećanje podataka postiže točnost usporedivu s točnošću nadziranih pristupa u tipološki sličnim jezicima.

Tehnike povećanja podataka imaju za cilj povećati veličinu korpusa za učenje na razne načine. Međutim, njihova primjena u obradi jezika s ograničenim podatkovnim resursima ostaje neistražena. U ovom dijelu predstavili smo naše rezultate temeljene na trima tehnikama povećanja podataka. Eksperimentirali smo s WordNetom i dvjema tehnikama temeljenim na transformatoru za sve jezike u istraživanju. Predložili smo dodatnu tehniku koja ne zahtijeva upotrebu dodatnih alata za obradu jezika. Nadalje, testirali smo različite količine proširenih podataka i prikazali rezultate klasifikacije mnijenja za četiri slavenska jezika s ograničenim resursima. Iako nam povećanje podataka omogućuje umjetno proširenje prostora instanci za metode dubinskog učenja, korištenje tekstualnog kodera temeljenog na transformatorima većinom ne dovodi do velikog napretka u performansama.

Na kraju je izvršena analiza pogrešaka te se ukazalo na razne pojave na koje smo naišli tijekom procesa evaluacije.

ABSTRACT

In this dissertation, we develop automated deep learning models for the task of sentiment analysis for low-resource languages. These models are built using transformer neural networks. To accomplish the task of sentiment analysis, formally known as the task of calculating the polar orientation of the text that is provided, we make use of the resources that are available in high resource languages.

In this dissertation, we develop and conduct experiments on a set of low-resource and high-resource South Slavic languages.

The dissertation is divided into three sections.

1) Using a probe mechanism, we conduct experiments in the first section to select a good pre-trained model from publicly available resources. We develop a simple scoring technique to correlate the performance of sentiment analysis and probing scores. To test our hypothesis that a model is appropriate for cross-lingual sentiment transfer, we compute scores before and after fine-tuning.
2) In the second section, we conduct numerous experiments employing Slavic and non-Slavic language datasets. We also examine the effect of Cyrillic and Roman scripts on the transfer of sentiment. We combine datasets from multiple languages and determine the optimal combination technique. We also propose a framework for multi-task learning for cross-lingual sentiment analysis.
3) In the third section, we examine the effect of augmenting low-resource sentiment analysis tasks using data augmentation techniques. We conduct an experiment utilising the existing data enhancement methods and propose two novel methods. Our proposed procedures do not rely on external oversight or resources. By analysing the results, we have determined that the transformer-based fine-tuning schemes do not benefit from augmented data because it is invariant to augmented instances.

KEYWORDS

Sentiment analysis, Low-resource, Multilingual, South-Slavic, Cross-lingual, Sentiment classification, Data augmentation, Multilingual, Multi-task learning, Probing, Negation, Semantic, Cross-family, Multi-source

LIST OF ABBREVIATIONS

NLP (Natural Language Processing)

CLSA (Cross-Lingual Sentiment Analysis)

PLM (Pre-trained Language Models)

MLLM (Multi-lingual Language Models)

MT (Machine Translation)

NLI (Natural Language Inference)

BERT (Bidirectional Encoder Representations from Transformers)

BiLSTM (Bi-directional Long-short Term Memory)

TABLE OF CONTENTS

1. INTRODUCTION

Human beings are very natural at giving opinions. The phenomenon is so natural to humans that its effect can be seen in written and spoken texts in a variety of formal and informal modes of communication. Such an enormous amount of data generation necessitates automated methods for processing these streams. The NLP community has reported an exponential increase in the number of methods performing automatic opinion analysis over the past decades. However, previous research has primarily focused on languages with abundant resources, ignoring those with limited resources. As a result, low-resource language processing has begun to emerge as the default hot area of research.

"Sentiment analysis, also known as opinion mining, is the field of study that analyses the opinions, sentiments, appraisals, attitudes, and emotions expressed in the written text regarding entities and their attributes" (B. Liu, 2012). An individual's decision-making process is influenced by his opinions and attitudes, and his decision-making procedures influence the perspectives of those involved in the decision. Typically, actions result from the processing of information (facts or biased-subjective data) gathered through interaction with the environment. This data facilitates the formation of opinions. Every piece of information we gather contributes to the formation of our opinions about the objects in our immediate environment. Social media, television, books, and newspapers are the primary sources for opinion formation. This interaction results in the dissemination of opinions. People enjoy sharing their opinions, and with the advent of social media, there has been a deluge of user-generated content on Twitter, Facebook, and Instagram, not to mention news websites, blogs, and forums. This allows for an extensive exchange of opinions in the form of media-disseminated information.

In recent decades, the field of sentiment analysis has received a great deal of attention. Sentiment analysis has been applied to customer reviews of restaurants, hotels, and movies, as well as reviews of more tangible objects such as electronic devices. For example, "*Pizza čapričioša vrlo dobra, dostava kasnila 10min*" (**Translation En**: *Capriccioso pizza very good, delivery was 10 minutes late*) is a user review for an online order left by a customer. In the example, the text's intricate details about the author's thoughts and feelings can provide a great deal of insight and information about various aspects. In this situation, therefore, automatic text analysis is essential for corrective maintenance.

In a more formal setting, this could be customer feedback gathered from Human Resources, banking, and retail (de Clercq et al., 2017). This topic has received considerable research attention in the field of Digital Humanities. When analysing different biases (Mejova et al., 2014), the text's information spread (el Ali et al., 2018), and detecting controversy (Garimella et al., 2018) in online news, news sentiment is an important factor, in addition to the news media's coverage of the event.

For the numerous sentiment analysis tasks and subtasks, languages such as English contain a vast amount of gold-standard data. The same cannot be said for other languages that lack annotated data or have small sentiment corpora. The creation of data resources is an expensive and time-consuming process. To have consistent annotation, annotation guidelines must be prepared, and annotators must be trained. While established methods can be used for sentiment detection in languages with abundant resources, these methods cannot be applied to languages with limited resources, necessitating more sophisticated approaches. Recent advancements in large language models based on the Transformers architecture have demonstrated an efficient method for utilising existing resources for downstream tasks such as sentiment analysis. This opens the possibility of investigating cross-linguistic techniques for sentiment analysis in low-resource languages, i.e., languages with few computational data resources.

This thesis focuses on the cross-lingual and the mono-lingual transfer of sentiment for languages with limited resources. This study's primary objective is to improve and/or develop sentiment analysis on low-resource languages for which there are insufficient annotated resources to train supervised deep learning algorithms. Our objective is to develop techniques for performing sentiment analysis by utilising fewer resources and state-of-the-art classification models.

This chapter summarises the research problem, research objectives, hypotheses, research questions and proposed methodology. We conclude the chapter by discussing thesis organisation.

1.1 Research problem

In the simplest supervised monolingual scenario, given a collection of training examples $X = \{x_1, x_2, x_3, \dots x_N\}$ and $Y = \{y_1, y_2, y_3, \dots y_N\}$ where Y is a label of the corresponding X instance, the goal is to solve a function $f(X;\ \Theta) \rightarrow Y$ such that, given an input x_i, the function predicts y_i, and the parameter Θ is learned.

For the model to perform well on unobserved instances, N must be large enough. This is a generalised representation of text classification that fits sentiment classification perfectly. We define cross-lingual sentiment classification as follows, assuming M << N:

Let $X_{source} = \{x_1, x_2, x_3, \dots x_N\}$ represent the training instances from a language with abundant resources, and $X_{t\arg et} = \{x_1, x_2, x_3, \dots x_M\}$ represent the training instances from a language with limited resources. The ultimate goal is to construct a model Θ such that $f\left(h\left(X_{source}, X_{t\arg et}\right);\ \Theta\right) \rightarrow Y_{t\arg et}$ such that the model learns to classify sentiments in the target language. Here, $h(X)$ is a function that makes use of source language resources and facilitates learning. The most straightforward illustration of $h(X)$ could be utilising the source instances without modification. To convert source data to the target language, a more sophisticated method could employ machine translation.

1.2 Research questions

The objective is to develop a model for sentiment analysis in EU-official Slavic languages with limited resources. With this as the ultimate objective, we posed the following research questions that will be revisited in subsequent chapters.

1. How can we select a good language model for cross-lingual sentiment analysis?

● Linguistic diagnostic classifiers, such as those for detecting negation and paraphrase, probe an existing model for sentiment transfer with high precision.

2. What effect do language similarity and the availability of resources have on MLLM (Multilingual Large Language Models)?

● Knowledge transfer between typologically similar languages is more successful than between typologically dissimilar languages. A large annotated dataset in a language from a distant family can overcome typological differences, in contrast to a small annotated dataset in a language from a close family.

3. Can data augmentation be utilised effectively for sentiment analysis in low-resource languages?

● The accuracy of the data augmentation technique is comparable to that of supervised methods in typologically similar languages.

1.3 Assumptions

Aspects considered for the overall study are described in the following section.

1.3.1 Defining low-resource languages

Hedderich et al. (2021) suggested three distinct dimensions to classify a typical circumstance with limited resources. The first dimension is the lack of task-specific data. The absence of unlabelled or domain-specific corpora is the second dimension. The third one is the unavailability of resources associated with supplementary tasks. For the objectives of this dissertation, we define a low-resource language as one for which there are insufficient monolingual or bilingual corpora or resources for developing statistical NLP applications. Our primary focus is on EU-official South Slavic languages. All official South-Slavic EU languages have few resources, except for Czech and Polish, which have more publicly available datasets for the sentiment task.

1.3.2 Parallel data

Even though we have moderate parallel data with the English language for the languages under study, the literature indicates that machine translation is not yet capable of handling, preserving, and translating the sentence's semantics, at least in a language as complex as Serbian (Lohar et al., 2019). As a result, the words selected by the MT (Machine Translation) system do not accurately convey the original meaning and are therefore incorrect. As a result, with the exception of probing experiments, we do not use parallel data or a machine translation system to train a sentiment analysis system, but would like to investigate it further in the future.

1.3.3 Document-level sentiment analysis

Text sentiment analysis is typically performed at three levels: document, sentence, and aspect (B. Liu, 2020). At the document level, the objective is to determine whether an entire opinion document expresses a positive or negative sentiment (Pang & Lee, 2008; Turney, 2002). The document-level sentiment analysis implicitly assumes that the entire document expresses an opinion about a single entity and does not apply to documents that signal views about multiple entities, as noted by B. Liu (2012). In such instances, additional processing is necessary. Our research focuses on the document level. This topic will be discussed in greater detail in Chapter 2.

1.4 Research proposal

Our target languages are South-Slavic languages with very few labelled examples for sentiment analysis tasks. Languages within the same family typically share a subset of vocabulary and typological characteristics. Cognates (Crystal, 2011), which are sets of words in different languages that have been directly inherited from an etymological ancestor in a common parent language, are one such phenomenon. For instance, the Proto Slavic word **noktь** (*night*) has equivalents in other languages such as **ночь** (*nočʹ*) (Russian), **ніч** (*nič*) (Ukrainian), **ноч** (*noč*) (Belarussian), **noc** (Polish, Czech, Slovak), **noč** (Slovene), **ноћ/noć** (Serbo-Croatian), **нощ** (*nosht*) (Bulgarian), **ноќ** (*noḱ*) (Macedonian). Recent Transformer-based language models that have demonstrated efficacy in supervised downstream tasks are prime candidates for low-resource NLP. A study (Chi, Hewitt, and Manning, 2020) has shown that Multi-lingual BERT exhibits cross-lingual clustering that is largely consistent with UD (Universal Dependencies) dependency labels in English and French. We believe that resources from the same family languages are better suited for cross-lingual knowledge transfer due to the aforementioned factors.

With the task of cross-lingual sentiment classification for low-resource settings in sight, this thesis investigates three distinct problem areas. There are numerous PLM options available for a given task. They vary in size and specifications. They can vary based on (1) the languages used during training, i.e., monolingual or multilingual (bilingual, trilingual, etc.), (2) the number of network parameters, and (3) the modalities (text plus image or video), to name a few.

1.4.1 Probing language models

The first problem we face is selecting the PLM (Pre-trained Language Models) that will work best for sentiment analysis (SA) tasks in low-resource environments. We believe that models that initially perform moderately well on a specific NLP task will improve once exposed to additional learning with essential data. Important SA tasks include semantic textual similarity in monolingual contexts (paraphrase detection) and bilingual contexts (bitext detection). This leads us to our first hypothesis, which states that PLMs that perform well in detecting negation and paraphrasing are superior at cross-lingual sentiment classification. The configuration proposes combining datasets from the negation, bitext, and paraphrasing tasks to score a PLM. To obtain this score, we created new manually annotated datasets for the language for which we were unable to obtain it directly and repurposed datasets for other languages. We utilised existing resources for the tasks of paraphrasing, Natural Language Inference, and translation. All probing tasks utilised cosine similarity to evaluate the model. This was done before and after fine-tuning the language model using sentiment datasets in each language separately. In addition, the models were trained in each language using data from three different probing tasks in a multi-tasking fashion. Each of these enriched models was subsequently utilised in the sentiment analysis phase of fine-tuning. According to the experimental findings, there is a moderate correlation between the cosine of the negation task and sentiment classification scores. For upcoming experiments, a suitable PLM was chosen based on empirical findings. We also find that the correlation between the bitext and paraphrase similarity scores and the sentiment analysis score is weak.

1.4.2 Cross-lingual sentiment analysis – same family vs distant family

In the absence of sufficient labelled instances, joint training is an alternative method for training classifiers that combines data from multiple sources. In this configuration, we combined resources from multiple languages in their original distribution for joint training. We utilised both high-resource distant family languages and same family languages to examine the impact on final performance. In addition, we proposed a framework for treating multiple labels of the same dataset as distinct tasks. In conclusion, we demonstrated the efficacy of the MTL (Multi-task Learning) by comparing it to a non-MTL version. We discovered that the transfer of sentiment knowledge is enhanced between languages of the same family, i.e., the larger the dataset, the easier it is to transfer sentiment knowledge from one language to another. We also discovered that a large training dataset from a distant language family can outperform smaller datasets from similar languages. Consequently, datasets from the same language family as well as those from distant language families can be utilised to combat the problem of data scarcity.

1.4.3 Data augmentation for sentiment analysis in low-resource settings

Data augmentation is a technique for increasing the number of training examples (Simard et al., 2012). This can serve as a viable replacement for additional manual data or data from other family languages in environments with limited resources.

For each language, WordNet, the Masked Language Model, and the Causal Language Model are used to supplement the data. In addition, a simple technique based on permutation and combination was proposed for expanding data without additional resources. The method's rationale is predicated on the hypothesis that every sentence within a positive review is also positive, and vice versa for negative reviews. Thus, it is possible to generate a new training instance using sentences from completely polar classes. We trained a sentiment classifier using each of the enumerated techniques with training sets of varying sizes. Using augmented data with a Transformer-based encoder does not result in significant gains, as demonstrated by the empirical validation of the hypothesis and the experimental results.

1.5 Thesis outline

The remainder of the thesis is structured as follows:

The **second chapter** examines the background of sentiment classification, text classification, and multi-task learning, as well as sentiment-related concepts.
In section 2.1, we concentrate on fundamental and related sentiment analysis concepts.
In section 2.2, we analyse in detail the previously presented approaches for sentiment analysis in monolingual, cross-lingual, and multilingual contexts. For each method, we segregate, aggregate, and classify the pros and cons of each reported method for English and South-Slavic languages. In section 2.3, we examine the data requirements and data availability of a variety of low-resource languages. In the final section, we discuss the fundamentals of the neural techniques utilised in the thesis.

Chapter three introduces the issue of selecting a candidate language encoder from a variety of alternatives. In this section, we discuss how various smaller datasets can be used to assess the sentiment capabilities of an existing PLM. Then, we evaluate multiple PLMs and contrast and correlate their performance with a straightforward classification method.

Chapter four examines our datasets for sentiment classification across languages. In this chapter, we demonstrate the effectiveness of the MTL setup across all languages by experimenting with the use of resources from distant and same-family language families.

Chapter five assesses the data augmentation techniques for low-resource languages. We propose a simple data augmentation technique inspired by combinatorics and compare the results to other prominent DA techniques.

Chapter Six summarises the thesis' findings and discusses potential future research directions.

2. RELATED WORK

With more than 2,760,000 Google scholar hits as of today[1], sentiment analysis continues to be one of the most researched topics. The subject has been studied both independently and in conjunction with other disciplines. The advancements in machine learning, data mining, and deep learning have had a significant positive impact on the investigation of sentiment analysis. In exchange, the emphasis on text classification tasks is driving the development of more sophisticated methods. The accessibility of user comments and opinions on public Internet domains, including social media, has also been a significant contributor.

This chapter provides an overview of the field of sentiment analysis. The essential experimental terminology and associated processes are discussed first. Following this, we list various prior works in the various subject areas of sentiment analysis. Finally, we present research-relevant work that is relevant to the field.

2.1 Background

2.1.1 Primary definitions

Subjectivity is not susceptible to evaluation and verification, whereas objectivity is. This is because subjective statements are composed of an individual's experiences, beliefs, and emotions. This is subjective to the subject in the truest sense. While tasks such as Information Retrieval and Topic Modelling have dealt with objective statements, the development of human-like subjectivity analysis has led to the development of multiple tasks, each of which solves a specific problem.

Most work in SA has focused on classifying the text's polarity as positive or negative (or neutral). In the context of sentiment analysis, the terms "emotion" and "opinion" have been used interchangeably. Affect, feeling, and emotion are additional synonymous terms found in the context of classifying sentiments, with subtle distinctions between them. All these terms refer to distinct phenomena with intricate distinctions, but they have been used interchangeably due to improper nomenclature and inconsistent usage (Munezero et al., 2014).

[1]Google Scholar, *Google Scholar* [website], (accessed 18 July 2022)

Affect is defined as positive and negative evaluations of an object, behaviour, or concept, accompanied by intensity and activity dimensions (Thoits, 1989; Shouse, 2005). Affect is "the predecessor to feelings and emotions" (Munezero et al., 2014). **Feelings** are expressions of affect. Emotions (Thoits, 1989) are defined as culturally determined feelings or affects. Feelings are the result of past experiences and are unique to each individual. **Emotions** are culturally/socially constrained expressions of affect (Calvo & D'Mello, 2010). **Sentiments** are partly social constructs of emotions that develop over time and are enduring. The duration of emotions and sentiments is experienced differently. Moreover, unlike sentiment, emotions may not necessitate an object of focus. **Opinions** are personal interpretations of information that may or may not be emotionally charged. Due to their close relationship, sentiments are most frequently substituted for opinion (S.-M. Kim & Hovy, 2004). The term *opinion* (B. Liu et al., 2010) is mathematically defined as the quintuple $<o; f; s_o; h; t>$, where *o* is an object; *f* is a feature of the object *o;* so is the orientation or polarity of the opinion on feature *f* of object *o* (positive, negative, or neutral); *h* is an opinion holder; *t* is the time when the opinion is expressed.

In the field of natural language processing, sentiment analysis has primarily been associated with categorising text into binary or ternary polarities, such as positive, negative, and neutral. Another classification system employs a numeric scale ranging from -1 (negative) to 1 (positive). Motivated by affective computing (Picard & Healey, 1997), texts have also been marked with positive or negative valence and arousal/intensity. In addition to positive, negative, and neutral (no-sentiment) labels, annotation scheme introducing mixed-class labels have also been proposed (Mohammad, 2016).

2.1.2 Sentiment classification of documents

Sentiment analysis can be applied at the level of the word, the sentence, and the document. This research focuses on the document level. "Document sentiment classification assumes that the opinion document *d* (for example, a product review) expresses opinions about a single entity *e* and contains opinions from a single opinion holder *h*." (B. Liu et al., 2010).
As a direct result of this assumption, the classification becomes restrictive as the opinion is tied to a single entity, which may or may not be true; for instance, "*The food is delicious, but the delivery was late.*" We also observe that the majority of publications do not use neutral class to simplify the modelling task.

Analysing the sentiment of text from diverse domains and writing styles requires caution. Texts from various domains have a distinguishing feature that is unique to that domain. As the tweets are brief and devoid of context, Twitter data can be treated as a distinct type of information. They frequently contain sarcasm and irony and are often concise, employing contractions, emoticons, and informal language usage. Twitter bots and fake accounts contribute to this. Regarding services associated with the product, product reviews contain coherent language. Movie and book reviews, on the other hand, typically include a variety of components. In film reviews, the author may discuss scenes, characters, and personnel including the director and screenwriter. These reviews provide additional context and premise.

2.2 State of the art

2.2.1 Monolingual sentiment analysis

Knowledge-based, machine-learning-based, and hybrid-based approaches to sentiment analysis can be distinguished. The knowledge-based approaches are lexicon-based and can be subdivided further into dictionary-based and corpus-based approaches. The knowledge-based/lexicon-based techniques utilise a compiled list of emotion terms. Unsupervised sentiment analysis (Paltoglou & Thelwall, 2012) is another name for the process of implementing a sentiment classification system by making use of an existing polarity lexicon. Existing polarity lexicons for English include SentiWordNet (Esuli & Sebastiani, 2006), which tags WordNet synsets with positivity and negativity scores, WordNet Affect List (Strapparava & Valitutti, 2004), which tags WN synsets with emotions, and others. This is accomplished by examining the lexicon for the polarity of the individual words and aggregating the parts to obtain the final score. When utilising multiple lexicons, lexeme size is increased via synonymy and antonymy relations. However, the approach has flaws (Das & Bandyopadhyay, 2011). 1) There is no context information (Pang et al., 2002) captured, 2) There are no domain knowledge associations (Aue & Gamon, 2005), 3) There is no information about time (Read, 2005), and there are no language/culture properties (Strapparava & Ozbal, 2010; Wiebe & Mihalcea, 2006).

In corpus-based approaches, a seed list is employed to tag a corpus with initial points. Using a similarity metric, new candidate words are searched for, and the word list is expanded. Tags can then be assigned using rule-based/semantic or statistical techniques.

Volkova et al. (2013) utilised the process of bootstrapping lexicons to tag social media text iteratively. Using a simple check for the presence of lexemes in the text, a lexicon was used to identify subjective and objective statements. Using the same lexicon, the text is classified into subjective classifications based on the number of positive and negative terms. The tagged tweets are used to calculate the probability that an unknown word is positive or negative. The newly tagged, non-lexicon-present words are added to the list. The procedure repeats until no new words remain. In a similar work by Banea et al. (2008), instead of PMI (Pointwise Mutual Information), the authors used LSI (Latent Semantic Analysis) to rank the candidate list of words with the original seeds. The threshold for filtering candidates from the ranked list is determined empirically, and a lexicon-based rule-based classifier is developed.
Although all these methods are simple to implement and do not require sophisticated text processing tools, the overall process can be enhanced by incorporating Part of Speech and lemmatisation features to reduce false positives.

Pang et al. (2002) classified movie reviews using a variety of features, such as unigrams, bigrams, and part-of-speech tags, as well as combinations of these and other features. The authors reported that the unigram features on SVM were the most effective machine learning method. Cui et al. (2006) pointed out the drawbacks of previous works employing a small amount of data and the inefficiency resulting from the use of more n-grams. The authors found no statistically significant gain when they correlated the performance scores with the top 50k, 100k, and 200k n-gram features selected by chi-square scores.

Wilson et al. (2005) suggested a two-step procedure for determining the prior and posterior/contextual polarity of phrases. The authors utilised a two-step methodology. First, lexicons were used to classify the phrase as neutral or polar, and then it was classified as positive, negative, or neutral. To compute the test score, the MPQA dataset was enriched with additional subjective expression annotation layers. An agreement study revealed that inter-rater agreement was 82%.

Among the most popular SA features are term-frequency and TF-IDF, Part of Speech, and sentiment shifters (e.g., negation, intensifiers). Mejova and Srinivasan (2011) demonstrated that a classifier trained with small features ranked by mutual information outperforms one trained with all features. This suggests that feature selection should follow feature engineering. The author tested stemming, term frequency, binary weighting, negation-enriched features, n-grams, and phrases-based features.

Word embeddings are another set of feature learning techniques that have been extensively studied in the domain of text classification. In these methods, each vocabulary term is assigned a vector in hyperspace so that words with similar meanings are grouped together. The vectors are learned by using a large corpus to train a neural network. The network is trained to predict a word given a small window of a predetermined size or to predict the context given a single word. CBOW and Skip-gram are two of the earliest methods for generating word embeddings.

Word2Vec (Mikolov, Sutskever, et al., 2013) is a 2-layer, shallow neural network that uses individual words as its vocabulary. It has been implemented using CBOW and SG techniques, the latter of which performs better with large datasets. Glove (Pennington et al., 2014) is a technique for word embedding based on the co-occurrence matrix of words within a corpus. Fast Text is a technique that uses CBOW and SG with the sum of character n-grams of a given word to compute un/known words in a given language. All of the aforementioned word embedding techniques assume a linear relationship between two words and train the model using linear classifiers. Recently proposed ELMO (Peters et al., 2018a) improves word representation through the use of two bidirectional LSTM as pre-trained neural language models that represent words as a function of the entire input sequence. The method provides embeddings based on context and has proven successful at capturing meaning, particularly in cases of polysemy.

The models that use deep learning to solve problems can be roughly categorised based on their architecture.

Convolutional Neural Networks (CNN)

CNNs were used for text classification by Y. Kim (2014a). The author hypothesised that CNNs, like those used for image classification, can be combined with word embeddings to learn text classification features. Each sentence of length n was represented by a vector of length *R*. The input was fed to a simple CNN layer with multiple filters, followed by a max-pooling layer and a fully connected classification layer with softmax output. It was discovered that fine-tuning static word embedding enhances performance.

Recursive Neural Networks

Recursive neural networks are a set of networks that discover a relational representation of the input text. The relational representation is a directed acyclic graph, specifically a tree data structure. A recursive neural network uses word embedding and relational information provided as a parse tree to recursively learn parent representations using a bottom-up strategy. For recursive input processing, the same weights are utilised. Consequently, the tokens are combined to create phrases, which are then combined to form a sentence. The representations can then be used as input for a classifier. Since this network processes phrases, we can provide each parent node with sentiment information via the softmax layer. A requirement for training this network is a tree-structured dataset with appropriately labelled nodes. Consequently, Stanford Sentiment Treebank (Socher et al., 2013) was developed to train and comprehend these networks.

Recurrent Neural Networks

Recurrent neural networks (RNN) process sequences, where the elements are indexed by time (or, in case of language, by sentence position). The network receives the input as a sequence of elements. When introduced into the network, a single element stores information in its internal states. The subsequent input is processed with both the current element data and the previously stored hidden states. Thus, the output at any given time depends not only on the current input but also on previous inputs. The following are several RNN variants that have been widely employed in sentiment classification tasks.

Long Short-Term Memory (LSTM)

Q. Huang et al. (2017) experimented with CNN and LSTM (Hochreiter & Schmidhuber, 1997) and proposed their combination for sentiment classification, combining context-dependent and global features. A single convolutional layer is followed by two LSTM layers in the architecture. The CNN layer is applied with a window to produce n-gram features. Multiple feature maps are generated by the layer and fed into the LSTM layer. The features of the second LSTM layer are fed to a sigmoid layer for classification. The authors noted that CNN or LSTM alone cannot achieve the desired results and that CNN-LSTM configuration requires two layers of LSTM rather than a single layer. A study by Hassan & Mahmood (2017) demonstrated that a single convolution layer with LSTM as the pooling layer can achieve good results with improved hyperparameters. This is because CNNs are better at extracting local features, whereas LSTMs capture long-term sentence dependencies.

Gated Recurrent Units (GRU)

Tang et al. (2015) presented a method in which CNN/LSTM was used to model every word in a sentence to obtain sentence representation. The sentence representations were fed into a bidirectional Gated Recurrent unit to generate document representations. As features, the convolution layer extracted unigrams, bigrams, and trigrams from the text.

T. Chen et al. (2017a) classified sentences within a text document containing reviewes into non-target, one-target, and multiple-target sentences by extracting target expressions using a BiLSTM-CRF model. Using a 1d-CNN, the final sentiment class of the sentences was determined.

Several studies have focused on the attention mechanism in the context of sentiment classification. RNNs are known to extract a great deal of information from the text provided as input. The attention mechanism attempts to concentrate on relevant portions of a text rather than the entire input.

T. Chen et al. (2017b) proposed a Feature-enhanced Multiview Co-Attention Network for Sentiment analysis by using POS and word position features for learning word embeddings and separate LSTM networks for modelling sentiment words, target words, and the context. The CNN network is layered on top of word embeddings in order to obtain features that are passed to the LSTM network. Additionally, multi-view attention is constructed in order to discover attention matrices for each of the three types of words. The resulting matrices are combined with embedding representation to compute the final representation, which is then passed to a softmax classifier for classification.

Yuan et al. (2018) selected domain-discriminative features using a Domain Adaptation Module that exploits domain classification to obtain a document-level context vector. These features were utilised in the attention mechanism alongside a sentiment classification module to construct a multi-task multi-domain classification model.

Basiri et al. (2021) demonstrated that an Attention-based Bi-CNN RNN model could improve the feature extraction process during network training. Using a word-embedding and two independent Bi-LSTM and Bi-GRU branches, context-sensitive features were extracted. Utilising CNN layers with global and average pooling, dimensionality reduction is carried out. The combined features are then passed to a fully connected layer for classification.

2.2.2 Cross-lingual sentiment analysis

Cross-lingual techniques aim to diminish the language gap between the source and target languages. This is accomplished by mapping the source language to the target language. Machine Translation is one of the most desired methods for achieving this goal. The second class of methods, known as representation learning, aims to discover common feature representations across multiple languages. Bilingual word embedding is an example of such a method.

In an effort to develop a subjectivity classifier, Mihalcea & Banea (2007) translated the source language lexicon into the target language and used the resulting target language lexicon to construct a classifier. The source lexicon was translated with the aid of bilingual dictionaries. The authors reported problems such as the loss of subjective meaning during lemmatisation and translation, as well as the lack of information regarding word sense. The resultant lexicon was used to develop a rule-based classifier that utilised heuristics based on the absence/presence of subjective clues. In the second approach, the English-Romanian parallel corpus was labelled using automatic source-language tools. The projected labels are subsequently utilised to train a Naïve Bayes model.

Banea (2008) experimented with English, Spanish, and Romanian as target languages for the subjectivity analysis task. The study consisted of four experiments. In the initial experiment, source language training data is translated by an MT system and then used to train a Machine Learning model. In the second experiment, it was assumed that there is no annotated corpus for the source language, but that a tool for annotating the raw source text is available. The corpus was initially annotated using the tool, then machine translated into the target language, followed by training with an algorithm for machine learning. In the third experiment, the raw text in the target language is translated into the source language, followed by the application of an annotation tool to label the translated text. The labels are projected back onto the text in the target language, and the corpus is used to train a subjectivity classifier. In the concluding experiment, the authors reported a scenario in which the target text is translated into the source language and then annotated using an annotation tool. The final experiment evaluated the resources generated in the target language during the preceding three experiments.

A co-training strategy presented in Wan (2009) utilised a multi-view representation for English and Chinese review classification. Utilising a machine translation module, the method converts English-labelled reviews into Chinese. There exists an additional set of unlabelled Chinese reviews that, along with English reviews translated into Chinese, comprise the Chinese perspective. The second view includes both labelled English reviews and unlabelled Chinese reviews that have been translated into English. A separate SVM classifier is trained for each language and used to predict unlabelled reviews. Taken from both languages, the intersection of the most confidently predicted reviews is added to the set of labelled training reviews. Even during the prediction phase, the method relies heavily on machine translation and requires input in both languages.

A co-regression algorithm for cross-lingual rating prediction was proposed by Wan (2013). The authors utilised machine-translated source texts represented by term frequency as features and an SVM linear kernel regressor as an algorithm for machine learning. The training setup is similar to that of Wan (2009), but the classification task has been replaced with regression.

Zhou et al. (2016a) introduced a hierarchical attention mechanism in the LSTM network in order to capture long-term dependencies in the texts. In the configuration, machine translation was utilised to generate parallel documents. The overall neural network was comprised of word-embeddings, bidirectional LSTMs, and hierarchical attention mechanisms for words and sentences so that the network can learn to focus on sentiment-bearing sentences in the document and polar words in the sentence. In addition to classification loss, an additional Euclidean loss was incorporated to align parallel sentences.

Q. Chen et al. (2015) refuted the claims made in Duh et al. (2011) that MT is ready for CLSA, as Q. Chen et al. (2015) demonstrated that the sentiment polarity of the translated text differed from that of the original text. This is a consequence of the noise introduced by the MT system. In addition, Q. Chen et al. (2015) addressed the issue of filtering out noisy knowledge introduced by incorrect translations and incorrectly classified instances of source language classifiers. Similar to Wan (2009), the experimental setup trains a single classifier using an additional knowledge validation function. A classifier is initially trained by identifying, validating, and recommending knowledge. Source and train data are updated for subsequent iterations utilising pseudo-parallel data and the validation function for target language knowledge.

Zhou et al. (2016b) reported extending the paragraph vector model and employing it to jointly discover a bilingual embedding space. An additional constraint is imposed to place polar documents on opposite sides of the hyperplane. The documents in the source language are translated into the target language using machine translation mode. The documents in each language, along with their sentiment labels, are trained for classification loss, bilingual Euclidean distance loss, and a loss function that makes document vectors with the same sentiment class closer. Upon completion of the representation learning step, the features are used to train a logistic regression classifier.

S. M. Mohammad et al. (2016) showed that automatic translation of texts and lexicons does improve performance while working on a sentiment analyser for Arabic social media posts. The study compared existing SA techniques in various settings. Furthermore, the authors confirm that the translated text does cause label shift (polar to neutral). Additionally, the change in polarity is caused by poor translation quality. In the study that was conducted, an Arabic text was translated into English. The original and the translated text were manually annotated and compared for concordance. A similar analysis was conducted on the lexicons of both languages. The authors observed that MT errors influence human judgement, and that the sentiment analysis of a machine-translated text is less prone to error. In addition to cultural bias playing a role in the annotation, various linguistic phenomena such as word-reordering, sarcasm, and metaphoric experimentation are common causes of misclassification.

For the subjectivity classification task, Nandi et al. (2021) conducted experiments with state-of-the-art AdaSent (Zhao et al., 2015), context-independent (Word2Vec, Glove) and context-dependent (ELMO, BERT-Base) embedding models with an LSTM layer. BERT's capacity to capture bidirectional content information allowed it to approach AdaSent’s efficiency while outperforming other techniques.

Xu et al. (2010) proposed an extension to the AdaBoost algorithm that employs a re-weighting strategy to learn the source and target language data jointly. In this method, if a source language instance is incorrectly classified, it is given less weight because it cannot be used for learning the target language. Alternatively, if a target language instance is misclassified, it is given greater weight. The training is conducted on both the source and target datasets, but the error is only calculated for the target language. To prevent the loss of source instances due to the model's inability to correctly classify instances early on, the authors proposed a weighting scheme that reduces early discarding.

2.2.3 Transfer-learning approaches

Transfer learning is one of the AI learning regimes in which a previously trained model is used to solve a different problem. In zero-shot, the source model is directly applied to the target problem, while additional data is used to fine-tune the model. Transfer learning has previously been used in sentiment analysis to solve the problem of domain adaptation, which involves knowledge transfer between source and target domains. Using a three-part linear setup, Ganin et al. (2016) proposed a domain-adversarial neural network (DANN) to solve the problem of domain knowledge transfer. The authors suggested an architecture consisting of a feature extractor, label predictor, and domain discriminator. The training is conducted to minimise label classification loss and maximise domain classification loss, enabling domain-independent feature extraction.

Meng et al. (2019) suggested training two multi-layer CNNs by distributing weights between the source and target domains. Each network has its own classification heads for the source and destination domains, but all networks share the same backbone. First, a convolution-Relu-max pooling network with an embedding layer derived from Word2Vec is trained using a source language dataset. A target domain-specific classification layer of the network is fine-tuned with a small amount of data from the target domain in the second stage. The authors conducted experiments with data samples ranging between 200 and 4,000. The method proved superior to other machine learning and domain adaptation techniques due to its simplicity and the absence of pivot queries.

Gupta et al. (2021) compared the outcomes of task-specific pre-training for code-switched sentiment analysis. The authors emphasise the significance of target language presence during the pre-training phase of the contextualised model. The author conducted task-specific pre-training with the source language, followed by fine-tuning with the target language dataset.

In a monolingual environment, a single classifier functions for a single language, necessitating a separate classification model for each language. As reported in Xu and Wan, (2017), a classifier comprising labelled data in English and unlabelled parallel data in a few language pairs was developed. The process was predicated on the premise that sentiment transfer can be accomplished using pivot languages. The model learns sentiment-aware word embeddings from parallel data to ensure that similar words in different languages have identical representations. For the model to utilise parallel pivot language data, additional constraints are imposed. In addition to the size of parallel data, quality and genres play an important role in the system's final performance, as noted by the authors.

Krchnavy and Simko (2017) conducted experiments on Slovak using LEX (lexicon-based), Support Vector Machine (SVM), Naïve Bayes (NB), and Maximum Entropy (ME) with four pre-processing parameters involving emoticons, diacritics, lemmatisation, and negation. The authors reported the existence of the double negation ("*ne*") phenomenon and identified negation detection as a crucial task in Slovak. All pre-processing operations yielded the best results for the ME. In the case of the LEX approach, special negation processing does not improve performance, according to the authors.

In Polish, Bartusiak et al., (2015) presented a straightforward method for employing unigram and bigram features for cross-domain transfer learning. The authors trained an SVM model using both unigrams and long words. The dictionary resulting from vectorisation is used in conjunction with the trained model to classify sentiment labels for data from another language in Polish.

Inspired by Zhang & LeCun (2015), Mršić et al. (2017) compared the performance of sentiment classification in Croatian and English using a deep convolutional network. The author conducted experiments with various activation functions and concluded that the sigmoid function produced the highest test scores.

In cross-lingual settings, Přibáň and Steinberger (2022) utilised English and Czech datasets along with multilingual BERT and XLM-R-Large. According to the authors, larger multilingual language models are superior to smaller monolingual language models. When the source and target languages were combined, performance in the target language declined. Furthermore, a smaller manually annotated dataset is preferable to a large automatically tagged dataset for cross-lingual studies.

Robnik-Šikonja et al. (2021a) conducted experiments on a Twitter sentiment dataset in 13 languages using two transfer learning mechanisms. The first approach employed a word embedding constructed from parallel or comparable corpora. Due to very low self-agreement and low inter-annotator agreement, the authors noted that the annotations in the dataset are of poor quality. The factor of agreement (self or inter-annotator) on annotations is extremely important, as it has been empirically demonstrated that combining datasets from the same language family with conflicting annotations leads to a performance decrease. Using datasets from different language families for joint training led to a performance decline. The study also discussed the number of instances in the target language dataset used in the joint training, which, if appropriate, does not require additional training instances, while the addition of other language datasets will decrease performance.

2.3 Deep learning

The fundamental element of a neural network is the perceptron (Rosenblatt, 1958). The perceptron accepts multiple inputs. The weights are multiplied by the input to produce a value that is then added to produce the final output. If the value meets the threshold and the condition, the neuron is activated and sends 1 as output; otherwise, it sends 0. Due to the inability to model certain functions, multilayer perceptrons consisting of an input layer, a hidden layer, and an output layer were developed.

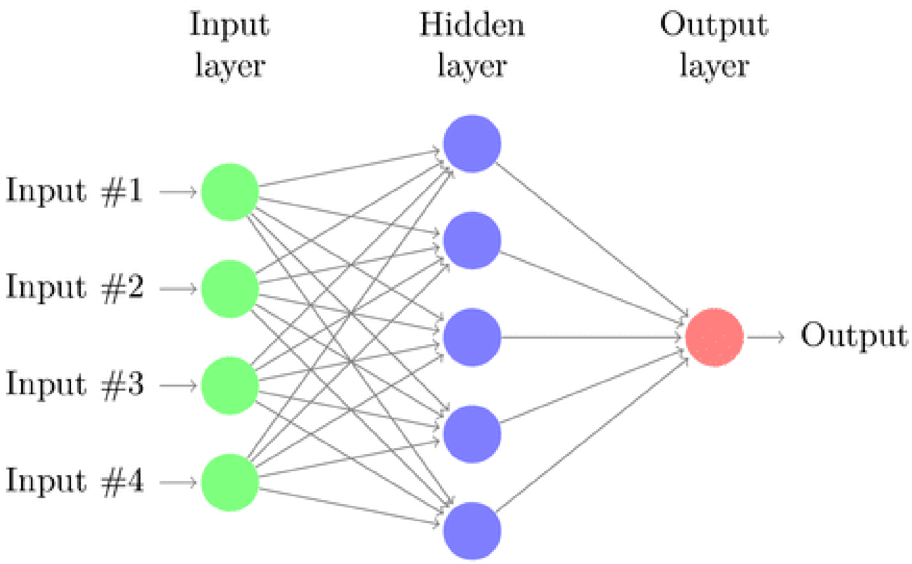

Figure 2.1 Depiction of a feed-forward neural network.

The following equation characterises a neuron:

$$z = \sum_{i}^{N} x_i w_i + b \qquad 2.1$$

$$y = F(z) = \sum_{i}^{N} x_i w_i + b \qquad 2.2$$

The output decision in the above equation is determined by the weighted sum and bias term. Thus, a function is introduced to implement a decision-making functionality. Here the function takes in the weighted sum and bias and decides whether to activate the neuron or not.

$$y = 1,\ if\ F(z) \geq 0\ else\ y = 0$$

A rectified linear unit (ReLu) activates the neuron using the following function.

$$y = z,\ if\ z > 0\ else\ 0$$

A sigmoid function is defined as:

$$y = \frac{1}{1 + e^{-z}}$$

2.3

A hyperbolic tangent function is defined as:

$$y = \frac{e^z - e^{-z}}{e^z + e^{-z}}$$

2.4

A softmax function is defined as:

$$y = \frac{e^z}{\sum_{i=1}^{K} e^z}$$

2.5

Commonly, the softmax function is used in the output layer for classification because it transforms unnormalised probability, or logits, into a natural probability distribution for a given instance over the set of labels. In a supervised setting, neural networks are trained using a collection of (*x*,*y*) pairs, where *x* is the input and *y* is the class label. The training step consists of parts i.e., a forward and reverse pass. During the forward pass, the input *x* is fed to the network's input layer and a prediction $y^\wedge$ is computed using the final layer. $y^\wedge$ and Y are utilised in the loss computation $H(y, y^\wedge)$. The loss term is used to compute the derivative of each weight with respect to the input during the backward pass. The backpropagation algorithm is then employed to adjust the weights. The objective is to minimise the value of the loss term by determining the optimal weights given the input. Passing the entire dataset through the network once signifies a single epoch. Although the entire dataset can be passed and weight update can be deferred until the end of the epoch, min-batching is typically employed to compute the error and update the network value for each batch of *n* examples. As updating a network for a single value is not the optimal strategy, batch-based training provides a stable training setup.

The term $H(y, y^\wedge)$ is the loss term and is calculated for classification using cross entropy as follows:

$$H(y, y^\wedge) \; = \; -y \log y^\wedge \; - (1 - y) \log(1 - y^\wedge)$$

2.6

During the training phase, overfitting of the network to the training samples is another phenomenon observed. Overfitting is the phenomenon that occurs in data modelling when a function aligns too closely with a minimal set of data points. As a result, the model performs exceptionally well on the training data but cannot generalise to new data. Such a model memorises the dataset and is flawed by design. Regularisation is used to solve this issue by penalising models that attempt to overfit network weights. ***L2*** is a frequent term for regularisation used during training. The ***L2*** term is the sum of squares of all the weights of a model. When computing loss, the sum of the squared norms from the model weights are added to the error term. The lambda is a hyperparameter that controls the loss as well the weights assigned to the model. If the lambda is large, then the weights of the network will be closer to zero as larger model weights values will lead to larger loss and in turn leading the model to opt for smaller values.

$$Loss = Error(y, y^{\wedge}) + \lambda \sum_{i=1}^{N} w_i^2 \tag{2.7}$$

2.3.1 Multi-task learning

A single-task learning setup is exemplified by a neural network trained to perform a single task. Multi-Task Learning aims to solve two or more tasks using information shared across multiple layers. Given a smaller number of data instances and the network capacity defined by the number of hidden units, multi-task learning utilises the data from multiple tasks not only to learn the parameters useful for all the multiple tasks, but also to prevent overfitting during training. MTL has demonstrated success in natural language processing.

In addition to modelling the neural schema architecture for the classification task, the successful application of MTL is contingent on a number of other variables. First, how similar the tasks being completed are, and second, how the network parameters are shared. It has been demonstrated that hierarchically dependent tasks form better task pairs than those that do not. A model trained with Named Entity Recognition and Part of Speech has excellent synergy in the MTL setup, for example. Similarly, Emotion detection and Sentiment classification share a large number of features that can be modelled in MTL through parameter sharing.

Marasović and Frank (2018) classify MTL networks into three categories based on the shared parameters.

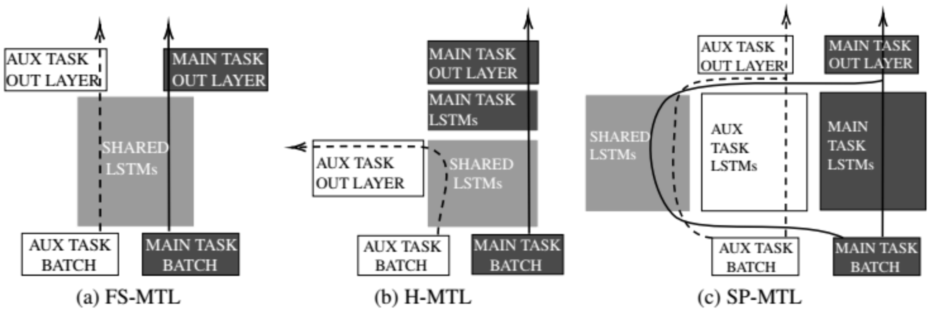

Figure 2.2 Parameter sharing between MRL networks. *(Marasović & Frank, 2018)*

- **Fully Shared MTL model.** The **FS-MTL** (Collobert et al., 2011) shares all parameters between the main task and the auxiliary task with the exception of the input and output layers. There is no interaction between the output layer of one task and the input layer of another task.
- **Hierarchical MTL model (H-MTL).** The **H-MTL** (Søgaard & Goldberg, 2016) model hierarchically organises tasks on the assumption that one task benefits from the other. Consequently, the principal task learns from the representation learned by another task, while maintaining a separate list of parameters for learning its own task. This is a combination of cascaded and multi-task learning.
- **Shared-Private MTL model (SP-MTL).** The **SP-MTL** (P. Liu et al., 2017) model employs a distinct parameter space that is shared by the primary and auxiliary tasks. Each task has the option of utilising either the shared parameters or the private parameters from the respective branch. In our study, we did not use any other task besides sentiment classification. Instead, we relied on the dataset labels and treated them as separate tasks. Moreover, we do not differentiate between our primary and secondary tasks.

2.4 Evaluation metrics

The majority of previous research on sentiment analysis describes the performance of the technique using either accuracy or F1 score, or both. The most sorted metric for reporting binary sentiment classification on English datasets has been accuracy. However, accuracy cannot account for the class imbalance in the reported test set. Thus, macro-F1 is a viable option. The F1 score is the harmonic mean of precision and recall, defined as precision by equation 2.8, recall by equation 2.9, and F1 by equation 2.10. The value of F1 ranges from 0 to 1, and the higher the score, the better the performance of the classification.

$$precision_{class} = \frac{\#correct_{class}}{\#assigned_{class}} \tag{2.8}$$

$$recall_{class} = \frac{\#correct_{class}}{\#total_{class}} \tag{2.9}$$

$$F_{class} = 2\,\frac{\#precision_{class} * recall_{class}}{\#precision_{class} + recall_{class}} \tag{2.10}$$

2.4 Hypothesis testing

When comparing the performance of two models, statistical significance testing is employed since the performance differences may be caused by the proposed adjustment or by random noise. Dror et al. (2018) provide an overview of statistical significance testing in natural language processing for a variety of tasks.

The statistical significance testing confirms that the difference between the two models is not significant by establishing a null hypothesis. It is designated as an alternative hypothesis that there is a considerable difference. The models are tested in the form of distributions using parametric or non-parametric methods. If the difference is greater than a particular threshold, the null hypothesis is rejected and the alternate hypothesis is accepted.

Although parametric tests are more reliable, one fundamental criterion must be met: the data must follow a specific distribution. Because this distribution is unknown in NLP, non-parametric approaches are frequently used. Noreen (1989) defined approximate randomisation testing as a non-parametric technique to statistical significance testing that employs computationally demanding randomisation testing.

3. PROBING LANGUAGE MODELS FOR CROSS-LINGUAL SENTIMENT TRANSFER

Cross-lingual sentiment analysis aims to leverage existing resources from high-resource languages and utilise classification performance for low-resource languages. The source language's resources have an immediate impact on the performance of the target language. We hypothesise that by selecting a strong initial resource, we will be able to improve the overall performance. In this chapter, we aim to use negation as a probe to select a good candidate to act as a pre-trained language model for subsequent fine-tuning. This study has three components. The initial objective is to evaluate a language model's negation-capture performance. Second, we aim to determine how well sentiment analysis performs with language models that perform well with negation. Finally, we enrich the model with the highest performance through multi-task training. Experiments demonstrate that negation is a moderate signal for probing a pre-trained language model (PLM) for cross-lingual sentiment transfer.

3.1 Introduction

The mainstream research in Natural Language Processing (NLP) has focused on a handful of high-resource languages, with an emphasis on English, while ignoring thousands of languages around the world (Bender, 2019). The accuracy of NLP tasks has improved as a result of the development of resources for deep learning techniques in high-resource languages. However, this is not true for languages with limited resources. Therefore, natural language processing for low-resource languages remains an open problem in language processing research.

While high-resource deep learning NLP utilises large annotated datasets, low-resource NLP requires an alternative approach. Data augmentation, distant supervision, cross-lingual projections, pre-trained language models (along with embedding), adversarial training, and meta-learning are possible approaches. Traditional word vectors (Mikolov, Chen, et al., 2013; Pennington et al., 2014) are static; each word is represented by a fixed vector. In contrast, contextual word representations (CWR) assign each word a vector based on the entire input to the model. This process results in different representations for each word based on its position, making the overall context of the input significant for vector generation. The vast majority of these contextual representations are derived from language models that have already been trained on vast amounts of data (Devlin et al., 2019; Peters et al., 2018a), which has a substantial impact on the performance of various NLP tasks.

Contextual word representations have the inherent ability to capture various language characteristics, such as syntax trees (Hewitt & Manning, 2019). This results in various cross-lingual syntactic categories sharing the same cluster in a multilingual setting (Chi et al., 2020). While multilingual PLM has proven useful in languages with limited resources and limited task-specific data, the large number of publicly accessible models raises questions about their selection criteria. There are currently more than 15,000[2] publicly usable models available. These PLMs vary on multiple levels. For instance, the languages used during the pre-training phase, the number of pre-training or fine-tuning tasks in the training cycle, the difference due to the unsupervised objective used on the neural schema, and the number of tuned or learned parameters. The trend of transformer models with a higher number of parameters linked to better downstream performance has resulted in the development of models with 175 billion parameters, such as the Generative Pre-trained Transformer (GPT-3) (T. Brown et al., 2020).

Recent research on analysing the black-box behaviours of deep learning models has led to a search for the captured knowledge within such models. Various supervised learning objectives have been proposed in addition to the self-supervised objective of pre-training PLM, which has produced the best results for downstream tasks. Thus, a PLM can be trained sequentially or concurrently on multiple of these objectives. Each combination yields a specific output, pointing us in the direction of an investigation. "In general, probing is the process of testing for a specific pattern, such as local syntax, long-range semantics, or compositional reasoning, by constructing inputs whose expected output cannot be predicted without the ability to detect that pattern." (Wallat et al., 2020). The probing method examines the specific phenomenon or information embedded in a resource, such as a pre-trained model or word representations, using existing tasks. For instance, Ettinger et al. (2018) evaluated compositional meaning information in sentence embedding, whereas Petroni et al. (2019) utilised a "*fill in the blanks*" style knowledge completion task. In this chapter, we studied how a simple probe can be used to select a candidate model for a downstream task, as well as the behaviour of the PLM before and after fine-tuning.

[2]Huggingface, *Pretrained Models* [website], https://huggingface.co/models, (accessed 12 June 2021)

"Negation is in the first place a phenomenon of semantic opposition" (Horn & Wansing, 2020). This phenomenon plays an important role in sentiment analysis when the opinion of the sentence depends on negation (Dadvar et al., 2011; Jiménez-Zafra et al., 2021). Negation cue and scope detection received particular focus in biomedical text processing (Dalloux et al., 2019; Hagege, 2011; Nawaz et al., 2013).

This chapter seeks to establish a correlation between the phenomenon of negation and pre-trained language models. In addition, we include the tasks of bitext and similarity scoring for paraphrasing. This research was conducted in a multilingual environment. Therefore, a dataset for the three tasks in all languages was required. A manually validated negation dataset was compiled for seven Slavic languages, namely six official European languages and Russian. We proposed a gold standard negation dataset creation workflow and one silver standard negation dataset creation workflow for all languages included in the study. The primary objective was to identify the indicator of a promising PLM model candidate for further fine-tuning. We used cosine similarity as a metric to detect a model's ability to transfer sentiment. We conducted all our research using both pre-existing PLMs and modified models. Even though vanilla PLMs must be fine-tuned, we attempted to correlate the current ability of the model with low-resource language sentiment knowledge transfer. The purpose of this chapter is to determine the optimal backbone encoder for cross-lingual sentiment analysis. We utilised negation datasets in target languages for testing. First, we established the premise of investigating models and their relationship to the sentiment analysis task. In section 3.4, we describe the steps we took to generate our dataset for subsequent model analysis. Section 3.5 enumerates numerous models. It is followed by several datasets and their respective descriptions. The experimental setup is described in Section 7 along with a discussion and conclusion.

Experiments conducted on all six datasets of South Slavic languages reveal a moderate correlation between the sentiment analysis score and the model's ability to score negation. Nevertheless, this is not the case for bitext and paraphrases. The performance analysis indicates that a model with a multilingual PLM backbone performs better in a zero-shot scenario. The multi-task learning (MTL) enrichment setup using PLM degrades the overall classification performance.

3.2 Related work

3.2.1 Sentiment analysis

The field of sentiment analysis is dynamic and constantly strives to enhance performance and address previous obstacles. Research in techniques and their application in a new domain (Blitzer et al., 2007), new languages (Dashtipour et al., 2016), or a new environment such as a low-resource setting are examples of recent challenges (Xia et al., 2021). Due to the extensive literature on sentiment analysis, we would like to refer the reader to a more thorough survey (Dashtipour et al., 2016; R. Liu et al., 2019).

3.2.2 Language models

Recent efforts to transfer knowledge from language models have made substantial progress (T. B. Brown et al., 2020; Devlin et al., 2019; Y. Liu et al., 2019; Peters et al., 2018a). Recently, (Merchant et al., 2020) demonstrated that fine-tuning is a conservative process that does not result in catastrophic forgetting. Jiang et al. (2020) showed, on the contrary, that aggressive fine-tuning that overfits the trained data can be generalised using regularisation and optimisation techniques. Raffel et al. (2020) applied a pre-trained encoder-decoder model to a variety of unsupervised and supervised tasks.

3.2.3 Cross-lingual representations

In a monolingual context, word vectors (Mikolov, Chen, et al., 2013) map words with similar meanings closer together in embedding spaces across languages through simple linear association (Glavaš et al., 2019; Vulić et al., 2019). In a multilingual setting, this is accomplished by mapping multiple languages into one subspace using a multi-adversarial setup (H. Wang et al., 2021). Several of these cross-lingual word embeddings require supervision, but unsupervised methods have been proposed (Artetxe et al., 2017).

A simple sentence encoder works by averaging word embeddings and has been studied in both monolingual (Cer et al., 2018) and multilingual settings (Chidambaram et al., 2019). The network operates by training an encoder on a variety of tasks, including semantic similarity, conversational response prediction, quick thought, and natural language inference. Using existing multilingual PLMs, such as multilingual BERT (mBERT), has produced acceptable zero-shot learning performance; however, mBERT is not trained with explicit cross-lingual signals and has non-aligned multilingual vector spaces (Kulshreshtha et al., 2020). This lack of cross-lingual supervision is presented as an intermediate supervised task (Reimers & Gurevych, 2019) whose loss can range from cross-entropy, mean-squared error, to a triplet objective. Applying a pooling operation to generate a fixed-size sentence representation accomplishes this. These representations are trained on additional NLP tasks. Following this work, knowledge extraction from these existing monolingual to multilingual models has been performed successfully (Reimers & Gurevych, 2020).

3.2.4 Probing

N. F. Liu et al. (2019) investigated linguistic knowledge from diverse PLMs by training a linear model on a frozen backbone for sixteen distinct tasks. They also investigated the transferability of knowledge between various layers. Tenney et al. (2019) investigated various sub-sentence tasks using edge probing tasks. Additionally, previous studies analysed the performance of sentence vectors (Adi et al., 2017; Conneau et al., 2018). Wallat et al. (2020) investigated BERT in order to conclude that fine-tuning objectives influence catastrophic forgetting when utilising knowledge base completion tasks. Petroni et al. (2020) examined the knowledge present in every BERT layer. Chi et al. (2020) investigated subspaces of mBERT in order to retrieve syntactic tree and dependency tree distances in numerous languages. They utilised a structural probe (Hewitt & Manning et al., 2019) that identifies a linear transformation where squared L2 distance encodes the distance between words in the parse tree.

In contrast to previous research (Wallat et al., 2020), we have not restricted our investigation to a single PLM such as Bert or Roberta, but rather to a large number of language models that have been trained on self-supervised as well as multiple supervised objectives. Our work is based on Kassner and Schütze (2020)'s research. Our experiments, however, vary in how they probe the language model. Using a different type of dataset, we expanded the work to multilingual contexts. On the negation task, our probe interacts with the final layer representations at the sentence level. We have not probed for factual or general knowledge, although the dataset used for probing is factually verified.

3.3 Probing language models

Transformers-based language models (Vaswani et al., 2017) trained on the massive text and then tuned on downstream tasks have demonstrated state-of-the-art performance on a variety of downstream tasks (Devlin et al., 2019). Numerous models have been presented (Clark et al., 2019; Y. Liu et al., 2019). These models differ in how they encode the input data (byte pair, sentence piece, or word piece), have encoder and decoder or encoder-only architecture, various unsupervised pre-training tasks (next word prediction, next sentence prediction, masked word prediction), and are trained in the left-to-right or both text directions. The number of tuneable model parameters is a further consideration. In addition to technical differences, the languages used during pre-training have a significant impact on the performance of subsequent tasks.

Recent research has resulted in the availability of an enormous number of PLMs to the research community. This raises the crucial issue of selecting the optimal backbone model for text representation, particularly in the cross-lingual context of low-resource languages. This study investigates the language transfer capability of numerous PLMs for text classification. We first probe the PLM with a scoring function. The next step is to fine-tune the sentiment analysis model. The trained model was evaluated using the same scoring function. To improve the word representations, we reapplied the multi-tasking system with multiple tasks. The setup for fine-tuning and probing is repeated for PLMs that have been enriched. In this section, we propose a simple cosine distance score for choosing a candidate for cross-lingual sentiment transfer. This is investigated further in a multi-task enrichment environment. The sections that follow describe the approaches in greater detail.

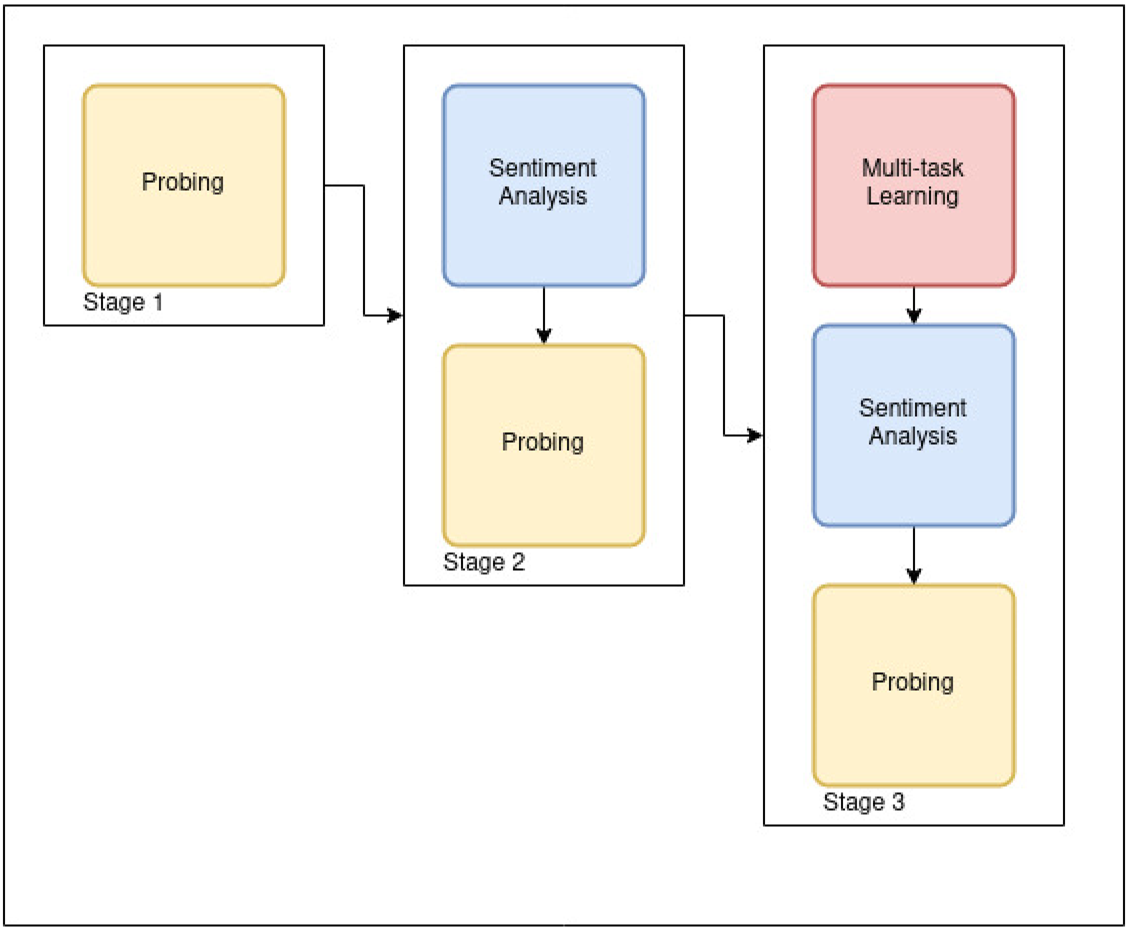

Figure 3.1 Methodology.

3.3.1 Probing

Numerous linguistic phenomena have been demonstrated to be encapsulated by language models. We are more concerned with the semantic nature of acquired knowledge. We hypothesise that the model that accurately captures negated and affirmative sentences will also serve as an effective backbone model for cross-lingual knowledge transfer. Intuitively, similar sentences should be closer together in higher-dimensional space, while negated sentences should be farther apart. As a measure of similarity, the cosine distance was used.

Assuming negation to be the task for probing the model, we computed a similarity metric between each pair of negated and affirmative entries. The metric connects the contextual representations of negated sentences ***t*** to a word vector ***e*** representing affirmative sentences. The two vectors were compared using cosine similarity, which is defined as follows: A low cosine value indicates that an entry is not closely related to the other vector and is, therefore, a good candidate with the ability to differentiate semantically opposing concepts. As a result, this model was employed as the foundation for fine-tuning.

$$cos(t,e) = \frac{te}{\|t\| \cdot \|e\|} = \frac{\sum_{i=1}^{n} t_i e_i}{\sqrt{\sum_{i=1}^{n} t_i^2} \sqrt{\sum_{i=1}^{n} e_i^2}} \quad 3.1$$

Therefore, negated statements should have lower cosine similarity scores, while semantically similar sentences should have higher cosine similarity scores. We proposed three distinct tasks for capturing PLM's semantic capability. To compute the scores, we utilised:

1. a negated dataset (a dataset consisting of affirmative and negative statements), for example, *Let's do that today*. ≠ *Let's not do that today*.
2. a bitext dataset (a dataset consisting of parallel sentences), for example, (Eng) *I have to go to sleep*. ≡ (Cro) *Moram ići spavati*.
3. and paraphrase dataset (a dataset comprising semantically similar sentences), for example, *You're so naive.* ≡ *You are so gullible*.

By stacking Transformer blocks, Transformer-based language models convert unprocessed text into contextualised embedding vectors. Typically, these models use a [CLS] token to represent a sentence or group of sentences. We used this special token to calculate the distance between sentences in the three previously mentioned datasets.

3.3.2 Fine tuning

Either fine-tuning or task-specific pre-training, which can optionally be followed by fine-tuning, can be used to optimise a pre-trained language model for the final downstream task. One of the practices that significantly improves state-of-the-art performance is the fine-tuning of previously trained language models. This is done by adding a new task-specific output layer on top of the original output layer. By applying training data to the pre-trained model and the task-specific layer, we modified the parameters of the new layers and the weights of the pre-trained model. In our case, the subsequent task was the analysis of sentiment. In the zero-shot setup, the entire network was trained using only the source language (English), and the test scores were reported in the target languages. We examined the fine-tuned PLM using a probing task to determine whether the orientation of the representations changed after fine-tuning. The results of the experiments were then compared, contrasted, and confirmed.

3.3.4 Multi-task learning

Before performing fine-tuning, the effect of having three tasks was examined. We retrained the existing PLM with new natural language processing tasks during this phase. These tasks involved the evaluation of bitext, negation, and paraphrase. Based on the task scores, the setup sequentially trained a model. We referred to this process as enrichment because we added knowledge of negation, paraphrase, and bitext to the core model. This was then followed by probing the enriched PLM and fine-tuning the enriched model for the subsequent sentiment classification task. For correlation values, the enriched, fine-tuned model was probed using the same datasets described previously.

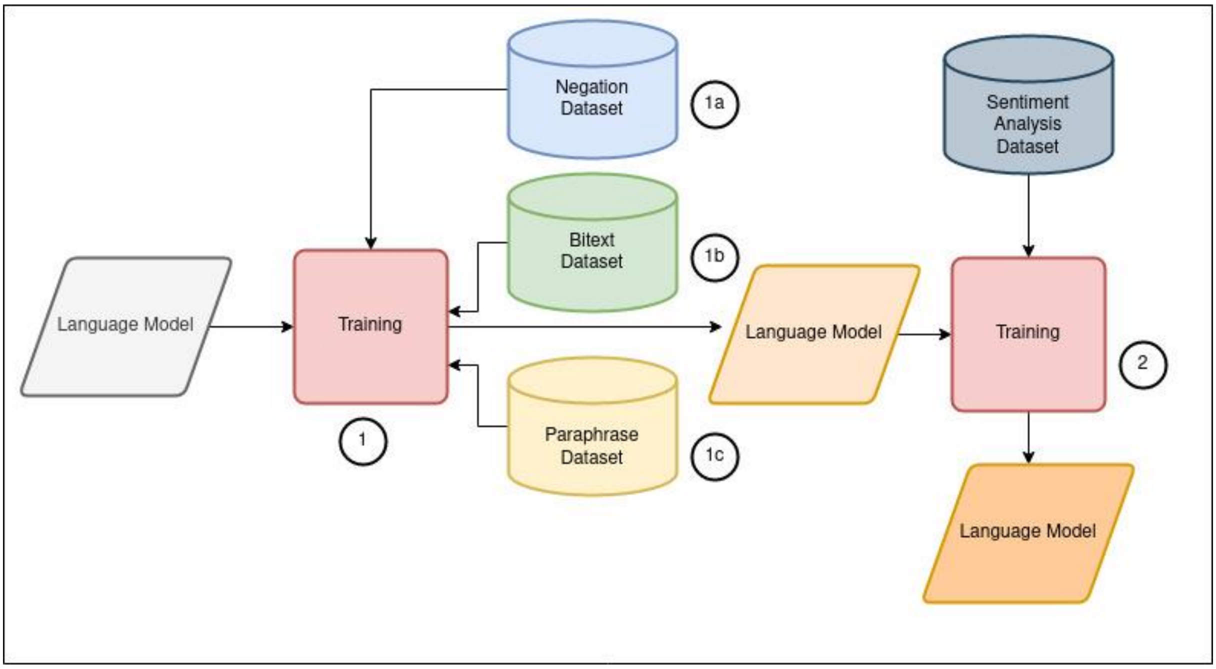

Figure 3.2 Multi-task learning.

3.3.5 Languages in this study

To investigate cross-lingual sentiment transfer, we selected six official South Slavic languages of the European Union: Bulgarian, Croatian, Czech, Polish, Slovene, and Slovak. Except for Czech and Polish, the remaining languages have limited resources for sentiment analysis. In addition to English, which we refer to as the source language, we also included Russian in our test set as a high-resource language. The Russian dataset was left out of the training process.

3.4 Negation dataset creation

3.4.1 Probing strategy

To assess an existing PLM's semantic representation capabilities, we required a dataset for negation, bitext, and paraphrase tasks. All languages had access to bitext and paraphrase, but negated sentences were unavailable. Using the workflow described below, we created a gold-standard dataset for the test and a silver-standard dataset for training.

Round 1: To generate new datasets, we utilised the TaPaCo (Scherrer, 2020) dataset and the Tatoeba[3] database as our primary datasets. The TaPaCo dataset is a corpus of sentential paraphrases for 73 languages derived from the Tatoeba database. We chose this dataset because it contains simple parallel sentences. Using a lookup table, each English sentence is connected to its translation in multiple languages. The corpus is separated into groups of sentences from various languages that are paraphrases of one another. Each sentence has a unique sentence identifier. This sentence ID is derived from the database at Tatoeba. The procedure is as follows:

1. We curated the English negation cue list manually (*"aren't", "can't", "cannot" "couldn't", "didn't", "doesn't", "don't", "hadn't", "hasn't", "haven't", "isn't", "mustn't", "needn't", "negative", "never", "not", "oughtn't", "shan't", "shouldn't", "wasn't", "weren't", "won't", "wouldn't", "no"*). Extraction was restricted to only explicit negation.
2. Using a list of negation cues, English sentences were filtered. To avoid cases of double and triple negation, sentences with a single negation cue were selected. Sentences containing the words "*yet*," "*but*," "*just*," and "*anyone*" were avoided.
3. The explicit negation cues were replaced with their affirmative equivalents, for example, "*do not*", "*don't*" => "*do*". This creates a potential affirmative sentence candidate from the negated sentence.
4. The TaPaCo dataset was searched using affirmative sentences from the previous step. If present, an association is made with the negated sentence.
5. Associations from the TaPaCo dataset were utilised in conjunction with negated and affirmative sentences to identify associated sentences from South Slavic languages.

Step 4 ensures that the sentence created in Step 3 is valid and logical since it is reasonable to assume that the sentence in the TaPaCo dataset is valid. We observed that the initial data distribution of the original corpus influences the final dataset generated when the previously mentioned steps are applied. As shown in Table *3.3*, the data instances for low-resource

[3] https://tatoeba.org/

languages are extremely low. The result is a large volume of negation data in English and Russian, as shown in Table *3.1*. Table *3.2* demonstrates the train-test distribution.

Language	**Number of instances**
English	18,914
Russian	7,446

Table 3.1 Distribution of gold-standard sentences for English and Russian - Round 1.

Language	**Train**	**Validation**	**Test**
English	11,405	1,268	6,243
Russian	4,490	500	2,459

Table 3.2 Train-test distribution.

Language	**Number of instances**
Bulgarian	4
Croatian	0
Czech	12
Polish	0
Slovak	1
Slovene	0

Table 3.3 Distribution for Slavic languages - Round 1.

Round 2. Since the TaPaCo dataset is derived from the Tatoeba database, we redesigned our workflow to extract sentences for Slavic languages for this study. Instead of TaPaCo, the following steps were performed on the Tatoeba database.

1. Searched for and extracted English phrases containing explicit negation cues.
2. Chose a sentence from each Slavic language that matched each English sentence.
3. Replaced the negation cue with the affirmative variant to form an affirmative sentence, for example, (Eng) *That won't happen.* = (Cro) *To se neće dogoditi.* ≠ (Cro) *To se će dogoditi.*

Language	Number of instances
Bulgarian	6,837
Croatian	3,058
Czech	21,937
Polish	27,264
Slovak	2,887
Slovene	1,086

Table 3.4 Silver-standard distribution for Slavic languages - Round 2.

Language	Number of instances
Bulgarian	456
Croatian	608
Czech	617
English	6,241
Polish	464
Russian	2,457
Slovak	634
Slovene	766

Table 3.5 Distribution of negation dataset.

Language	Train	Test
Bulgarian	13,832	1000
Croatian	63,463	1000
Czech	37,291	1000
Russian	406,839	1000
Polish	53,170	1000
Slovak	10,298	1000
Slovene	95,559	1000

Table 3.6 Distribution of the bitext dataset.

Language	Train	Test
Bulgarian	4,375	2,083
Croatian	198	131
Czech	4,493	2,297
English	178,133	46,691

Polish	18,543	18,543
Russian	346,649	62,101
Slovak	278	255
Slovene	82	77

Table 3.7 Distribution of paraphrase dataset.

In each language, this step produced affirmative and negated sentences. Table *3.4* presents the instances' statistics. We searched the Tatoeba database for auto-replaced sentences, but there were no results because the number of sentences in the Tatoeba database for Slavic languages is not comparable to high-resource languages such as English and Russian. This approach generated illogical or grammatically incorrect sentences, as opposed to the first approach, which searched the database for affirmative sentences, thereby ensuring a perfectly valid and meaningful sentence. As a result of replacing negation cues without understanding the context, the following issues arose:

1. **Sentence structure and construction**: The structures of the sentences are incorrect, for example, *I don't expect anything from you ≠ I do expect anything from you*.
2. **Proverbs**: Proverbs with explicit negation cues cannot be easily negated in meaning, for example, *No gains without pains.*
3. **Incorrect word order**: *He didn't do it on purpose. = Nije to uradio namjerno. ≠ je to uradio namjerno.* The appropriate phrase is "*uradio je to namjerno*".

For languages where the genitive case is linked to negation, such as Polish (Przepiórkowski, 2000) and Slovenian (Pirnat, 2015), the generation of silver-standard datasets presents an additional challenge. This phenomenon was not observable in other Slavic languages.

1. **English** *I don't like coffee*.
2. **Croatian** *Ne volim kavu*.
3. **Slovak** *Nemám rada kávu*.
4. **Polish** *Nie lubię kawy*.
5. **Slovenian** *Ne maram kave*.

Therefore, sentences were provided to native speakers for grammatical and semantic review. This resulted in the development of a small test set (450+ instances for each Slavic language) from the silver-standard dataset. Table *3.5* displays the final distribution of the gold-standard dataset.

3.5 Language models and datasets

3.5.1 Pre-trained language models

For our investigation, we selected several publicly accessible, pre-trained language models. This included models trained on architectural approaches such as word2vec (Mikolov, Chen, et al., 2013), in which words are represented by a fixed-length vector, as well as contextualised language models. The models explored include both monolingual and multilingual models. We utilised BERT (Devlin et al., 2019), DistilBERT (Sanh et al., 2019), and RoBERTa (Y. Liu et al., 2019). The **APPENDIX** contains the complete list and additional details.

3.5.2 Datasets

This section provides an overview of the datasets utilised in our experiments. The negation datasets used in the probing experiments are described in Section 6.1. Section 6.2 provides statistics regarding the sentiment analysis datasets used for tuning. Section 6.3 describes the various datasets utilised for multi-task enrichment of language representation models.

3.5.2.1 Probing datasets

Negation This is a collection of datasets according to the workflow described in Section 3.4. It includes both the gold and silver-standard datasets. Table *3.5* lists the size of each language's dataset.

Bitext For bitext analysis, we utilised Tatoeba and WikiMatrix's existing collections. WikiMatrix is used in addition to Tatoeba to supplement language pairs with limited resources, such as Croatian and Slovene. Table *3.6* depicts the distribution of the train-test split.

Paraphrase For paraphrasing, we derived the paraphrase dataset for all languages using the Tapaco dataset. Table *3.7* displays the statistics of the dataset.

3.5.2.2 Sentiment analysis datasets

Bulgarian The Cinexio (Kapukaranov & Nakov, 2015) dataset is comprised of movie reviews with 11-point star ratings: 0 (negative), 0.5, 1,...4.5, 5 (positive). Other meta-features included in the dataset are film length, director, actors, genre, country, and various scores.

Croatian Pauza (Glavaš et al., 2013) contains restaurant reviews from Pauza.hr[4], the largest food ordering website in Croatia. Each review is assigned an opinion rating ranging from 0.5 (worst) to 6 (best). User-assigned ratings are the benchmark for labels. The dataset also contains opinionated aspects.

Czech The CSFD (Habernal et al., 2013) dataset was influenced by Pang et al. (2002). It includes film reviews from the Czech Movie Database[5]. Every review is classified as either positive, neutral, or negative.

English The Multilingual Amazon Reviews Corpus (MARC) is a large collection of Amazon reviews (Keung et al., 2020). The corpus contains reviews written in Chinese, English, Japanese, German, French, and Spanish. Each review is assigned a maximum of five stars. Each record contains the review text, the title, the star rating, and product-related meta-data.

Polish The Wroclaw Corpus of Consumer Reviews Sentiment (Kocoń et al., 2019) is a multi-domain dataset of Polish reviews from the domains of schools, medicine, hotels, and products. The texts have been annotated at both the sentence level and the text body level. The reviews are labelled as follows: [+m] represents a strong positive; [+s] represents a weak positive; [-m] represents a strong negative; [-s] represents a weak negative; [amb] represents ambiguity; and [0] represents neutrality.

Russian The ROMIP-12 dataset (Chetviorkin & Loukachevitch, 2013) is comprised of news-based opinions, which are excerpts of the direct and indirect speech published in news articles. Politics, economics, sports, and the arts are just some of the diverse subject areas covered. This dataset contains speech classified as positive, neutral, or negative.

Slovak The Review3 (Pecar et al., 2019) is comprised of customer evaluations of a variety of services. The dataset is categorised using the 1-3 and 1-5 scales. The Sentigrade[6] dataset

[4] http://pauza.hr
[5] http://www.csfd.cz
[6] https://sentigrade.fiit.stuba.sk/data

contains 1,588 Slovak-language comments from various Facebook pages. The annotations on the texts range from -2 to +2.

Slovene The Opinion corpus of Slovene web commentaries KKS 1.001 (Kadunc & Robnik-Šikonja, 2017) includes web commentaries on various topics (business, politics, sports, etc.) from four Slovene web portals (RtvSlo, 24ur, Finance, Reporter). Each instance within the dataset is tagged with one of the three labels (negative, neutral, or positive).

Label transformation Because not all datasets contain the same number of labels, we evaluated them in three distinct scenarios:

a. a scale with five points ranging from 1 to 5,
b. a three-class scale that labels negative, neutral, and positive sentiments.
c. a two-class label prediction scale (positive and negative).

In the case of Croatian, the Pauza dataset contains 11 classes (0–6) that are mapped to five classes using the following formula:

$$t = \frac{x - a}{b - a} \tag{3.2}$$

$$\boldsymbol{v} = (\boldsymbol{B} - \boldsymbol{A}) * \boldsymbol{t} + \boldsymbol{A} \tag{3.3}$$

Where **A** - new min, **B** - new max, **a** - old min, **b** - old max, **x** is the value mapped and **v** is the new value. Table *3.8* provides a summary of all datasets with the corresponding train-test split distribution.

Language	**Dataset**	**Train**	**Val**	**Test**
Bulgarian	Cinexio	5,520	614	682
Croatian	Pauza	2,277		1,033
Czech	CSFD	63,966	13,707	13,707
English	MARC	200,000	5,000	5,000
Polish	all2	28,581	3,572	3,572
	all4	6,771	846	846
Russian	ROIMP 2012	4,000	260	5,500
Slovak	Reviews3	3,834	661	1,235
	Sentigrade	1,143	127	318
Slovene	KKS	3,977	200	600

Table 3.8 Distribution of sentiment analysis datasets.

3.5.2.3 Multi-task learning datasets

In addition to the negation, bitext, and paraphrase datasets, we utilised the Semantic Textual Similarity STS (Cer et al., 2017) dataset, which measures sentence meaning similarity. Every sentence pair is given a score between 0 (no meaning overlap) and 5 (meaning overlap) (almost identical in meaning). ALLNLI is comprised of two distinct natural language inference (NLI) datasets, namely SNLI Stanford NLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018). NLI is concerned with the "task of determining the inference relationship between two (short, ordered) texts: entailment, contradiction, or neutrality" (MacCartney & Manning, 2008). We also experimented with the Amazon reviews compiled by Prettenhofer and Stein (2010), which consisted of 6,000 reviews mapped to the values 0 (negative) and 1 (positive). In our experiments, we refrained from utilising the negation silver data generated for the EU Slavic languages because we did not want the noise to be present during training. However, we did conduct tests with the translation of negated English sentence pairs into six languages.

3.6 Experiments and results

3.6.1 Implementation details

Probing: We computed the cosine similarity on text pairs for each language and each probing task using the top 600 instances from the respective test sets for each language. The PLM model was loaded and vector representations were extracted in order to obtain the word representation. In instances where the [CLS] token was unavailable due to the language model's architecture, the final representations were derived by applying mean-pooling to the individual token vectors. For each model, the texts were truncated and padded to the maximum input size the model supported.

Fine-tuning: Following the method described in Devlin et al. (2019), we used a PLM as the backbone and stacked a single 5-class softmax classification layer (0–4 stars). We fine-tuned a classification layer using the Adam optimiser and a constant learning rate of 2e-5 for four epochs. We utilised a batch size of 16 with automatic mixed precision. With the Pytorch-lightning framework and a single RTX 3090, 24 GB GPU, each experiment required ≈ 2 hours. We selected model checkpoints using the source language's development set.

Multi-task learning: We trained the backbone model on three different tasks. In this phase, we conducted sequential training on the PLM, with each task having a predetermined objective. We sampled a single task and a single batch from the corresponding dataset. Each task was trained with its separate Adam optimiser for 16 epochs and a learning rate of 2e-5. For the warm-up, we utilised a batch size of 8 and 10% of the training data. Each task was evaluated using the Spearman correlation constant for the labels and their cosine scores. Using the development set, the best model checkpoint was chosen. As with the experiment on fine-tuning, the automatic mixed precision was used for training. Contrastive loss (Hadsell, Chopra and LeCun, 2006), which increases or decreases the distance between two embeddings based on the label, i.e., 0 (negation) or 1 (positive) (bitext, paraphrase), was used for each task. We utilised English and Russian resources for the negation task because we lacked gold-standard data for other low-resource languages. While the negation task was trained in two languages (English and Russian), the bitext task was trained in seven language pairs (Bulgarian, Czech, Croatian, Polish, Russian, Slovenian, and Slovak) with English as the source language. The same situation existed for the task of paraphrasing in eight languages. The PLM was trained on a negation dataset with labels 0 and a contrastive loss, so it was anticipated that the word embeddings of affirmative and negative sentences would be further apart. This is anticipated to aid language representations. The label presented in the case of bitext is 1. Therefore, after training, embeddings were anticipated to be closer. This addresses the interlingual situation. Training for the paraphrasing task involves bringing word vectors with similar meanings closer in intralingual semantic space.

Extra tasks: Aside from the three primary tasks, we conducted additional experiments and added additional tasks to multi-task learning setup. However, we were unable to train an exhaustive list of all possible combinations. This is because our computational infrastructure lacks the computational capacity to manage large models and their parameters. The STS task taught a network to bring two sentences with scores ranging from 0 (not similar) to 5 (almost identical) closer or further apart based on the label. ALLNLI is a classic NLI classification problem involving three classes with softmax loss. Similar to ALLNLI, we introduced a sentiment analysis task with softmax loss, though it was not entirely trained on a massive dataset. To evaluate the impact of having negation datasets in Slavic languages, we incorporated a Google-translated version of the English negation dataset as the final task.

3.6.2 Results of probing

	Model	**Negation**	**Bitext**	**Paraphrase**
1	LaBSE	0.8525	0.8891	0.9252
2	allenai-specter	0.9708	0.7340	0.9503
3	bert-base-nli-cls-token	0.9113 ⊖	0.4104	0.9384
4	bert-base-wikipedia-sections-mean-tokens	0.9988 ⊖	0.9855 ⊕	0.9982 ⊕
5	german-roberta-sentence-transformer-v2	0.7920 ⊕	0.9300 ⊕	0.9213
6	msmarco-roberta-base-ance-fristp	0.9975	0.9769	0.9953 ⊕
7	nli-bert-large-cls-pooling	0.9253	0.4713	0.9517
8	nli-bert-large-max-pooling	0.9342	0.5826	0.9539
9	nli-bert-large	0.9342	0.4553	0.9409
10	nli-distilbert-base-max-pooling	0.9420	0.5839	0.9568
11	paraphrase-xlm-r-multilingual-v1	0.8734	0.9454	0.9375
12	xlm-r-100langs-bert-base-nli-mean-tokens	0.5772 ⊕	0.9642 ⊕	0.9625
13	xlm-r-100langs-bert-base-nli-stsb-mean-tokens	0.6003 ⊕	0.9465 ⊕	0.9456
14	xlm-r-distilroberta-base-paraphrase-v1	0.8734	0.9454 ⊕	0.9375
15	clip-ViT-B-32-multilingual-v1	0.9913 ⊖	0.9853	0.9899 ⊕
16	distilbert-base-nli-stsb-mean-tokens	0.8802	0.0959 ⊖	0.8704
17	distilbert-multilingual-nli-stsb-quora-ranking	0.9438 ⊖	0.9778	0.9803 ⊕
18	distilroberta-base-msmarco-v2	0.9094	0.0634 ⊖	0.8426
19	msmarco-distilbert-base-v3	0.8864	0.0518 ⊖	0.8140
20	quora-distilbert-multilingual	0.9438	0.9778	0.9803 ⊕
21	xlm-r-large-en-ko-nli-ststb	0.6025 ⊕	0.7746	0.8756
22	stsb-xlm-r-multilingual	0.6003 ⊕	0.9465	0.9456
23	average_word_embeddings_glove.6B.300d	0.6344 ⊕	-0.0309 ⊖	0.6132 ⊖
24	average_word_embeddings_glove.840B.300d	0.6093 ⊕	-0.056 ⊖	0.5750 ⊖
25	average_word_embeddings_komninos	0.6269 ⊕	0.2036	0.6102 ⊖
26	average_word_embeddings_levy_dependency	0.5876 ⊕	0.2449	0.5574 ⊖
27	CroSloEngual BERT	0.9369	0.4795	0.8988
28	xlm-roberta-base	0.9982 ⊖	0.9955 ⊕	0.9979 ⊕
29	bert-base-multilingual-cased	0.8950	0.5427	0.8505

Table 3.9 Cosine similarity scores. ⊕ best and ⊖worst.

Table *3.9* compares three tasks for each PLM. Using cosine similarity scores for each of the three tasks, we identified the models with the highest performance for each task. No model performed optimally in every task. The best scoring models for the negation task are xlm-r-100langs-bert-base-nli-mean-tokens and word embeddings, respectively. The PLM with the lowest efficiency is bert-base-wikipedia-sections-mean-tokens. The model with the best performance in bitext was xlm-roberta-base. Consequently, the performance of the average word embedding model is the worst. We observe bert-base-wikipedia-sections-mean-tokens as the model with the best performance for the task of paraphrasing, while average word embeddings perform poorly.

3.6.3 Results of fine-tuning

Language	**5 star F1 micro**	**5 star F1 macro**	**3 star F1 micro**	**3 star F1 macro**	**2 star F1 micro**	**2 star F1 macro**
Bulgarian	⊞ 0.533	⊞ 0.464	⊞ 0.797	⊞ 0.654	⊞ 0.863	⊞ 0.80
English	⊞ 0.596	⊞ 0.597	⊞ 0.772	⊞ 0.772	⊞ 0.912	⊞ 0.912
Croatian	⊠ 0.72	⊠ 0.531	⊠ 0.865	⊞ 0.678	⊠ 0.914	⊠ 0.881
Czech			⊞ 0.613	⊞ 0.593	⊡ 0.974	⊡ 0.493
Polish	△ 0.369	⊞ 0.257	⊞ 0.542	⊞ 0.514	⊞ 0.854	⊞ 0.853
Russian			⊟ 0.736	∩ 0.338	∪ 1.0	∪ 0.5
Slovak	⊡ 0.514	⊞ 0.399	⊞ 0.628	⊞ 0.559	⊞ 0.967	⊞ 0.944
Slovene			◂ 0.686	⊞ 0.427	⊡ 0.963	⊡ 0.49

Table 3.10 Fine-tuning scores. ⊠ CroSloEngual BERT ($p<0.05$) ⊞ xlm-roberta-base ($p<0.05$) ⊟ bert-base-wikipedia-sections-mean-tokens ⊡ nli-bert-large-cls-pooling ◂ distilroberta-base-msmarco-v2 ($p<0.05$) ∩ nli-bert-large-max-pooling ($p<0.05$) ∪ allenai-specter † bert-base-nli-cls-token △ nli-bert-large. The models were checked for statistical significance at $p<0.05$.

In Table *3.10* we present the outcomes of fine-tuning PLM and evaluating it on the respective datasets from the languages in this study. The table lists the optimal values for each test scenario (5-class, 3-class, and 2-class). The model xlm-roberta-base performs the best for Bulgarian. For the English test set, the scores matched the mBERT encoder from the original paper (Keung et al., 2020), whereas the xlm-roberta-base version improved by two points. In the majority of test cases, CroSloEngual BERT outperformed all other models for Croatian. Due to the 3-class nature of the Czech dataset, 5-class scores were not reported. The best classification models include xlm-robert-base, crosloengual-bert and nli-bert-large-max-pooling. For Polish, the majority of metrics were scored by xlm-roberta-base. In Russian, Allenai-spector outperformed other models. The zero-shot performance scores for Russian and Polish are inadequate. The xlm-roberta-base model performed exceptionally well for Slovak. Slovene performed well using distillroberta-base-msmacros-v2 for 3-class.

3.6.4 Results of multi-task learning

Language	Model	Metric	Score
Bulgarian	negation-clip-ViT-B-32-multilingual-v1	f1-micro-5	50.4
	negation-sentiment-stsb-xlm-r-multilingual	f1-macro-5	31.8
	negation-clip-ViT-B-32-multilingual-v1	f1-micro-3	71.5
	negation-sentiment-stsb-xlm-r-multilingual	f1-macro-3	50.8
	negation-sentiment-average-word-embeddings-levy-dependency	f1-micro-2	82.6
	negation-bitext-paraphrase-xlm-r-100langs-bert-base-nli-stsb-mean-tokens	f1-macro-2	68.9
English	negation-sentiment-bert-base-nli-cls-token	f1-micro-5	58.1
	negation-sts-allnli-sentiment-bert-base-nli-cls-token	f1-macro-5	57.6
	negation-bitext-paraphrase-bert-base-wikipedia-sections-mean-tokens	f1-micro-3	76.1
	negation-bitext-paraphrase-sts-allnli-sentiment-bert-base-nli-cls-token	f1-macro-3	71.6
	negation-sts-allnli-bert-base-nli-cls-token	f1-micro-2	90.4
	negation-sts-allnli-bert-base-nli-cls-token	f1-macro-2	90.3
Croatian	negation-bitext-paraphrase-distilbert-base-nli-stsb-mean-tokens	f1-micro-5	59.1
	negation-bitext-paraphrase-nli-distilbert-base-max-pooling	f1-macro-5	29.8
	negation-negoogle-nli-bert-large	f1-micro-3	73.4
	negation-bitext–paraphrase-sts-allnli-nli-distilbert-base-max-pooling	f1-macro-3	46.4

	negation-bitext-paraphrase-sts-allnli-sentiment-negoogle-nli-distilbert-base-max-pooling	f1-micro-2	78.6
	negation-bitext-paraphrase-nli-bert-large	f1-macro-2	68.9
Czech	negation-sentiment-xlm-r-100langs-bert-base-nli-stsb-mean-tokens	f1-micro-3	44.4
	negation-sentiment-quora-distilbert-multilingual	f1-macro-3	41.0
	negation-bitextnli-bert-large	f1-micro-2	1.0
	negation-bitext-bert-large	f1-macro-2	50.
Polish	negation-bitext-distilbert-base-nli-stsb-mean-tokens	f1-micro-5	44.8
	negation-sentiment-xlm-r-100langs-bert-base-nli-stsb-mean-tokens	f1-macro-5	24.5
	negation-bitext-xlm-r-100langs-bert-base-nli-stsb-mean-tokens	f1-micro-3	50.3
	negation-bitext-xlm-r-100langs-bert-base-nli-stsb-mean-tokens	f1-macro-3	50.5
	negation-sentiment-xlm-r-100langs-bert-base-nli-stsb-mean-tokens	f1-micro-2	71.7
	negation-sentiment-xlm-r-100langs-bert-base-nli-stsb-mean-tokens	f1-macro-2	70.2
Russian	negation-sentiment-bert-base-nli-cls-token	f1-micro-3	73.6
	negation-bitext-paraphrase-sts-bert-base-nli-cls-token	f1-macro-3	37.3
	negation-bitext-nli-bert-large	f1-micro-2	1.0
	negation-bitext-nli-bert-large	f1-macro-2	50.
Slovak	negation-bitext-training-average-word-embeddings-levy-dependency	f1-micro-5	55.9
	negation-bitext-xlm-r-100langs-bert-base-nli-mean-tokens	f1-macro-5	30.0
	negation-bitext-paraphrase-sts–distilbert-base-max-pooling	f1-micro-3	62.1
	negation-bitext-paraphrase-sts-distilbert-base-max-pooling	f1-macro-3	57.5

	negation-bitext-training-nli-bert-large	f1-micro-2	1.0
	negation-bitext-distilbert-base-nli-stsb-mean-tokens	f1-macro-2	72.0
Slovene	negation-nli-bert-large-max-pooling	f1-micro-3	72.6
	negation-bitext-paraphrase-bert-base-nli-cls-token	f1-macro-3	42.6
	negation-bitext-training-nli-bert-large	f1-micro-2	1.0
	negation-bitext-training-nli-bert-large	f1-macro-2	0.5

Table 3.11 MTL sentiment analysis score.

Similar to the fine-tuning setup, we list the micro and macro F1 scores for the best-performing models for the respective setting in Table *3.11*. When the MTL system is connected to a training setup, the overall performance values consistently decline. To validate our claim regarding the relationship between sentiment score and three probing tasks, we calculated the Spearman rank correlation coefficient $\boldsymbol{\rho}$. Across all languages and models, the correlation between negation and the F1-score of the sentiment classification is statistically significant and moderate (≈ 0.38). The relationship between the bitext and paraphrase scores does not hold. Figure 3.3 displays the cosine scores for the negation task on PLMs in ascending sentiment score order. The red cosine score represents the initial cosine scores, while the light red score represents the after-sentiment cosine score. Figure 3.4 displays the refined MTL models alongside their negation and sentiment scores. We also find it noteworthy that the bitext score correlates with the classification score, whereas the paraphrase score does not.

Figure 3.3 Cosine score before and after fine-tuning vs sentiment score.

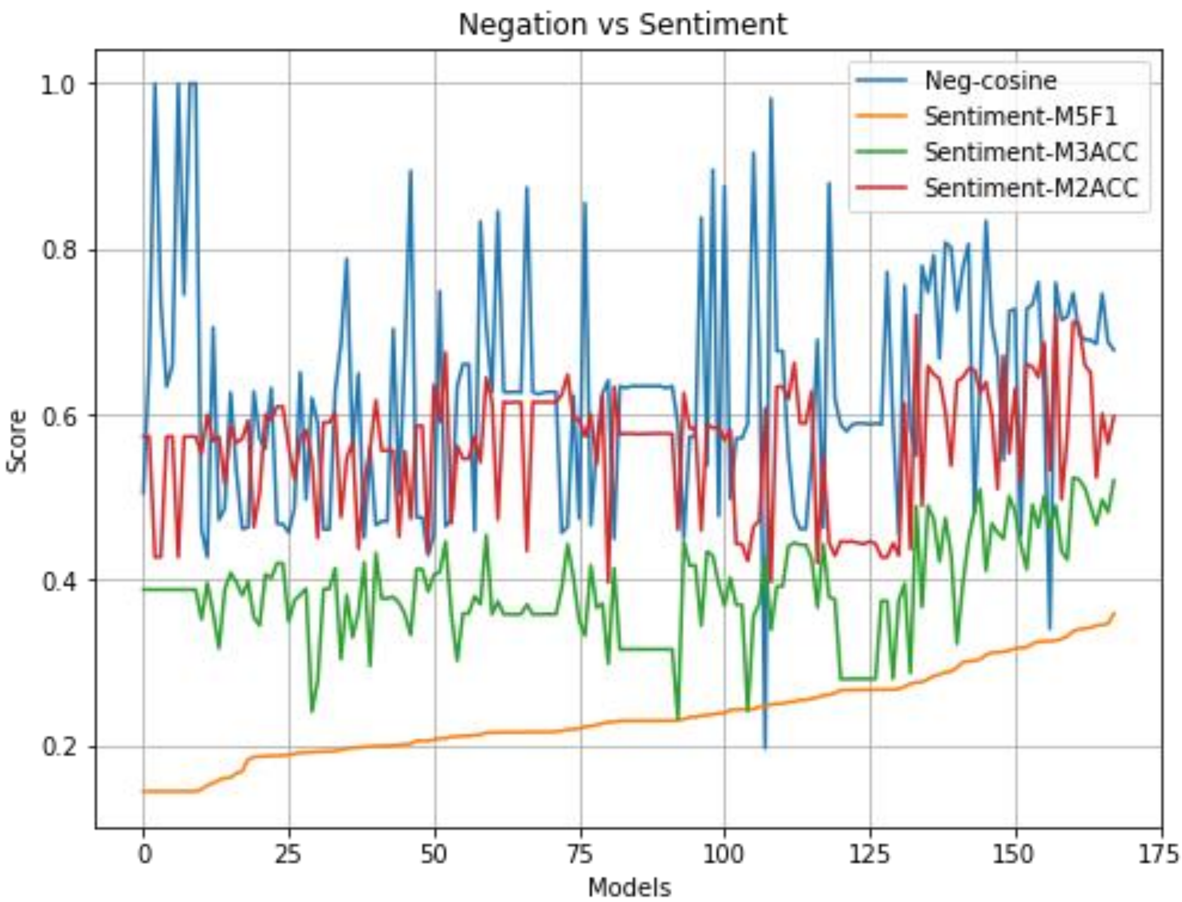

Figure 3.4 Cosine score vs sentiment scores for MTL

No clear winner exists for the fine-tuning setting. The xlm-roberta-base model appears most frequently in Table *3.10*. This could be attributed to its multilingual nature and number of parameters. Ulcar and Robnik-Sikonja (2020) reported that a multilingual PLM with fewer languages improves performance in closely related languages. This is evident in the CroSloEngual BERT model's higher scores for Croatian and can be attributed to the fact that the model only contains data for Croatian, English, and Slovene. S. Wu and Dredze (2020) state that "better models for low-resource languages require more efficient pre-training techniques or more data" and recommend using models trained on similar languages. In other words, the cross-lingual transfer setup necessitates models that are better trained in languages with limited resources. According to Kassner and Schütze (2020), the majority of PLMs are unable to distinguish between negated and non-negated sentences. The low number of negated sentences in the pre-training period had a direct correlation with poor performance. One option for resolving this issue is to teach the model negation through supervised instruction.

We conducted hypothesis testing utilising the random approximation test for the cosine and sentiment classification scores in order to establish the statistical significance of the data obtained from numerous experiments. Additionally, following the approach outlined in Tesfagergish et al. (2022), we used the Friedman test (Friedman, 1937) and the posthoc Nemenyi test (Hollander et al., 2013) to assess the effectiveness of the scores of various sentiment classification models. For the random approximation test, we performed 10,000 iterations and used a fixed seed for reproducibility.

For negation before and after the first SA fine tuning, there is a strong correlation (85.64), with statistical significance, but there is not an absolute difference or improvement between the negation values. A similar situation exists with the bitext and paraphrase detection values. There is no significant change in the before and after scores of the negation, bitext, and paraphrase once trained on sentiment. There is a statistically significant correlation between negation scores and the models trained on MTL and their corresponding sentiment analysis scores. There is a statistically significant correlation between paraphrase and the models trained on MTL and their corresponding sentiment analysis scores.

	Experiments	**Is significant?**
Sentiment vs probing tasks	Sentiment vs Negation	38.63*
	Sentiment vs Bitext	no
	Sentiment vs Paraphrase	no
Among probing tasks using vanilla model	Negation vs Bitext	no
	Negation vs paraphrase	47.06*
	Bitext vs Paraphrase	66.12*
Among probing tasks using FT sentiment model	FT Negation vs FT negation	85.64*
	Bitext vs FT Bitext	84.22*
	Paraphrase vs FT paraphrase	94.60*
FT Sentiment vs FT probing tasks	Negation FT vs Sentiment	44.03*
	Sentiment vs FT Bitext	no

	Sentiment vs FT Paraphrase	43.77*
MTL probing task vs MTL sentiment	MTL Sentiment vs MTL Negation	0.23*
	MTL Sentiment vs MTL Bitext	-0.19*
	MTL Sentiment vs MTL Paraphrase	0.38*

Table 3.12. Summary of correlations among various entities that are statistically significant *(p<0.05).

3.7 Conclusion

In this chapter, we described our methodology for probing a PLM for cross-lingual sentiment analysis. In the initial phase, we examined the correlation between the various pre-trained language models and sentiment analysis tasks using simple tasks. There is a moderate correlation between the negation and semantic comprehension for the cross-lingual sentiment analysis score. We discovered that the correlation between the bitext and paraphrase similarity scores and the sentiment analysis score is not significant. This suggests that simple negation can be used to choose a good PLM for the subsequent task of sentiment analysis. The performance of the XLM-Roberta-base model is superior to that of the other models. The findings indicate that PLMs with target languages should be utilised. The plans for future research are as follows: (1) to investigate the intermediate layers of the multi-layered pre-trained language models (van Aken et al., 2019); and (2) to enhance the performance of cross-lingual sentiment analysis for low-resource languages. Our objective is to probe the models for each layer and delve deeper into the xlm-roberta-base for cross-lingual sentiment analysis.

4. TRANSFERRING SENTIMENT CROSS-LINGUALLY WITHIN AND ACROSS SAME FAMILY LANGUAGES

Natural Language Processing for languages with limited resources is hampered by a lack of data. Using English as a hub language for such languages, cross-lingual sentiment analysis has been developed. The sheer quantity of English language resources raises questions about its status as the primary resource. This research aims to examine the impact on sentiment analysis of adding data from same-family versus distant-family languages. We analyse the performance using low-resource and high-resource data from the same language family (Slavic), investigate the effect of using a distant family language (English) and report the results for both settings. Quantitative experiments demonstrate that adding a large quantity of data from related and distant family languages is advantageous for cross-lingual sentiment transfer.

4.1 Introduction

The earlier chapter briefly introduced the premise of cross-lingual sentiment analysis. We used the probe to identify the candidate backbone model. The performance of the pre-trained models in a zero-shot setting was computed using the target language test set. Subsequently, the best models for transferring sentiment knowledge between source and target languages were identified. The purpose of this chapter is to delve deeper into the effects of language family and the number of supervised resources on cross-lingual sentiment transfer.

Classification of sentiment is essential to text analysis. It is concerned with the automatic extraction of subjective data from text sources. The data provides a clear picture of the entities of interest, such as people, products, aspects, or concepts. The process assigns labels with varying granularity depending on the task. For instance, labels can be positive-negative (Go et al., 2009), positive-neutral-negative (Nakov et al., 2013), positive-negative-mixed-other (Shamma et al., 2009), or positive-negative-neutral-mixed-other (Saif et al., 2013). Earlier work concentrated primarily on extracting well-designed features (Agarwal et al., 2011; Wilson et al., 2005, 2009). Recent work with neural networks simplifies the feature engineering required to extract features from input text (Socher et al., 2011). The main challenge posed by such deep neural networks is the requirement for training data (supervision). The availability of such supervised resources is a challenge for languages with limited resources. Cross-lingual Sentiment Analysis (CLSA) makes use of resources from high-resource languages to construct a sentiment analyser for low-resource languages. For instance, the simplest configuration translates instances of data from the target language to the source and applies the classifier trained on the high-resource source language (Wan, 2009). Using Machine Translation (MT) systems to translate the resources from the source language (annotated datasets or lexicons) into the target language is an alternative method (Balahur & Turchi, 2012; Banea et al., 2008). However, such an accurate translation system is not always available for language pairs with limited resources. Similarly, the prior attempt employed parallel data (A.R. et al., 2012).

Recent research utilising word-embeddings and context-sensitive representations, such as GPT (Radford et al., 2019), ELMO (Peters et al., 2018b), BERT (Devlin et al., 2019), and ROBERTA (Y. Liu et al., 2019), has improved overall classification performance. Pre-training on large corpora is used to acquire these representations. In a multilingual environment, multiple languages are trained collectively on a single model. Downstream tasks like Named Entity Recognition and Classification (NERC) (Grancharova & Dalianis, 2021) or Question Answering (QA) (Z. Wang et al., 2019) refine the pre-trained language models. The primary challenge with multilingual pre-trained language models is the effect of similar languages and representation in the learned space; for instance, the Multilingual Bert (MBERT) (Devlin et al., 2019), which has been trained in 104 languages, does not represent each language in terms of training corpus proportionally. Fine-tuning MLLM for under-represented languages results in poor performance for the target language. In addition to underrepresentation, many languages are absent from these PLMs. All of these conditions are a result of the lack of data for languages with limited resources.

Sentiment classifiers trained on pre-trained language modelling tasks have demonstrated cutting-edge performance (Jiang et al., 2020; Z. Yang et al., 2019). Even though these approaches have been investigated in a cross-lingual context, their application to low-resource languages, particularly languages within the same language family, remains to be investigated. Our analysis investigates the transfer of knowledge between languages of the same language family. We seek the optimal means of combining source-language and target-language data sources. This chapter describes all the techniques and experimental analyses for combining high-resource languages with low-resource languages.

Throughout the past decade, cross-lingual sentiment classification has remained an active field of study. Das and Sarkar (2020) classify cross-lingual processing approaches as either model transfer or annotation adaptation. Model transfer utilises language-independent features. One of the ways to learn these characteristics is through adversarial training (X.

Chen et al., 2019; Kandula & Min, 2021). These cross-lingual representations are optimised for the final task, such as the recognition and classification of parts of speech or named entities. Methods for annotation projection utilise massive parallel corpora between the source and target languages. They exploit the semantic similarity between the parallel corpora. Using the source-trained classifier on a machine-translated view of the target dataset is the simplest approach. As previously observed (Lohar et al., 2019), machine translation introduces noise into the translation, altering the final output's meaning. The classification of noisy input does not guarantee its conformance to the target instance class. The second class of methods (X. Chen et al., 2018) combines model transfer and annotation adaptation into a single unit. The configuration simultaneously trains the shared encoder and parallel corpora for alignment and classification tasks.

4.2 Research questions and hypotheses

Empirically, we pose the following question for our proposed study:
Q. What is the effect of language similarity and available resources inside of MLLM?
We hypothesise that the following:

- A cross-lingual transfer is more successful for typologically similar languages than typologically different languages.
- A large, annotated dataset in a distant family language can overcome typological differences, unlike a small, annotated dataset in a close family language.

To answer the research question, which was to examine the effect of typology on the performance of cross-lingual sentiment analysis, we trained models using English and Slavic language datasets. The training involved the combination of diverse language datasets. We calculated the effect of utilising a language during training and its effect on final performance in several different combinations. The outcomes were compared to previously published research. We determined the optimal language combination for sentiment transfer.

This chapter's contributions are as follows:

- Initially, we propose a framework for unified deep learning that utilises existing data labels from high-resource languages on low-resource datasets. We conduct rigorous experiments on languages within the same language family. We investigated how effectively sentiment classification capabilities could be transferred.

- Second, we demonstrate that, given multiple large-scale training datasets, our framework is superior to a straightforward setup for fine-tuning.
- Finally, we devised the optimal method for jointly training sentiment analysis systems in order to address the issue of insufficient resources for target languages.

4.3 Languages in this study

A language family is a collection of languages that share a common ancestor. English, for instance, is a member of the Indo-European (IE) language family. The languages share characteristics such as phonology, morphology, and syntax. The language family is subdivided into branches that are categorised as subsets. For instance, one of the branches of IE, Balto-Slavic, has a Slavic branch that is subdivided into West, South, and East subgroups (Sussex & Cubberley, 2006): Russian, Belarusian, and Ukrainian (of the East group), Polish, Czech, and Slovak (of the West group), and Bulgarian and Macedonian (eastern dialects of the South group), and Serbo-Croatian and Slovene (western dialects of the South group). We chose to concentrate on three West Slavic languages (Czech, Slovak, and Polish), three South Slavic languages (Croatian, Slovene, and Bulgarian), and one East Slavic language (Russian). Czech and Slovak have the highest degree of mutual intelligibility, followed by Croatian and Slovenian (Golubović & Gooskens, 2015). Except for Bulgarian and Russian (which use the Cyrillic script), all languages use the Latin alphabet. Russian has a complex case system, whereas Bulgarian has lost almost all of its case declension (Townsend & Janda, 1996).

4.4 Related work

4.4.1 Sentiment analysis

Turney (2002) extracted phrases containing adverbs and adjectives by focusing on consecutive words within the context. Patterns were applied to this phrase extraction to eliminate the influence of proper names. 'Excellent' and 'poor' were used to calculate the semantic orientation (SO) of the phrase. The final review score was determined by averaging the semantic orientation of the phrases. The author noted that a text from a particular domain has a distinct writing style that can mislead the final grade. Vanilla sentiment lexicon-based methods employed either the presence or absence of words or the scoring of individual words in the text (S.-M. Kim & Hovy, 2004), ultimately averaging the final score. The authors chose a list of verbs, adjectives, and nouns as a starting point and expanded it using WordNet. Using the synsets from WordNet, a word's polarity score was calculated. The final class was derived from emotionally charged words. Using negation, intensifiers, and diminishers, the lexicon-based technique (Polanyi & Zaenen, 2006) was investigated. The combination of positive and negative words inverts the overall evaluation. In contrast, negative phrases and negation result in a positive final evaluation. The use of modal operators establishes a context for the possibility of necessity. Therefore, realis and irrealis events should be treated differently, as irrealis situations do not necessarily reflect the true attitude of opinion holders toward a concept, as they do in the realis context. Other linguistic structures mentioned by the authors include presuppositional items (such as ‘it is barely sufficient’), connectors, and irony. The earliest attempts were rule-based methods with a high degree of precision (Riloff & Wiebe, 2003) that relied heavily on subjective lexicons and patterns. Results were obtained using two classifiers that relied on the presence and absence of subjective clues for subjective and objective classification. The initially classified sentences are then subjected to pattern extraction and iterated in a bootstrapping process to increase the classifier's lexicon size and coverage. The training dataset was used to train a Naïve Bayes classifier for ranking unlabelled text corpora and passed through the initial pattern extraction procedure to enhance the self-training procedure. Several sentiment lexicons include SentiWordNet (Esuli & Sebastiani, 2006), General Inquirer (Stone & Hunt, 1963), SenticNet (Cambria et al., 2010), and AFFIN (Nielsen, 2011). Traditional machine learning models such as Naïve Bayes and Support Vector Machines (SVM) have played essential roles in classification. These methods (Mullen & Collier, 2004; Wilson et al., 2005) utilised feature engineering. Mullen and Collier

(2004) used Turney's features and lemma to conclude that the calculation of Pointwise Mutual Information (PMI) could be supplemented with domain information when searching the web for the context window, assuming that domain information did not reduce the hit count. Wilson et al., (2005) compiled a list of subjectivity clues and expanded it using additional lexicons, including General Inquirer, a dictionary, and a thesaurus. The methodology was based on the prior lexicon-based polarity classifier. This was refined through a two-step process based on intensive feature engineering to distinguish contextual polarity. McDonald et al. (2007) conducted experiments with cascading sentences and document labels. Together, the document and sentences are trained for the classification task. The sentence classification feature space included unigram, bigram, trigram, and POS tags. The inference is performed using the Viterbi algorithm to calculate the document's final score based on the scores of its sentences. Paulus et al. (2014) integrated phrase-level predictions into global belief recursive neural networks to provide feedback to words. This is accomplished by incorporating a backward pass that propagates from the parse tree's root to its leaves. The GB-RNN employs both forward and backward parent nodes, whereas the Bi-RNN employs only forward parent nodes. This method necessitates a parser for the tree structure. In addition to supervised and unsupervised techniques, research also focuses on semi-supervised methods. Read and Carroll (2009) created domain-independent polarity classifiers using word similarity techniques in a semi-supervised setup. The authors described numerous matrices of word similarity. First, the lexical association is calculated using PMI to determine the similarity between two words. Second, semantic spaces represent a collection of conceptually similar words. Last but not least, distributional similarity defines the similarity between two words based on the words in their context. A large unsupervised dataset was utilised to compute the co-occurrence and occurrence frequencies required for the aforementioned matrices. Moraes et al. (2013) compared the performance of SVM and ANN (Artificial Neural Networks). The authors discovered that ANNs statistically outperformed SVM when combined with the information gain-based feature selection method. Nonetheless, the results demonstrated that SVMs are less susceptible to noisy terms in the presence of data imbalance. Several authors (E. H. Huang et al., 2012; Socher et al., 2012, 2013) investigated recursive style neural networks for learning vector representation for a sentence. These methods abandon single-word features in favour of a vector-based strategy. The authors' proposed recursive neural network learns the vector representations of phrases in a tree structure. It assigns a vector and a matrix to each node in a parse tree in order to capture its influence on the surrounding words. The recursive neural tensor network computes higher node representation using leaf-level word vectors.

These procedures utilised parse trees. CNN's semantic modelling of sentences was investigated (Kalchbrenner et al., 2014; Y. Kim, 2014b). The CNNs presented by the author are not parse-tree-based. Utilizing filter pooling operations, relations between discontinuous phrases were captured. In addition to using a single neural schema such as unidirectional LSTM (X. Wang et al., 2015) or bidirectional LSTM (L. Dong et al., 2014), authors have mixed and matched networks such as CNN–LSTM (J. Wang et al., 2016) and CNN and RNN (X. Wang et al., 2016). CNN is used to acquire regional characteristics, while the recurrent network learns the interdependencies between these regional characteristics. These methods consistently outperform feature engineering techniques. During backpropagation, which retrofits these representations for sentiment analysis, the word embedding used as input layers is also fine-tuned. The task-specific knowledge eventually aids during the time of inference. To prevent overfitting, they require an extensive training set.

4.4.2 Sentiment analysis in Slavic languages

Kapukaranov et al. (2015) provided a dataset of movie reviews with fine-grained scores, which was a significant contribution to sentiment analysis in Bulgarian. Georgieva-Trifonova et al. (2018) compiled a dataset containing customer feedback derived from online store reviews. Lazarova et al. (2015) classified movie reviews using a semi-supervised multi-view genetic algorithm. Osenova et al. (2012) described the creation of a corpus of Bulgarian political speeches. The classification of Bulgarian tweets was performed by Smailović et al. (2015). Hristova (2021) provides a concise overview of the text-analytic work in Bulgarian.

Steinberger et al. (2012) created sentiment dictionaries for multiple languages, including Czech, that are multilingual and comparable. Veselovská (2012) compiled a corpus of annotated opinion articles from the Aktualne.cz news website. This was supplemented with supplementary data derived from domestic appliance reviews on the Mall.cz retail website. The dataset of Czech movie reviews was compiled by Habernal et al. (2013). The authors iteratively examined the Maximum Entropy classifier and Gibb's sampling to determine the desired probabilities. Çano et al. (2019) evaluated supervised machine learning algorithms using the Mall.cz and Facebook datasets. Bert-based models for Czech sentiment have also been attempted (Klouda & Langr, 2019; Sido et al., 2021; Straka et al., 2021; Vysušilová & Straka, 2021).

Agić et al. (2010) developed grammar-based rules for determining the overall sentiment of Croatian financial news texts. Agić et al. (2012) have created rule-based techniques for detecting sentiment in horoscopes published on news portal websites. Jakopović et al. (2016) evaluated a lexicon-based method for analysing user comments in the transportation domain. Glavaš et al. (2013) presented aspect-based domain-specific sentiment analysis for the Croatian language. Mozetič et al. (2016) and Rotim and Šnajder (2017) have studied the sentiment analysis of Croatian social media texts. Robnik-Šikonja et al. (2021) compared the Slavic and Germanic language families for the Twitter sentiment analysis task.

Lula Pawełand Wójcik (2011) discussed theoretical and practical aspects of Polish consumer opinions. Haniewicz et al. (2013) presented the first attempt to create a polarity lexicon that is accessible to the public. They utilised readily available resources such as dictionaries, thesauri, and existing open-source initiatives. Other attempts at solving SA in Polish primarily include lexicons (Rybiński, 2017), WordNet features (Zaśko-Zielińska et al., 2015), and unigrams/bigrams (Bartusiak et al., 2015). Numerous authors (Kocoń et al., 2019; Wawer & Sobiczewska, 2019) have compared and contrasted machine learning and deep learning techniques for sentiment recognition, including Naïve Bayes, SVM, BiLSTM, and BERT.

Rules (Kuznetsova et al., 2013), machine learning techniques (Chetviorkin & Loukachevitch, 2013), and deep learning approaches have been described in previous work on the Russian language. Using various neural techniques, Golubev et al. (2020) improved the scores on multiple Russian sentiment datasets. This work posed sentiment classification as a task of natural language inference and improved final scores. Golubev et al. (2021) continued the same work with three-step sequential training and achieved state-of-the-art results. Smetanin et al. (2021) identified multiple datasets and baselines for the sentiment analysis task in Russian.

Machová et al. (2020) translated an English lexicon into Slovak and combined it with a particle swarm optimisation algorithm to construct a lexicon-based sentiment categorization system. Bučar et al. (2016) annotated and evaluated five distinct classifiers for Slovenian web media content. Various attempts have been made at sentiment analysis in Slovenian news texts (Bučar, 2017; Pelicon, Pranjic, et al., 2020; Pelicon, Pranjić, et al., 2020; Žitnik, 2019). The corpus of web commentary was examined by Kadunc & Robnik-Šikonja (2017). Offensive language detection in Slovene (Evkoski et al., 2021; Ljubešić et al., 2021) is an active area of research.

4.4.3 Cross-lingual sentiment analysis

In a cross-lingual multi-task learning setup, Cotterell et al. (2017) performed morphological tagging and language identification by jointly training a BiLSTM with character embeddings. The tagger shared the same tagsets for all languages. Lin et al. (2019) studied optimal transfer language selection but did not include sentiment transfer in their setup.

In the earliest work in cross-lingual sentiment analysis, Mihalcea et al. (2007) utilised resources such as bilingual dictionaries, subjectivity lexicons, and manually translated parallel corpora. Rather than relying on manually translated parallel corpora, Benea et al. (2008) investigated this further with automatic translation and cross-lingual projections of subjectivity annotations. It was observed that translating the target dataset into the source language was the preferred approach to training a classifier with translations of source language data into the target language.

Feng et al. (2019) employed adversarial training and multilingual language modelling. The English and French language representation models were shared, and language-specific decoders, sentiment classifiers, and language discriminators were trained jointly (DVD and books). Earlier cross-lingual sentiment analysis research has focused primarily on translation. In such a scenario, the objective was to translate the target language instances into the source language and perform inference using the source language classifier. The translated instances were also used to train a language tagger with limited resources. Kanayama et al. (2004) introduced the machine translation methodology. Galeshchuk et al. (2019) demonstrated the efficacy of using machine translation systems when there is insufficient data for the target language. These translations necessitate the existence of a reliable translation system. It has been demonstrated that such systems introduce semantic modifications and errors (Lohar et al., 2017, 2018, 2019). Subjectivity indicators used by humans can be lost in translation. Wan (2009) merged two distinct perspectives by using Chinese and English translations for a co-training setup. For the task of bilingual lexicon extraction, Vulic et al. (2013) used language models trained on comparable corpora to identify and extract words with similar meanings. This was based on the theory that two words are identical if their top semantic word responses are identical. In lexicon-based approaches where supervised resources are scarce, such words are crucial resources.

According to Conneau et al. (2020), multilingual pre-trained models utilising shared transformers are superior to shared softmax, shared BPE, and anchor points for cross-lingual representations. Cross-domain sentiment analysis research focuses on the acquisition of shared representations across domains and is closely related to cross-lingual sentiment analysis. Li et al. (2017) performed domain-independent feature extraction using domain classification and sentiment classification. Conditional Domain Adversarial Networks (Long et al., 2018) incorporated multi-linear conditioning of features to enhance the discriminator's performance. Using multi-view representations and a six-layered transformer model with shared encoder and decoder and adversarial training, Fei et al. (2020) aligned data from two distinct languages. The configuration also captures the cross-lingual and cross-domain aspects. The author used Wikitext to train the model. Compared to Romance languages, the model's performance for Japanese was the worst.

Previous research (Cotterell & Heigold, 2017; D. Dong et al., 2015; M. Johnson et al., 2017) has demonstrated that selecting a hub language from the same language family or one that is closer to the target language in the language family tree facilitates knowledge transfer. Dong et al. (2015) utilised fewer instances from Latin languages (French, Spanish, and Portuguese) to improve the performance of machine translation using large parallel English corpora. This also improved performance in the Germanic Dutch language. They did not, however, investigate the correlation with a distant family language. The selection of a transfer language based on the linguistic properties pertinent to the specific task is another important consideration. Lin et al. (2019) identified many heuristics for choosing a transfer language. A few indicators include lexical overlap and the quantity of available training data.

4.5 Data

Our supervised resources include datasets in eight distinct languages, seven of which are official EU languages. We considered English to be the source language for all pairs of languages. Bulgarian, Croatian, Czech, Polish, Slovak, and Slovene are the target languages.

A single dataset was selected for each language in the study. In Table *4.1,* we present the sizes of the datasets' training, development, and test splits.

4.5.1 Sentiment analysis datasets

.

Language	Dataset	Train	Val	Test
Bulgarian	Cinexio	5,520	614	682
Croatian	Pauza	2,277		1,033
Czech	CSFD	63,966	13,707	13,707
English	MARC	200,000	5,000	5,000

Polish	all2	28,581	3,572	3,572
Russian	ROIMP 2012	4,000	260	5,500
Slovak	Reviews3	3,834	661	1,235
Slovene	KKS	3,977	200	600

Table 4.1 Distribution of sentiment analysis datasets.

Language models XLM-Roberta is a language model that has been pre-trained in 100 different languages. Our earlier experiments demonstrated that XLM-R performed better than other pre-trained Slavic language models. We chose this model as the foundation for the procedure of fine-tuning.

4.6 Methodology

Phylogenetic similarity, typological properties, lexical overlap, and the size of the available data all contribute to the final performance of cross-lingual transfer. Lin et al. (2019) posed the selection of optimal transfer languages as a ranking issue. Previous research (McDonald et al., 2011) has demonstrated that a single or multiple similar languages provide adequate performance in languages with limited resources. For the final performance metric, we carefully analysed the various datasets and their presence in the training phase. We examined single-source versus multiple-source transfer in zero-shot and few-shot situations. The following training regimens were implemented: For each study language, a dataset from the target language is:

- used directly to train the model Here, the source language serves as the target language as well (such as Bulgarian).
- combined with a single dataset from a distant language family (such as English).
- combined with a single dataset from a different subbranch of the same language family (such as Russian, Polish, or Czech).
- merged with a number of low-resource language datasets (Croatian, Slovak, and Slovene).

We completed another training session by converting Bulgarian and Russian from Cyrillic to Latin. The datasets were merged with other language-specific datasets.

4.6.1 Model details

Transformer-based neural networks are the current gold standard for classification tasks (Jiang et al., 2020; Thongtan & Phienthrakul, 2019; Wolf et al., 2020). Taking cues from previous work (Thakkar et al., 2021; Wu & Saito, 2017), each of the fine-grained labels (1 (worst) to 5 (best)) and their corresponding coarse-grained labels (positive, neutral, and negative) were treated as two distinct tasks. The model was trained to perform both tasks simultaneously. Not all datasets in our study employ the same annotation scheme. This prompted us to conduct an annotation projection from fine-grained labels (such as 5-star or 11-star ratings) to coarse-grained labels (three-class, i.e., positive, negative, and neutral). Our model is based on the multi-task transfer learning setting (Collobert et al., 2011) for training a sentiment classifier with multiple datasets. The model is a hierarchical network that performs end-to-end training and stacks two classifiers on top of one another. The encoder is shared by all classifier layers.

We framed cross-lingual sentiment classification as a problem of multi-task learning. We aimed to jointly learn a set of neural network parameters for classifiers in the source and target languages. This was accomplished by jointly optimising a loss function that took coarse and fine-grained labels and resources from both languages into account. A transformer-based model fits a parameterized model to maximise the conditional probability of a target label y given a source sentence x, i.e., z = argmax $p(y|x)$ given a training instance x, y. By combining training data from various sources and languages, learning is extended to multiple languages. The objective function we optimised is the sum of the conditional probabilities of different datasets from different languages based on the representations obtained using a shared pre-trained language model.

$$\begin{aligned} \mathrm{L}_{\mathrm{multi}}(\theta) = \\ &\underbrace{\sum_{D_s} \log_{P_\theta}(y^5|x^{L1}, \theta, \omega)}_{1} + \\ &\underbrace{\sum_{D_t} \log_{P_\theta}(y^3|x^{L1}, \theta, \phi)}_{2} + \\ &\underbrace{\sum_{D_s} \log_{P_\theta}(y^5|x^{L2}, \theta, \omega)}_{3} + \\ &\underbrace{\sum_{D_t} \log_{P_\theta}(y^3|x^{L2}, \theta, \phi)}_{4} \end{aligned} \quad 4.1$$

In objective equation 4.1, four distinct loss terms share a common parameter, θ. In addition, language-independent classifiers share the parameters ω and ϕ, which are label-specific. The first and third terms optimise the source language loss for coarse- and fine-grained labels, respectively. Similarly, the second and fourth terms enhance performance in the target language. At two points, parameterisations are performed. First, we modified PLM for sentiment classification in source and target languages jointly. Second, there are two distinct parameters for labels. The global loss function is capable of both cross-lingual and hierarchical classification.

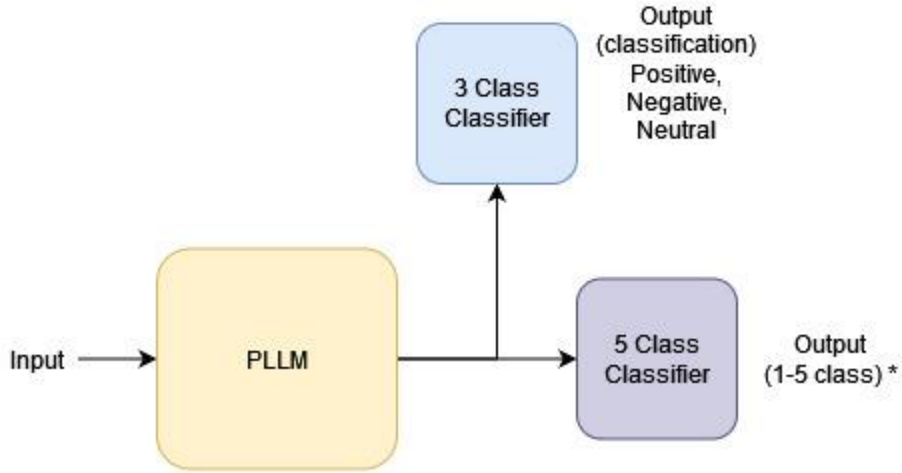

Figure 4.1 Schematic diagram of the neural network.

Consider the training examples x and y, where y is a five-class label (1–5). The labelled five-class dataset is also realisable as a three-class dataset. One and two are mapped to the negative category, three to the neutral category, and four to the positive category, in that order. For a particular training instance, $\langle x, \langle y3, y5\rangle\rangle$, where $y3$ is a three-class label that can be positive, negative, or neutral based on the five-class label, i.e., 1, 2, 3, 4, 5, the objective is to jointly maximise the conditional probability such that instances that belong to the negative class also receive a lower rating and vice versa. This is performed to optimise the model uniformly for various languages and labelled datasets. Two feed-forward neural networks with a softmax output layer define the architecture. Consequently, we have two classifiers trained on the same text but with distinct labels. The first label is coarse, while the second is fine. This is known as "pseudo-multi-task learning" because two tasks are simultaneously trained on a shared representation from a single training instance.

4.6.2 Training

A typical dataset is typically divided into train, test, and validation sets in the ratio of 8:1:3. This partitioning is preferred when the dataset used to train the system is extensive. Low-resource languages have a paucity of examples (of the size of thousands). By separating the few training data samples into test and validation sets, the training set is reduced further; therefore, we conducted cross-validation. K-fold cross-validation randomly divides all dataset instances into K groups, where K is a predetermined number. Folds are used to denote each group. One-fold from K is selected as the test set, while the remaining K-1 subsets are used for training. This process is repeated until each fold in the dataset has been utilised as a test set. The training process will therefore be repeated K times. In our case, K was assigned the value 5.

The most prevalent pattern is the transfer of knowledge to a low-resource language task using data from high-resource tasks. We investigate the use of multiple datasets from low-resource languages to enhance the performance of target languages. We conducted experiments in the subsequent environments:

1. Use only source-language data for fine-tuning. This is the conventional transfer learning setup performed by a source-language fine-tuning classifier. A zero-shot test is administered to the trained model using a test of the target language. We guided the training process using the target language's validation set. We projected labels from a fine-grained class of 5 classes to a coarse-grained class of 3 classes due to the possibility that the target language dataset labels do not match the source language.
2. Fine-tuning with a single source and target language: We sampled training sets from multiple languages and jointly trained the classifier. We utilised datasets from distantly related languages and vice versa.
3. Fine-tuning using multiple datasets derived from a single source and target language. This is a multilingual environment with multiple sources.
4. Fine-tuning with the Latin versions of the Bulgarian and Russian datasets.

In Table *4.2* and Table *4.3*, we list an assortment of experiments. The first section of Table *4.2* depicts the combined instruction of Slavic and English languages. The second section substitutes Russian for English. The third section only utilises language data from a single source. The following section illustrates the compilation of various low-resource languages. The next two sections combine Czech and Polish with Slavic languages with limited resources. Bulgarian is ultimately selected as the source language for the combinations. In Table *4.3*, Latin transliterations of Bulgarian and Russian with other language combinations are displayed that are pertinent to our study.

4.7 Experimental setup

4.7.1 Training details

The model was trained on a 24 GB Nvidia RTX 3090 using the Pytorch-lightning Python library. To ensure reproducibility, the standard techniques for fine-tuning as described by Devlin et al., (2019) were utilised along with a constant seed of 0. We used 10% of the training data as a warm-up alongside Adam optimiser with a 1e-5 learning rate. Each run of 5-fold cross-validation utilised a batch size of eight. The experiment was terminated early when the validation loss did not improve after three iterations. A [CLS] token was extracted from the encoder and passed through a fully connected network (FC1) 3 class softmax layer for each training instance. The encoder's output is routed through a second fully connected network (FC2) and then a five-class softmax layer. The encoder's features were discarded with a probability of 0.2. We sampled a mini-batch from each of the datasets for each training step, namely four English samples and four Bulgarian samples. The instances are then sent through the network, and the error is calculated separately for five and three classes before being summed. To update the parameters, the calculated error is back-propagated throughout the network. In cases where the lengths of the two datasets did not match, the smaller dataset was duplicated to match the larger length. Slovene was the only language for which we performed any pre-processing, replacing user mentions with placeholders and removing all URLs from the text. To compute the effect of having two classifiers, a simple three-class classifier was added to the pre-trained language model as a baseline. No additional hyper-parameters were altered. The results are displayed in Table *4.4*. We observe that cascading classifiers lead to improvements over our baseline in all languages besides Polish.

Source Languages			
1st	**2nd**	**3rd**	**4th**
Bulgarian	English		
Croatian	English		
Czech	English		
Polish	English		
Russian	English		
Slovak	English		

Slovene	English		
Bulgarian	Russian		
Croatian	Russian		
Czech	Russian		
Polish	Russian		
Slovak	Russian		
Slovene	Russian		
Bulgarian			
Croatian			
Czech			
Polish			
Russian			
Slovak			
Slovene			
Croatian	Slovene		
Croatian	Slovene	Slovak	
Croatian	Slovene	Slovak	Bulgarian
Czech	Bulgarian		
Czech	Croatian		
Czech	Slovak		
Czech	Slovene		
Polish	Bulgarian		
Polish	Croatian		
Polish	Slovak		
Polish	Slovene		
Bulgarian	Croatian		
Bulgarian	Slovak		
Bulgarian	Slovene		

Table 4.2 Language pairs combined in various combinations for joint training.

Source Languages		
1st	**2nd**	**3rd**
Bulgarian (Latin)		

Russian (Latin)		
Bulgarian (Latin)	Russian (Latin)	
Russian (Latin)	Croatian	
Bulgarian (Latin)	Croatian	
Russian (Latin)	Slovak	
Russian (Latin)	Slovene	
Bulgarian (Latin)	English	
Russian (Latin)	English	
Bulgarian (Latin)	Polish	
Russian (Latin)	Polish	
Bulgarian (Latin)	Czech	
Russian (Latin)	Czech	
Bulgarian (Latin)	Slovene	Slovak
Russian (Latin)	Slovene	Slovak

Table 4.3 Language pairs with Bulgarian (Latin) and Russian (Latin).

4.8 Results

We conducted experiments on each of the datasets described in Section 4.5 using the methodology described in Section 4.6. We reported on the accuracy and macro-f1 evaluation of five-class and three-class classifications. To verify the performance of one model over the other, we performed statistical testing using the Almost Stochastic Order significance test (Dror et al., 2019; Ulmer et al., 2022) implemented by del Barrio et al. (2018) and the random approximation test (Yeh, 2000). We ran the two tests for each model, as well as the corresponding five and three-class metrics (F1) scores.

We only have three-class scores for Czech, Russian, and Russian (Latin) because the datasets use three-label tagging schemes. Table *4.5* displays the results for the best-performing language pairs that are statistically significant. Each combination that outperformed the others on any of the four metrics has been listed. Bulgarian+English performed best for the five-class metric, while Czech+Bulgarian performed best for the three-class metric. Bulgarian+English and Bulgarian+Czech have no significance over each other for 3-class F1 but are statistically significant over the other combinations such as Bulgarian+Croatian. We did not find any decently performing non-Bulgarian combinations in the top 10 list. In a five-class setting, the combination of Croatian and English produced higher scores. The combination of Croatian+Czech, Croatian+Polish and Croatian+Bulgarian data proved advantageous in a three-class setting. One observation was made that combination of Croatian+Bulgarian performed statistically similar to Croatian+Bulgarian (Latin) data.

The baseline in Czech performs statistically better as compared to all the other cases. Combining Czech with other languages does not help Czech, which is a language with numerous training instances. Similar to Czech, adding other languages to English does not help in a 5-class setting, but we see 3-class F1 having significant improvement with the Czech 3-class data combination.

For the 5-star classification, we see no significant improvement for the combinations of Bulgarian (Latin) + Polish and Russian (Latin) + Polish. For 3-class, the Polish baseline performs better, and it too belongs to a larger data instance; adding more training instances does not help. The combination of other language data with Polish during training leads to a drastic drop in performance. When Russian is combined with either Croatian or Bulgarian, the results are about the same in both metrics, and statistically, this combination does better than others. The combination of Slovak and English performs better for 5-class F1, while Croatian+Slovak+Slovene works best for 3-class F1. For Slovene, a combination with a high-resource language provided better performance compared to the baseline. The best results were obtained when Russian data was converted to Latin script and combined with English data. Russian (Latin) + English, Russian (Latin) + Polish, and Russian (Latin) + Czech perform similarly but statistically significantly better than the other combinations. Although the Latin version of the Bulgarian dataset did not outperform its Cyrillic counterpart. Bulgarian (Latin) + English was the highest scoring combination for the 5-class metric in Latinized Bulgarian. Similarly, for the 3-class metric, Bulgarian (Latin) + English, Bulgarian (Latin) + Polish, and Bulgarian (Latin) + Croatian perform significantly better than others. Three-class metrics for Czech and Polish, two high-resource languages, did not improve from their initial scores.

In the case of Polish and Czech, the addition of the English dataset had no positive effect. While all other languages, i.e., Bulgarian, Croatian, Russian, Slovak, and Slovene, had improved performance with a large English dataset, we noticed that combining Slavic languages had slightly lower performance than English combinations. It was also observed that combining multiple languages (such as Bulgarian, Croatian, Slovene, and Slovak) did not outperform the five-class metric. A further observation is that, with the exception of Russian (Latin) and Slovak, none of the model combinations that scored over 80% on the three-class F1 value utilised English during their training phase. The performance of Bulgarian (Latin) is inferior to that of the Cyrillic version. In contrast, Russian (Latin) achieved the highest scores in each of the four metric classes. This may be due to the lack of train data in Bulgarian. When the languages are combined with English, it has resulted in superior performance in the majority of instances. Slovene was found to be the dataset/language with the lowest performance. This is because the Slovene dataset is derived from informal sources, such as news commentaries, which are noisy in nature. The Slovak dataset has fewer examples for training than the Slovene dataset, but these examples come from customer reviews.

Language	**Accuracy-3**	**F1-3**
Bulgarian	67.80(0.0076)	69.42(0.0046)
Croatian	62.37(0.004)	57.47(0.0053)
Czech	83.82(0.0037)	83.76*(0.0033)
English	68.15(0.0076)	67.85(0.0100)
Polish	87.70(0.0033)	87.57*(0.0039)
Russian	71.43(0.0013)	70.20(0.0030)
Slovak	81.60(0.0057)	79.75(0.0017)
Slovene	59.13(0.0180)	59.97(0.0307)

Table 4.4 Baseline three-class classification scores are averaged over 5-fold runs. The standard deviation is presented in the brackets to the right. * indicate no significant improvement was achieved when combined with other languages during training.

Target Language	Source Languages	5 class Accuracy	5 class F1	3 class Accuracy	3 class F1
Bulgarian	Bulgarian English ⊞	**53.37** (0.0123)	**54.60*** (0.0097)	72.73 (0.0142)	**74.22*** (0.7422)
Bulgarian	Bulgarian Czech	52.18 (0.0070)	53.14 (0.0106)	**72.79** (0.0098)	74.11** (0.0081)
Croatian	Croatian English ⊞	54.12+ (0.0186)	**53.80*** (0.0163)	74.07 (0.0121)	74.12 (0.0097)
Croatian	Croatian Czech	50.88 (0.0094)	50.12 (0.0251)	74.69 (0.0107)	**75.82*** (0.0106)
Czech	Czech Croatian			82.29 (0.0035)	82.24 (0.0036)
English	Czech English	56.22 (0.0099)	55.36 (0.0123)	**69.09** (0.0035)	**69.06*** (0.0043)
English	Bulgarian (Latin) English ⊞	56.91 (0.0031)	56.78 (0.0042)	68.36 (0.0086)	68.05 (0.0103)
Polish	Bulgarian (Latin) Polish	52.34 (0.0017)	52.28 (0.0012)	87.05 (0.0028)	87.15+ (0.0016)
Polish	Russian (Latin) Polish	52.19 (0.0010)	52.15 (0.0005)	86.92 (0.0016)	87.00+ (0.0007)
Russian	Bulgarian Russian			71.84 (0.0035)	71.31 (0.0022)
Slovak	Slovak English ⊞	**68.87** (0.0351)	**68.03*** (0.016)	83.51 (0.0182)	82.14 (0.0076)
Slovak	Slovak Croatian Slovene	64.47 (0.0135)	58.71 (0.0441)	**85.36** (0.0046)	**83.44*** (0.0064)
Slovene	Slovene English ⊞			**69.52*** (0.0203)	68.97* (0.0154)
Slovene	Slovene Czech			68.24* (0.0084)	**69.56*** (0.0078)
Bulgarian (Latin)	Bulgarian (Latin) English ⊞	50.73 (0.0094)	51.76 (0.0075)	70.30 (0.0093)	72.01 (0.0071)
Russian (Latin)	Russian (Latin) English ⊞			88.14* (0.0299)	**87.95*** (0.0290)

Table 4.5 Language pairs with Bulgarian (Latin) and Russian (Latin). + shows that ACO detected it as significant, but the permutation test rejected it.

Language	Metric	5 class	3 class	2 class
Bulgarian	MSE	0.666	0.141	
Croatian	F1			91.1
Czech	F1		87.08 ± 0.11	96.00 ± 0.02
English	ACC	56.5		
Russian	F1		72.69	87.04
Slovak [7]	F1		81.5	
Slovene	F1		65.7	

Table 4.6 Previously reported results for the languages in the study. ACC- Accuracy.

4.9 Analysis

Multiple top-performing models were selected, and their test-set predictions were analysed. In addition, we investigated how various languages are represented in the shared encoder.

4.9.1 Error analysis

We calculated the confusion metric for each fold for all models with the highest performance. It was observed that the Bulgarian+English model for target Bulgarian incorrectly classified a greater number of neutral and negative instances. The same effect occurs when predicting five classes, where zero to two classes are incorrectly predicted. The number of neutral and positive classes was overestimated by the Slovak+Slovak+Croatian model. In the scenario involving five classes, labels for classes two and three were exchanged. The negative instances were assigned to the neutral and positive classes by the model trained in Czech and Bulgarian. The neutral instances were incorrectly categorised as negative and (mostly) positive, followed by the negative class drifting into neutral. The Czech with Croatian training performed the best in two cases, namely Croatian and Czech. The negative class instances in Czech miscategorized into the neutral class. The neutral category instances

[7] http://arl6.library.sk/nlp4sk/webapi/analyza-sentimentu

were misclassified as negative and positive. The same can be said of Croatian. In Slovene, the negative was predicted to be neutral or positive. The neutral and positive comments were grouped with the negative ones.

4.9.2 Language representations in XLM-RoBERTa

The training setup consists of three components: the shared encoder, training data, and classifier heads. Using the training data, the classifier heads are trained. The shared model is a black box component of the entire system for representing multiple languages. The XLM-RoBERTa model was trained using 2.5T data from 100 languages. The various training dataset sizes for the languages under study are listed in Table *4.7*. The text is divided into tokens using a sentence-piece tokenizer. We conducted a simple study to examine these representations in different languages. We ran each dataset's training set through the XLM-R tokenizer. For the obtained sentence-piece tokens, we calculated the intersection of all possible language combinations. Table *4.8* indicates the number of common tokens in various languages. We observed that the best-performing language combinations have many shared tokens for a given target language's sentence fragments. In the case of Croatian, it shares 5,075 sub-tokens with Czech, allowing it to advance under the joint-training system. We would like to note that the Slovene dataset is comprised of comments from a news website and is therefore highly informal and noisy. Consequently, we hypothesise that, when combined, it adversely affects the Croatian performance metric. The performance of the Czech language decreases when it is combined with other languages. When English is combined with Czech, we observe a slight improvement over the baseline and other combinations. Russian (Latin), Bulgarian (Latin), and English combinations have higher scores for Polish. In the case of Slovak, training alongside Czech led to results that were comparable to those obtained with Slovak, Slovene, and Croatian combined. The Czech and Slovene shared the second-greatest number of tokens. Consequently, we hypothesise that sub-word token sharing indirectly influences the classification procedure. We can, therefore, assume that some dataset combinations belong to the same language family. In addition, distant high-resource languages (such as English) do not aid in the improvement of the performance of high-resource languages. Adding English data improves performance in five classes, whereas adding same-family language data improves performance in three classes. Although we observe that Bulgarian shares a large number of sub-words with Russian, Czech, and English, the languages with the most shared tokens, the precise classification behaviour of tokens requires further investigation.

Language	**Size** (Gb)	**Tokens** (Million)
Bulgarian	57.5	5,487
Croatian	20.5	3,297
Czech	16.3	2,498
English	300.8	55,608
Polish	44.6	6,490
Russian	278.0	23,408
Slovak	23.2	3,525
Slovene	10.3	1,669

Table 4.7 Data size used for training XLM-Roberta.

Languages	Croatian	Czech	Polish	Russian	Slovak	Bulgarian (Latin)	Russian (Latin)	Slovene	English
Bulgarian	130	**235**	90	**2,919**	123	261	126	122	**215**
Croatian		**5,075**	**2,881**	432	2,215	1,778	3,014	4,420	3,256
Czech			9,656	1,300	6,035	3,573	8,733	**10,075**	**15,122**
Polish				690	2,927	**2,207**	**5,075**	5,417	6,931
Russian					3,71	314	1,522	733	1,207
Slovak						1,616	2,923	3,412	**2,774**
Bulgarian (Latin)							2,689	2,655	**2,416**
Russian (Latin)								5,799	5,702
Slovene									**6,352**
English									

Table 4.8 Languages and number of shared tokens on their train set.

4.10 Conclusion

We have presented our framework for multi-task cross-lingual sentiment classifier transfer. We evaluated seven official Slavic languages using a model trained with multiple language resources. We discovered that the transfer of sentiment knowledge is enhanced within the same language family, i.e., the closer the language, the easier the transfer, given a large dataset. We also discovered that a large training dataset from a distant language family can outperform smaller datasets from similar languages. Consequently, datasets from the same language family and distant language families can be utilised to combat the issue of inadequate resources.

5. DATA AUGMENTATION

5.1 Introduction

In the previous chapter, we examined how the combination of diverse datasets affects sentiment classification. The cumulative effect of utilising resources from languages of the same and distant family was examined. Moreover, we determined the optimal combinations for a specific Slavic language. In this chapter, we will examine various techniques for augmenting data and how they can be applied to sentiment analysis. Recent developments in the field of NLP have resulted in breakthroughs that surpass the previous state-of-the-art. This has been attributed primarily to deep neural networks. To achieve the best results in the field, however, a large number of data points are required. They primarily work in the field in which they were trained, so they cannot be utilised in other fields, consequently needing domain generalisation algorithms (J. Wang et al., 2021).

The performance of the neural network is completely dependent on its hyperparameters and the training set-learned parameters. It is commonly believed that having more data points is the default method for improving performance. A direct approach requires running an annotation campaign, which is expensive, time-consuming, and labour-intensive in terms of annotation and training. Because these models rely on large parameters that necessitate a large number of training instances to perform the intended task, this requirement cannot be eliminated. In the reverse direction, new data points are generated from existing supervised or unsupervised text bodies. To date, numerous techniques for data generation have been identified. Kobayashi (2018) reported using contextual language under the assumption that sentences are invariant when original words are replaced by words with paradigmatic relations. When compared to original texts, in-context predicted words were deemed to be better options for creating data samples that vary in terms of pattern. Attempts have also been made at using data augmentation for different text classifications in large English-language datasets (J. Wang et al., 2021; Zhang et al., 2015). The augmentations were derived from an English thesaurus and then trained using various machine learning and deep learning algorithms. Wei et al. (2019) described simple augmentation operations (such as insertion, deletion, swap, and replacement) that produced comparable results when only half of the original dataset was used. In data-driven research, these techniques focus primarily on resolving low-data scenarios, mitigating the phenomenon of class imbalance, or serving as regularising terms to make systems more resistant to adversarial attacks.

Existing data augmentation strategies for other tasks in languages with abundant resources (especially English) have also been investigated. To detect event causality, Zuo et al. (2020) employed a remote annotator, followed by filtering, relabelling, and annealing on instances with noisy labels. For the common-sense reasoning task, Yang et al. (2020) used a pre-trained task model (XLM-R) and a generative language model (GPT-2) to generate synthetic data instances. Data selection was conducted using filtering functions that considered the quality and diversity of synthetic instances. The reported methods for languages with abundant linguistic resources are founded on linguistic resources. For a method to generate facts from Freebase, for instance, such a resource must exist in the target language. Therefore, a language with limited resources may lack these dependent resources, thereby rendering the method inapplicable. Empirical evidence regarding the effectiveness of these interventions in low-resource settings is still lacking. Even though data augmentation techniques such as EDA (Wei & Zou, 2019) are simple to implement, it is essential to conduct additional research on their applicability in low-resource settings.

This chapter compares data augmentation as a means of enhancing the performance of sentiment analysis for languages with limited resources. We hypothesise that DA strategies are equivalent to cross-lingual and cross-family configurations. For the task of sentiment classification, we experiment with various data augmentation techniques on a set of low-resource languages from the same language family (i.e., South Slavic languages). To analyse each of these facets, we employ three distinct data augmentation techniques that rely on synonymy (Miller, 1995) and pre-trained large language models (T. Brown et al., 2020; Vaswani et al., 2017). In addition, we propose a straightforward method of augmentation that requires no additional resources. To determine the effectiveness of these techniques on datasets with limited resources, we classified sentiments using them. With limited resources, experiments were conducted on South-Slavic languages (i.e., Bulgarian, Croatian, Slovak, and Slovene). To enable a three-class classification of the dataset for the Croatian language, we also conducted an annotation campaign to label instances that were claimed to be noisy by the original authors of the dataset.

Our findings indicate that augmentation methods do not contribute directly to sentiment classification. We find that the performance of augmentations based on pre-trained contextualised language models is inferior to that of methods constructed by combining multiple datasets from the same and different languages. Indirectly affecting the final classification score are factors such as noisy text and code-mixing. In addition, we find that WordNet-based augmentations are more effective than those based on the Masked Language Model or Causal Language. In seven instances, the expansion-permutation-combination technique resulted in an improvement.

5.2 Research question

Empirically, we pose the following question for our proposed study:
Q. Can data augmentation be utilised effectively for sentiment analysis in low-resource languages?
We hypothesise that the accuracy of data augmentation techniques is comparable to that of supervised methods when applied to typologically related languages.

In this study, we explore data augmentation methods as a means to artificially increase the instance space and compare the performance with that of using resources from the same language family. Some additional questions that we pose in this study concerning data augmentation are as follows:

1. Can the data augmentation technique improve the performance metric?
2. What is the effect of having augmented data generated from different techniques?
3. Can WordNet-based augmentation techniques work better with sentiment classification tasks?
4. Does training with Lemma-based instances work for Croatian?

5.3 Related work

5.3.1 Data augmentation

Distant supervision is a method for curating labelled data instances by utilising an existing knowledge base (Su et al., 2019). Mintz et al. (2009) reported the first instance of using distant supervision in NLP. The work entailed curating datasets for the task of relation extraction. The authors used Freebase, a large database that stores the relationships between two entities. The assumption was that any sentence containing two freebase entities could express the relationship. As a result, Freebase was used as an unsupervised lookup table. Various features were designed, ranging from POS tag, NER, and n-words within the context window. Su et al. (2019) introduced a similar approach in the BioNLP domain, in which knowledge from a database is used to label sentences containing two entities to generate a dataset based on remote supervision. In the same work, heuristics (trigger words and high confidence patterns) were proposed to reduce noise in the sentence augmentation process. A CNN trained with an automatically created dataset and then trained on a manually annotated dataset achieved the highest score. The authors hypothesised that the direct union of two datasets (distant supervision-based and manually annotated) is not advantageous because noisy datasets lead to a decline in the final performance.

Two types of augmentation methods for NLP can be distinguished broadly: 1) text-based augmentation, and 2) feature-based augmentation. The text-based enhancements operate at the text level. The process of augmentation can be implemented at various linguistic levels (morphological, syntactic, and semantic). Another branch of research focuses on adversarial attacks against the trained model. This is accomplished by generating text instances X' similar to the training data X such that the model attempting to perform the intended task fails. Instances X and X' should have identical human predictions, with X' containing minimal textual changes relative to the original instance. All adversarial attack techniques (Garg & Ramakrishnan, 2020; L. Li et al., 2020; Yoo & Qi, 2021) on classification tasks rely on text-augmenters as their primary component for supplying augmented instances for adversarial attacks.

Ren et al. (2019) experimented with various synonym replacement methods to generate adversarial samples. The synonyms were obtained from WordNet. The method for choosing a synonym for a word ranged from random selection to a more sophisticated method based on Word Saliency (Samanta & Mehta, 2017) score. Another way of finding a replacement for a given word is to use a pre-trained language model that uses context to predict the replacement word. Kobayashi (2018) altered the language model so that it integrates the label in the model along with the context during the word prediction stage. The language mode was trained on the WikiText-103 corpus of English Wikipedia articles. Garg et al. (2020) used contextual perturbations from a BERT masked language model to replace and insert tokens at masked locations. D. Li et al. (2021) extended the work using RoBERTa and three contextualised perturbations, i.e., replace, insert, and merge. All of these studies have been published on the basis of English datasets.

In the field of Neural Machine Translation (NMT), the technique of translating a target language into a source language is known as back-translation (Sennrich et al., 2016). The ultimate goal of this procedure is to increase the number of samples in the source language by translating the target language text obtained from the source language and the translation module back into the source language, thus generating paraphrases of the original text. The final system is trained using both the parallel synthetic corpus and the original training data. Although back-translation is an easy-to-use technique, it necessitates the training of a machine-translation model for low-resource languages, which may not be a viable option given the required volume of data. Edunov et al. (2018) showed through experiments that sampling and noisy beam outputs (delete, swap, and replace words) are better for making fake data than pure beam and greedy search.

Wei et al (2019) introduced EDA (Easy Data Augmentation), a set of augmentation techniques consisting of multiple processes including synonym replacement, random replacement, random swap, and random deletion. On five distinct datasets, the processes were executed and benchmarked. The authors conducted experiments with an augmentation parameter named $alpha$ whose values ranged from [0.05, 0.1, 0.2, 0.3, 0.4, 0.5] and discovered that small $alpha$ values provided greater gain than large values.

The same work was expanded by Longpre et al. (2020) to include two additional datasets for examining the impact of data augmentations using pre-trained language models (BERT, XL-NET, and ROBERTA). EDA and back-translation are two task-independent data augmentation techniques. According to reports, data augmentation methods do not provide any consistent improvement for pre-trained transformers. The authors attributed this phenomenon to large-scale, unsupervised, domain-spanning pre-training, although all datasets utilised in the study were English-based.

Consistency training is based on the premise that small changes or noise in the input should not impact model predictions. Xie et al. (2020) used data augmentation in place of noise signal to enforce consistency constraints during training. The overall loss consisted of classification loss and consistency loss between the original input and the enhanced version of the same. The consistency loss is only computed for instances in which the model has high confidence. The author used back-translation, RandAugment (for image classification), and TF-IDF word replacement for augmentations. A data filter within the domain was implemented to prevent domain mismatch.

Go et al. (2009) proposed the first method for classifying the sentiment of tweets using emoticons as remote supervisors. The technique was based on the premise that the emoticons "*:)*" and "*:(*" (and their variants) are poor indicators of positive and negative emotions. Therefore, each tweet containing these emoticons was tagged with their respective classes. There was an assumption that the statements in Wikipedia and newspaper headlines were neutral. The neutral class was not classified because it had no emoticons associated with it. The dataset was used to train the machine learning algorithms Naïve Bayes, Maximum Entropy (MaxEnt), and Support Vector Machines (SVM). The entire setup was studied using English as the study language.

Martinc et al. (2022) compared multiple data augmentation strategies (such as WordNet and Bert-based) for the generation of news headlines in Croatian, Finnish, and English. In addition to ROGUE, the authors employed two additional methods to assess the performance score. One technique was the computation of semantic similarity using a sentence transformer trained in the task of paraphrasing. The second method employs a metric based on natural language inference to quantify the similarity between the original and generated headlines. The authors did note that there was no NLI model covering Croatian and Estonian.

The other branch of data augmentation focuses on the latent space directly. Training as a whole is intended to add new latent information without altering the original class representation. This enables inducing difficult-to-input semantic cases with limited training data. Cheung et al. (2021) proposed that difficult-to-classify samples are the best candidates for data augmentation because they contain more information. Latent space augmentations were created using interpolation, extrapolation, noise addition, and the difference transform.

Previous research indicates that sentiment analysis with augmented data for low-resource languages has received little attention. As the grammars of these languages are not simple, and their morphology and inflection systems are complex, the situation is further complicated.

5.4 Data

We used sentiment classification datasets to answer our research questions, employing the data structure from the previous chapter. We targeted only low-resource languages in our experiment: Bulgarian, Croatian, Slovak, and Slovene. A single dataset was selected for each language in the study. In Table *5.1,* the sizes of the original training, development, and test dataset splits are displayed.

Language	**Dataset**	**Train**	**Val**	**Test**
Bulgarian	Cinexio	5,520	614	682

Croatian	Pauza	2,277		1,033
Slovak	Reviews3	3,834	661	1,235
Slovene	KKS	3,977	200	600

Table 5.1 The original distribution of sentiment analysis datasets.

5.4.1 Croatian dataset re-annotation

The authors of the Pauza dataset (Glavaš et al., 2013) eliminated reviews with a rating between 2.5 and 4.0 because these reviews were noisy. Therefore, ratings below 2.5 are considered negative, whereas ratings above 4.0 are considered positive. The reviews with ratings ranging from 2.4 to 4.0 have instances where the text is positive but has ratings that might tag it as a positive instance, and vice versa. We hypothesise that this might lead to the semantic drift, where the model might learn to classify instances incorrectly. Another major problem we will encounter will be in the augmentation step. In our methodology, we artificially augment the data using various methods, and a text with conflicting labels when augmented into large numbers, might lead the model away from learning. Therefore, we take up the activity of re-annotating our Croatian dataset.

We re-evaluated the ratings between 2.5 and 4.0 and asked three native speakers to annotate particular instances. Annotators were asked to classify the given text as positive, negative, or neutral/mixed. Only two annotators managed to complete the annotation of all the provided instances. Training and test sets were utilised in the annotation campaign. Those instances devoid of consensus were eliminated through filtering. Consequently, we have two datasets with re-annotated mixed instances: binary and ternary. The general inter-annotator agreement in terms of Cohen's kappa stands at 0.70.

“dostava brza, ćevapi solidni....” 4.0 the dataset has this entry, and if we follow the 2.5-4.0 logic, then it gets tagged as neutral, which might lead the model into the semantic drift.

The original dataset had around 224 instances of reviews which had ratings between 2.5 to 4.0, out of which 146 were retained as mixed/neutral, 50 were tagged as negative 19 were tagged as positive. The negatively tagged reviews spanned a wider range of star ratings (2.0 to 4.0).

“Sad je mješana pizza bila loša.” This was initially tagged with a rating of 3.0.

There were 9 instances of the text that were not included in the final set as the annotators had not reached collective agreement on these. Two instances from this lot include reviews where the users gave individual ratings to various aspects of the food and delivery.

The general pattern that can be seen for an almost positive review to get a 4.0 is the missing of an item or negative orientation towards an aspect of the order.

Similarly, for the test set, 115 ratings ranged between 2.5 to 4.0 score, 31 as negative ranged from a score of 2.5 to 3. 6 instances as positive and the remaining 78 were tagged as mixed/neutral.

5.5 Methodology

Two sections comprise the overall methodology: data generation and model training. First, we use tools for natural language processing and data augmentation to create samples of data. Then, we use the samples to train a transformer-based classification model on the data.

5.5.1 Data generation and augmentation

To answer the questions posed in earlier sections, we utilise four simple language processing techniques and three existing data augmentation methods. The aforementioned existing data augmentation strategies are used in adversarial attacks against trained classification models and can be utilised to obtain samples that are more semantically similar to the original dataset. Next, we describe the individual techniques for augmenting data and the overall procedure for augmenting and training the classifier.

- $Data_{lemma}$ based on lemmatisation.
- $Data_{expanded}$ based on sentence tokenisation.
- $Data_{expand-combined}$ based on sentence tokenisation.
- $Data_{expand-permuted}$ based on sentence tokenisation.
- WordNet (Ren et al., 2019).
- Masked Language Model (MLM) based Clare (D. Li et al., 2021).
- Causal Language Model (CLM)- based Generative Pre-trained Transformer (GPT)-2 (Anaby-Tavor et al., 2020)

5.5.2 Lemmatisation

By performing morphological analysis, the lemmatisation process returns the word's morphological base. The output is the canonical form of the original word. Since South-Slavic languages are rich in morphology, we decided to create a lemma-form variant of the original dataset. Previous studies (Bollegala et al., 2011; Gamon, 2004) had fed lemmas into machine learning classification algorithms as input features (such as Support Vector Machines and Random Forests). Transformers-based PLMs employ byte-pair encoding to reduce the vocabulary size, which is required to avoid sparse vector representations of the input text.

For instance, the word *running* is converted to **run** + **##ing** and the neural network learns to weight individual byte-pairs based on the dataset and the task's requirements. Therefore, the affixes may be useful for the task that takes into account the additional information. But this requirement has not been investigated in PLMs with languages that are rich in morphology or for sentiment analysis in particular. We made a lemmatized version of the original dataset to see how lemmatization affects the final performance of a language model that has already been trained.

- **Original HR**: super, odlicni cevapi.
- **Lemmatised**: super, odličan ćevap.

5.5.3 Expansion

Every labelled instance D^i from the train-set, i.e., document or text, consists of one or more sentences $D^i_{1..n}$ and a single instance $Di \in L$, where L can be negative, negative, or neutral/mixed.

$$D^i = D^i_{1..n} \tag{5.1}$$

$$D^i \in L \tag{5.2}$$

$$D^i_{1..n} \in L \tag{5.3}$$

$$D^1 D^2 D^n \in L \ \Rightarrow D^i \in L \tag{5.4}$$

From equation 5.4, it follows that each of the sentences ($D^1 D^2 D^n$) of a single training instance D^i can be weakly assumed to be labelled with the same class L. Therefore, every sentence from a review can be individually treated as a new labelled instance. For example,

- **Original HR**: “*Pizze Capriciosa i tuna, dobre. Inače uvijek dostava na vrijeme i toplo jelo”*.
- **Translated EN**: “*Pizza Capriciosa and tuna, good. Otherwise always delivery on time and hot food*”.

This example belongs to the positive class, and individual sentences may be treated as reviews of the positive class. Theoretically, this assumption may hold true for extremely polar classes, such as positive and negative, but it may fail for classes that are mixed or neutral. In practice, we are also presented with instances in which the service was poor, but the reviewer still awarded a high rating due to previous positive experiences.

5.5.4 Expansion-combination

Based on the previous technique for expansion, we propose a straightforward extension. Assuming that all individual sentences from all reviews for a given class also belong to the same parent class, we can now create a brand new dataset by randomly sampling from this set of individual sentences. Here, we consider the entire $D^i_{1..n}$ range to be the universal set. We obtained the new dataset by sorting the instances using combinations as denoted by the mathematical equation 5.5 For a more intuitive explanation, assume ABCD to be four positive sentences from various positive reviews. Combination ordering produces a new sampled dataset represented by combinations (**’ABCD’**, 2) –> **AB AC AD BC BD CD**. "Elements are treated as unique based on their position, not on their value. So if the input elements are unique, there will be no repeat values in each combination" (itertools combinations, 2022). This indicates that AB and BA will not be present in the final sampled dataset.

$$C^n_k = \frac{n!}{k!\,(n-k)!} - \text{combination} \tag{5.5}$$

5.5.5 Expansion-permutation

$$P^n_k = \frac{n!}{(n-k)!} - permutation \tag{5.6}$$

We also propose a second simple method that replaces previous combination sampling with a permutational process. Mathematically, this is denoted by equation 5.6 in which the universal set of individual sentences belonging to a single class can be combined as depicted by permutations (’**ABCD**’, 2**)—> AB AC AD BA BC BD CA CB CD DA DB DC**. According to the order of the input iterable, the permutation tuples are returned in lexicographic order. Therefore, if the input iterable is sorted, the output combination tuples will also be sorted. "Elements are treated as unique based on their position, not on their value. So if the input elements are unique, there will be no repeat values in each permutation" (itertools permutations, 2022). In other words, AB and BA will represent two distinct instances of the generated dataset.

5.5.6 WordNet augmentations

WordNet (Erjavec & Fišer, 2006; Koeva et al., 2004; Miller, 1995; Raffaelli et al., 2008) provides a straightforward formal synonym model for locating replacement words in context. This method replaces each word in a given text with its synonym. The assumption that a word's synonym will not affect the polarity of the given instance makes this one of the most straightforward data enhancement techniques. Synonyms are derived from synsets by querying WordNet with candidate keywords. The synset includes words with equivalent meanings. Notably, the word being searched may belong to multiple synsets, necessitating additional processing such as word-sense disambiguation to prevent incorrect synset selection[8].

- **Lemma HR***: jako dobar pizza*. (**Translation**: *very good pizza*)
- **Augmented HR**: *jako divan pizza.*
- **Augmented HR**: *jako krasan pizza.*

Here the word *dobra* ("*good*") has been replaced with its synonyms *’divan’* and *’krasan’*. WordNet entries are in lemmatized form, which is an important detail to note. Therefore, in order to obtain more results for the words in context, they must be lemmatized. The lemma can then be used to retrieve the synonym set. The results retrieved are also in lemma form. Although this is not a necessary condition, we can still obtain a significant number of terms to replace the words in the dataset. This is illustrated by the following examples:

[8] Due to the limited resources available, we did not pursue more sophisticated synset selection

- **HR**: *jako dobra pizza i brza dostava.* (**Translation**: *Very good pizza and fast delivery*).
- **Augmented HR**: *Jako dobra pizza i brza dostavljanje*.
- **Augmented HR**: *Jako dobra pizza i brza doprema.*

In order to prevent the semantic drift, no additional relations were employed.

5.5.7 MLM augmentations

CLARE (ContextuaLized AdversaRial Example) (Li et al., 2021) is an adversarial attack text generation technique. In this method, each word in the given sentence is greedily masked, followed by an infill procedure to obtain a replacement word for the masked word. The method permits data enhancement through replace, insert, and merge operations. This method is greedy in nature, as it replaces all the words in a sentence with substitutes. This typically results in augmentations with a different semantic meaning than the original, so it relies on multiple constraints to generate meaningful data. These constraints eliminate enhancements that do not meet the given criteria. Checking the semantic similarity of the augmented sentence with the original input using an existing process is one of these constraints. Using a neural network already trained on sentence similarity, cosine distance can be used to compute the semantic similarity in its most basic form. For computing similarity between the encoding of original sentences and augmentations, the authors utilised the Universal Sentence Encoder, a text encoder model that maps variable-length English input to a fixed-size 512-dimensional vector. In addition to the encoding model, there are dataset-dependent parameters such as minimum confidence, window size, and maximum candidates. We chose only the Replace method to prevent the semantic drift caused by random deletions and insertions.

- **HR**: *Ne narucivat chilly*. (**Translation**: *Do not order chilly*).
- **Augmented HR**: *Ne narucivat meso*. (**Translation**: *Do not order meat*).

5.5.8 CLM augmentations

Language generation tasks are competitively performed by causal language models such as GPT-2. During training, the model is tasked with predicting the next word in a text sequence. This causes the model to generate the next suitable word based on the previous words or context. During the inference stage, a model is fed an initial prompt and instructed to predict the next word. The entire procedure can be easily used to generate training resources for a model. This method was reported by Anaby-Tavor et al. (2020) using a small, supervised English dataset. Typically, a single model is trained with data from multiple classes in such a way that the generated text depends on the label. For instance, to generate a positive review, we instruct the model during training with the start token, class label, and text (i.e., ’<|**startoftext**|> |**review pos**|> **WHOLE TEXT** |**endoftext**|>’). During the inference, only a few initial words (such as ’|**startoftext**|> |**review pos**|> **PROMPT-TEXT**’) are needed to produce the entire text. Using a single model to generate data for all classes with a large amount of data is possible. After training in this environment, we noticed that the model began to generate negative reviews for the mixed/neutral class. Consequently, we trained three distinct models for each of the individual classes. Due to the fact that each class has its own model, the model can only generate text for the class in question. Since they are the ones discussed in reviews, we decided to use nouns as prompts to capture the overall context during the generation process; typically, it is food, such as pizza or risotto, or a service, such as delivery. Using morphosyntactic (MSD) tags, we extracted all nouns from the dataset. The nouns were manually inspected for pipeline-annotated false positive artefacts. The nouns obtained were then used as inputs for the three fine-tuned GPT-2 models to generate datasets.

- **HR**: *naručili salatu, dostava je bila na vrijeme, dostavljac simpatican.*
- **translation**: *ordered a salad, the delivery was on time, the delivery guy was nice.*

5.6 Experiments

Using a transformer-based classifier, we compared the efficacy of various data generation methods. Two distinct dataset versions were created: 2-class, which is the binary version (positive and negative), and 3-class, which is the ternary version (positive, negative, or neutral[9]). Using the various training sets, the parameters of entire networks were optimised. We trained a separate model for each language in the study and for each dataset generated using the previously described methods (including the original dataset), while maintaining the same network parameters. In cases where the dataset was unbalanced, class weight was computed using labels from the training set and used as a rescaling weight parameter in the cross-entropy loss. This allows for a greater penalty if a class with lesser number of instances makes an incorrect prediction. We trained the model with a learning rate of 1e-5, weight

[9] We refer to the class as neutral despite the fact that it consists of both positive and negative elements.

decay of 0.01, early stopping on validation loss, and a patience of four to five epochs. Utilizing the softmax classifier, the class probabilities are calculated. The final scores for the original set of manually administered tests associated with the dataset are reported.

5.6.1 Language tools

Each dataset for each of the four languages was required to undergo tokenisation, part of speech extraction and lemmatisation. The Classla[10] library was used for processing Bulgarian, Croatian, and Slovene, while the Stanza[11] library was utilised for Slovak[12]. We used the tokenised and lemmatised data to generate the lemmatised ($\mathrm{Data_{lemma}}$) and expanded ($\mathrm{Data_{expanded}}$) versions of the dataset. The expanded version was converted into $\mathrm{Data_{expanded-}}_{combined}$ and $\mathrm{Data_{expanded-}}_{permuted}$ by combining two individual sentences into a single training instance via sampling.

5.6.2 Data augmentations

5.6.2.1 WordNet

To reimplement a custom WordNet augmentor for each of the languages (Bulgarian, Croatian, Slovak, and Slovene), we used the textattack[13] library and derived a new class from the Augmentor[14] base class. In the augmentor, we introduced constraints to prevent modifications to stopwords and words that have already been modified. Based on the recommendation reported by Wei et al. (2019), the pct-words-swap parameter (i.e., percentage of words to swap) was set to 0.05, limiting the number of words to be replaced with synonyms. The number of augmentations per instance has been set at 16. We used Open Multilingual WordNet[15] to find replacements for synonyms.

5.6.2.2 Masked language model

Initially, we compared each augmentation to the original sentence using a second pre-trained language model. The authors suggested using the Universal Sentence Encoder, a pre-

---

[10] https://github.com/clarinsi/classla
[11] https://stanfordnlp.github.io/stanza/
[12] https://huggingface.co/stanfordnlp/stanza-sk
[13] https://github.com/QData/TextAttack
[14]textattack.augmentation.WordNetAugmenter
[15] http://compling.hss.ntu.edu.sg/omw

trained language model, to compute the similarity between the encoding of original sentences and augmentations. The Universal Sentence Encoder[16] has been trained in 16 languages, but none of them is South-Slavic; as a result, it is not a good candidate for encoding our data. Consequently, we utilised LaBSE[17], which has been trained in 109 languages. We used cosine scores as a similarity measure and eliminated all sentences that had a cosine similarity of less than 0.80. This was done to obtain augmentations that have the same class label as the original sentence due to their similar meaning. We implemented a custom MLM-CLARE augmentor with the constraints using the CLARE[18] base class from the textattack library. The percentage of exchanged words was set at 0.5 per cent. For Croatian, MLM augmentations were performed using a variety of pre-trained language models, including EMBEDDIA/crosloengual-bert, Andrija/SRoBERTa-F, macedonizer/hr-roberta-base, and classla/bcms-bertic. In terms of perplexity score, EMBEDDIA/crosloengual-bert, xlm-roberta-base, and Andrija/SRoBERTa-F performed the best. Ultimately, EMBEDDIA/crosloengual-bert was selected after examining its enhanced output. Similar procedures were repeated for additional languages.

Language	Method	Model name
Croatian	CLM	macedonizer/hr-gpt2
	MLM	EMBEDDIA/crosloengual-bert
Bulgarian	CLM	rmihaylov/gpt2-medium-bg
	MLM	rmihaylov/bert-base-bg
Slovak	CLM	Milos/slovak-gpt-j-405M
	MLM	gerulata/slovakbert
Slovene	CLM	macedonizer/sl-gpt2
	MLM	EMBEDDIA/sloberta

Table 5.2 Transformer models used in the training as base encoders for CLM and MLM.

5.6.2.3 Causal language model

Using the original and WordNet-augmented datasets, we optimised three distinct GPT-2 models for each of the three classes. The model was independently optimised for each dataset label to generate positive, negative, and mixed reviews. For the purpose of training the

[16] https://tfhub.dev/google/universal-sentence-encoder-multilingual/3
[17] https://tfhub.dev/google/LaBSE/2
[18] textattack.augmentation.CLAREAugmenter

language generator, we eliminated all reviews longer than 5 words. We utilised GPT-2 models trained in the respective languages as the initial backbone encoder. We optimised the model for the language generation task using a learning rate of 0.001, 1 epoch, a batch size of 4, and 1000 warm-up steps. We employed a decoding strategy with a penalty for bi-gram repetition and a beam search with five beams for text generation. Using this method, we made three different datasets that got bigger, so we could study the size of the corpus as a dependent feature.

5.6.2.4 Training set size

Table *5.3* displays the final distribution of the original, expanded-combined, and expanded-permuted datasets. For the expanded-combined and expanded-permuted, we varied the training set by sampling 10k, 20k, and 40k instances for each class. In the cases of WN, MLM, and CLM, the augmentation methods affected the final size of the training set, as the process of augmentation is influenced by several factors, including the nature of the original text, the matching of the words, WordNet, and semantic constraints. We obtained 10,000 and 20,000 (in some cases, 25,000 and 40,000) samples to be trained and tested for all languages, except for Bulgarian, where the number of instances remained low.

Language	Version	Train			Dev			Test		
		neg	pos	neu	neg	pos	neu	neg	pos	neu
Croatian	Original	467	1,586	145				236	719	78
	lemma	467	1,586	145				236	719	78
	expanded	1,523	3,979	436				742	1,787	254
Bulgarian	Original	864	3,898	710	96	436	80	107	486	88
	lemma	864	3,898	710	96	436	80	107	486	88
	expanded	1,435	6,321	1,060	154	686	116	185	803	133
Slovak	Original	297	1,337	1,926	46	211	265	80	416	545
	lemma	297	1,337	1,926	46	211	265	80	416	545
	expanded	879	2,493	2,397	136	352	326	279	841	627
Slovene	Original	2,722	749	506	138	37	25	431	112	57
	lemma	2,722	749	506	138	37	25	431	112	57
	expanded	13,676	2,165	2,073	559	170	141	2,183	400	229

Table 5.3 Train-development-test distribution of original and expanded dataset.

5.7 Results and discussions

The results of the experiments are shown in Table *5.4*, Table *5.5*, and Table *5.6*. The F1-score and accuracy values for the original, lemma, and expanded versions are shown in Table *5.4*. The performance of the original version of the dataset is superior to that of two other datasets. The performance of the binary-lemmatised version is 1% worse than that of the original dataset. This performance decline is greater in a three-class setting. This demonstrates that the pre-trained models, in this case, XLM-R, which was trained on unprocessed text, prefer a grammatically correct form over a lemma form for the given text. In contrast, separating reviews into individual sentences and using them for training did not perform better than the other two settings. In conclusion, treating opinionated text as a sum of parts does not work well in classification settings. In all languages besides Croatian, the *****nary-original *nary-lemmatised settings outperformed the simple expansion technique.

The results of using permuted and combined versions of the datasets are presented in Table *5.5*. Using the 20k/class version of the dataset yielded a slight improvement in the F1 score for Croatian over the original training dataset, based on the data presented in the table. There were no significant changes to the Bulgarian language. For Slovak, the expanded-permuted 10k-class version produced a four-point improvement for binary classification, but no improvement was observed for ternary classification. The performance of Slovene decreased when permuted and combined versions of the dataset were utilised. With the exception of Slovak, all other languages score higher on the expanded combined train set.

According to the data in Table *5.6*, training on the three augmented datasets did not improve the final classification scores. Some cells in the table were left blank because the augmentation technique did not generate the required number of training instances. In the final column, we present the scores for those data points per class that were either less than 10,000 or greater than 40,000. We performed random approximation tests (Yeh, 2000) using the *sigf* package with 10,000 iterations to determine the statistical significance of differences between the models. For all the languages, none of the models had a statistically significant improvement ($p < 0.05$) score over the model trained with the original data.

Language	**Version**	**Binary**		**Ternary**	
		F1	ACC	F1	ACC
Croatian	Original	94.11	95.86	75.04	88.18
	lemma	93.61	95.53	60.95	77.77
	expanded	73.99	78.76	73.31	86.93
Bulgarian	Original	90.00	94.43	72.90	83.55
	lemma	88.82	93.76	68.31	81.20
	expanded	84.44	91.09	65.89	80.55
Slovak	Original	94.83	97.17	79.50	81.07
	lemma	94.65	96.97	79.43	81.84
	expanded	88.07	90.98	71.60	72.46
Slovene	Original	80.92	87.84	68.70	79.33
	lemma	79.25	87.29	66.38	77.16
	expanded	68.05	85.63	49.96	67.03

Table 5.4 Results of original, lemmatised, and expanded versions of the dataset.

Lang	Ver	Binary		Ternary		Binary		Ternary		Binary		Ternary	
		F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC
Hr	expanded-combined	95.37	96.84	73.17	87.41	95.84	97.16	72.96	85.96	94.26	96.07	71.84	87.6
	expanded-permuted	95.53	96.84	73.87	87.99	94.79	96.4	68.72	84.99	93.06	95.31	71.63	86.93
Bg	expanded-combined	90.16	94.26	66.18	76.35	89.88	93.92	72.23	81.93	89.41	93.76	72.27	82.96
	expanded-permuted	89.85	94.26	71.7	80.91	89.17	93.76	71.69	81.64	89.08	93.76	70.5	79.29
Sk	expanded-combined	97.76	98.79	76.58	77.52	96.92	98.38	77.55	78.09	96.72	98.18	79.34	80
	expanded-permuted	98.12	98.99	76.4	76.94	97.37	98.58	78.31	79.05	97.8	98.79	77.86	79.05
Sv	expanded-combined	75.89	81.76	59.73	70.16	77.9	84.16	62.89	74.88	77.67	83.6	58.8	67
	expanded-permuted	75.57	81.21	53.66	60.16	74.07	79.92	54.62	59.33	77.84	83.24	61.5	73.5

Table 5.5 Results of expanded-combined and expanded-permuted for all languages.

		10k			
Language	**Version**	**Binary**		**Ternary**	
		F1	ACC	F1	ACC
Croatian	wn	94.18	95.96	71.9	87.12
	mlm	92.3	94.55	67.74	81.31
	clm	92.06	94.44	64.96	81.89
Bulgarian	wn				
	mlm				
	clm	87.07	92.58	61.87	79.73
Slovak	wn	96	97.78	74.86	79.82
	mlm	96.19	97.98	77.24	78.67
	clm	92.31	95.96	70.01	72.14
Slovene	wn	73.47	79.18	59.39	68.83
	mlm	63.02	66.11	62	72.16
	clm	74.29	81.03	55.16	65.33
		20k			
Language	**Version**	**Binary**		**Ternary**	
		F1	ACC	F1	ACC
Croatian	wn	93.09	95.31	68.73	84.8
	mlm	90.26	93.35	70.63	83.93
	clm	90.74	93.89	6235	81.8
Bulgarian	wn				
	mlm				
	clm	84.15	90.55	59.05	77.09
Slovak	wn	95.61	97.58	79.35	82.32
	mlm	94.93	97.17	76.49	76.75
	clm	90.54	94.55	69.8	71.85
Slovene	wn	78.25	84.71	53.33	65
	mlm	73.99	79.37	60.827	72.33
	clm	67.19	72.19	54.46	69.83
		25k			
Language	**Version**	**Binary**		**Ternary**	
		F1	ACC	F1	ACC
Croatian	wn				
	mlm	90.76	93.68	69.36	83.15
	clm				
Bulgarian	wn				
	mlm				
	clm	82.76	88.87	58.43	80.02
Slovak	wn				
	mlm	96.27	97.98	73.44	74.25

	clm				
Slovene	wn				
	mlm	76.152	82.13	56.11	64.16
	clm	68.02	73.66	56.38	65.83
		40k			
Language	**Version**	**Binary**		**Ternary**	
		F1	ACC	F1	ACC
Croatian	wn	94.2	95.96	61.78	84.31
	mlm				
	clm				
Bulgarian	wn				
	mlm				
	clm				
Slovak	wn	95.22	97.37	77.67	80.97
	mlm				
	clm	91.63	95.56	68.79	71.66
Slovene	wn	78.25	84.71	58.53	69.5
	mlm				
	clm				
		all			
Language	**Version**	**Binary**		**Ternary**	
		F1	ACC	F1	ACC
Croatian	wn	93.94	95.86	69.43	86.73
	mlm				
	clm	89.73	93.02	67.11	83.83
Bulgarian	wn	91.56	94.94	70.64	84.43
	mlm	88.73	93.76	70.07	81.49
	clm	84.1	91.23	58.35	76.65
Slovak	wn	97.37	98.58	76.5	78.96
	mlm				
	clm	91.4	95.16	68.66	70.5
Slovene	wn	77.83	86.37	59.87	73.5
	mlm				
	clm	65.89	69.98	47.68	57.83

Table 5.6 Results of using augmented datasets using WordNet, MLM and CLM.

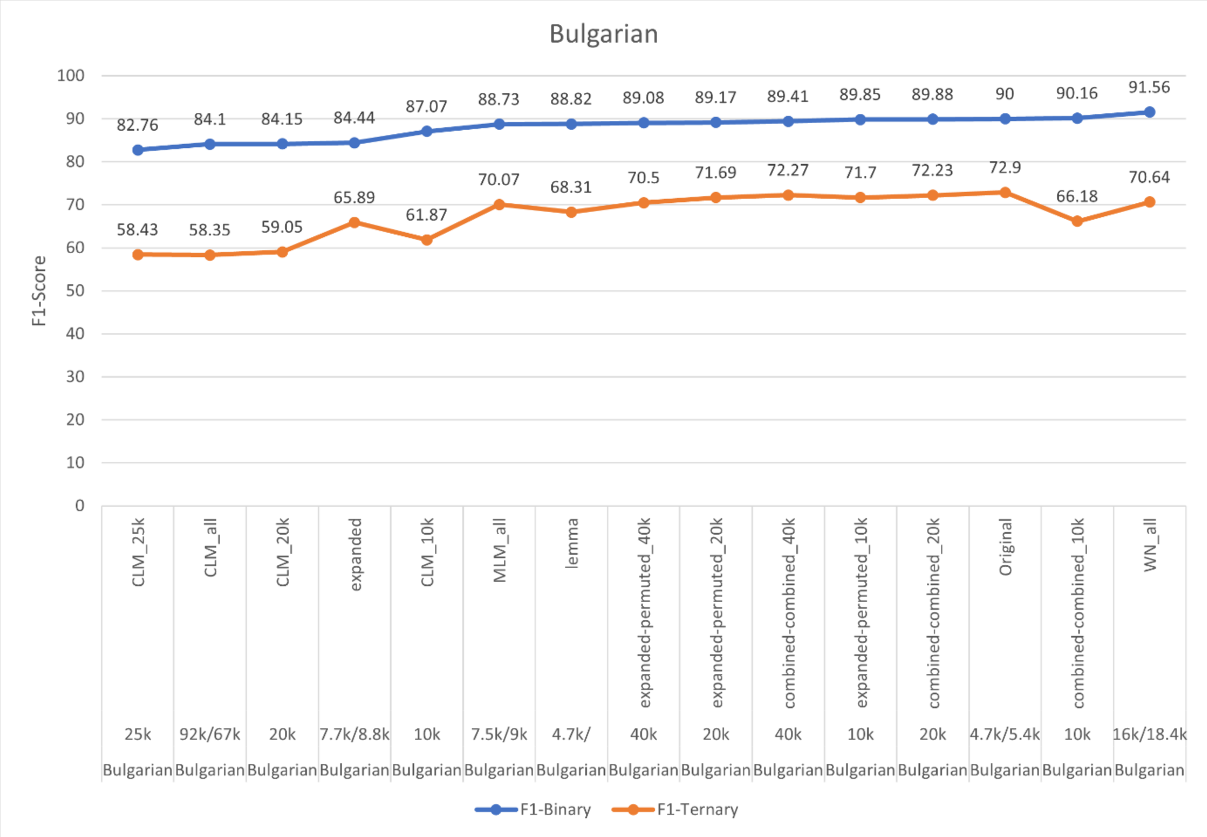

Figure 5.1 Comparison of F1 scores for Bulgarian datasets.

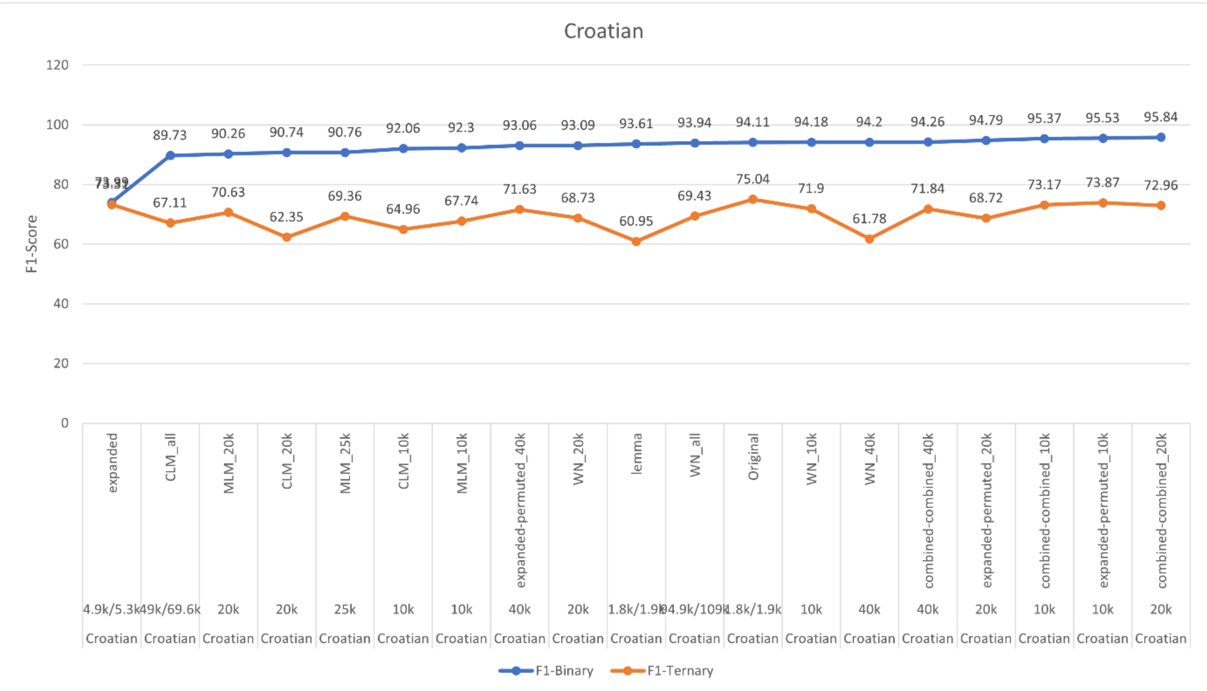

Figure 5.2 Comparison of F1 scores for Croatian datasets.

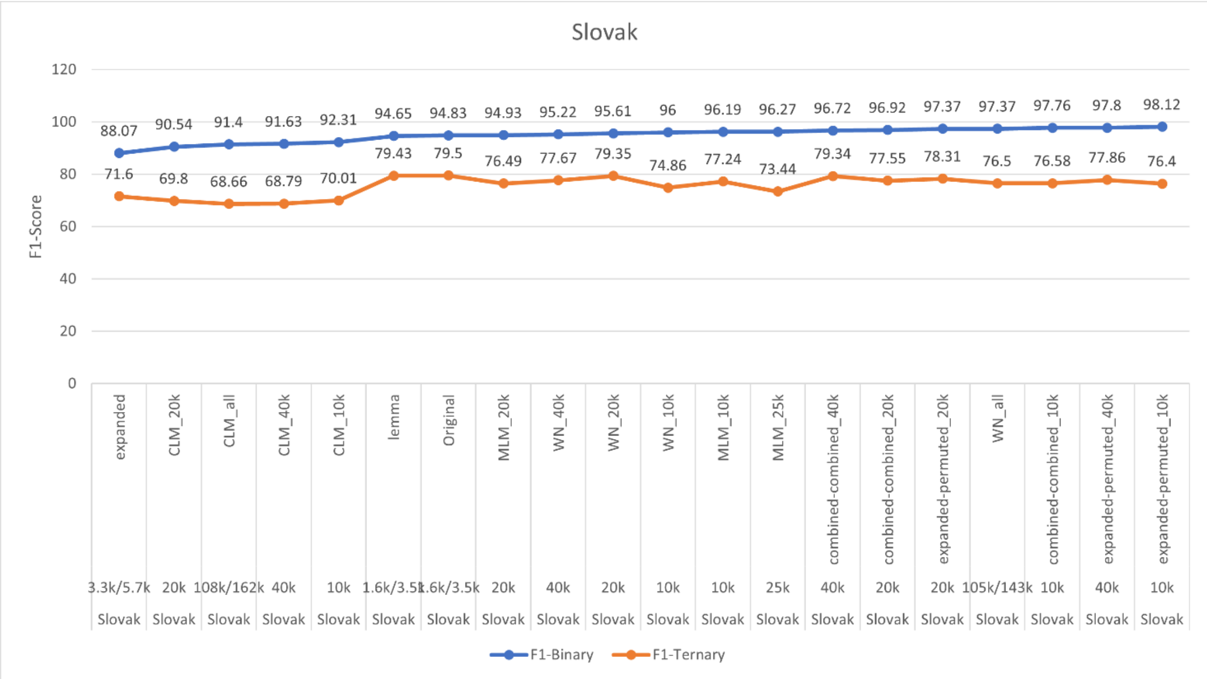

Figure 5.3 Comparison of F1 scores for Slovak datasets.

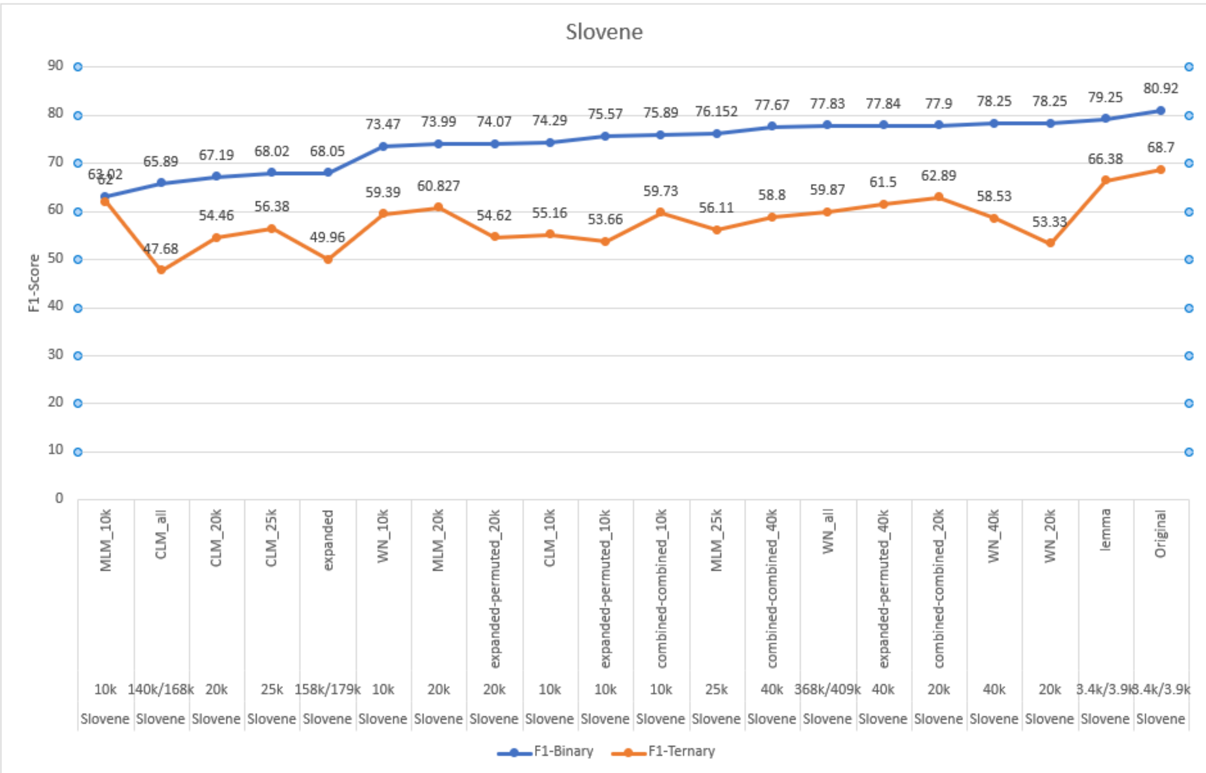

Figure 5.4 Comparison of F1 scores for Slovene datasets.

5.8 Error analysis

For the best scoring models, we randomly sampled incorrectly classified instances from the test set for each language. We manually examined the cases and present a summary of the results.

5.8.1 Text accompanied by additional context

In these instances, the statement starts with a premise or speculation (*I believe it will be good*) and is followed by the user's opinion (*But I did not like it*). Alternatively, the text may contain an opinion followed by speculation. The additional information may or may not justify the users' feelings. In the following example, the user discusses audience members leaving the theatre, followed by his own review. The author marked the review as positive, but the model categorised it as negative.

- (**Original BG**) *Половината салон си тръгна на 30тата минута. Аз следя сериала от както го има и филма ми хареса.*
- (**Transliteration BG**) *Polovinata salon si trgna na 30tata minuta. Az sledya seriala ot kakto go ima i filma mikharesa.*
- (**Translation EN**) *Half the salon left at the 30-minute mark. I've been following the series since it started and I liked the movie.*
- **Original label**: <u>positive</u>; **predicted**: <u>negative</u>.

5.8.2 Reviews with aspect ratings

In this type of text, each aspect is evaluated separately by the user. The current classifier fails to classify these formats, and a specialised process may be required to classify them.

- (**Original BG**) *1 за декорите... Начосът заслужава 5.*
- (**Transliteration BG**) *1 za dekorite... Nachost zasluzhava 5*
- (**Translation EN**) *1 for the decorations... The nachos deserve a 5.*
- **original label**: <u>negative</u>; **predicted**: <u>positive</u>

5.8.3 Mixed aspects

The majority of cases fall into this category. The text comprises a compound or a complex sentence having multiple targets.

- (**Original BG**) *Твърде много ненужно пеене,но всичко останало е супер!:)*
- (**Transliteration BG**) *Tvrde mnogo ne nuzhno peene,no vsichko ostanalo e super!:)*
- (**Translation EN**) *Too much unnecessary singing, but everything else is great!:)*
- **original rating**: negative; **predicted**: positive

5.8.5 Contradictory expressions

The conflicting sub-parts of a sentence are presented as a single unit rather than as a compound sentence, as in the previous error type.

- (**Original BG**) *Красив филм с безкрайно несъстоятелен сценарий*
- (**Transliteration BG**) *Krasiv film s bezkraino nesstoyatelen stsenarii*
- (**Translation EN**) *A beautiful film with an endlessly unworkable script*
- **original rating**: negative; **predicted**: positive

The neutral/mixed-class instances in the Croatian test set have the highest number of misclassifications. The text of binary-classified reviews consists of only positive or negative words. When used with the Transformer encoder, these polar words receive heightened focus, which ultimately determines whether the final classification is positive or negative. In the case of mixed-class sentences, the text is composed of both positive and negative polar words, with one group receiving a disproportionate amount of attention, resulting in an incorrect classification. We discovered that 'ali'-containing sentences were misclassified because the model could not identify compound sentences. As specified by B. Liu, (2020), dealing with mixed-class sentences is difficult because the assumption that the document or sentence has a single target is false.

Further examination of the test-set predictions and ground-truth labels yielded the following findings:

- Some reviews contain sentences that are lengthy. The XLM-R accepts 512 (-2) tokens that have been processed by a tokeniser. Due to the omission of these text tokens, the model performs poorly when the text is exceedingly long. This phenomenon is notable in the Slovene and Croatian datasets.

- There are cases in which the author gave the review a positive rating, but the text contains many unrelated negative statements. This occurs when the author rants about many other stores and writes one positive line about the target entity.
- We also found that the greater the distance between the negation cue and the scope of the negation, the less likely the model is to capture the negation. For example, "*Pizza dola mlaka, i ne ukusna*", vs "*Pizza dola mlaka, i ne ba ukusna*", and "*Pizza dola mlaka, i ne ba previe ukusna*". The first sample is correctly classified, but the second and third samples are not.
- People write negative reviews but rate the restaurant highly because they had a pleasant experience there.
- Code-mixing and English text in Croatian and Slovene. Additionally, we observe that customers rate the overall review positively even if something was missing from the delivery.
- *Brza dostava, ok hrana. Jedino kaj su zaboravili coca colu :(.* (**Translation EN**) *Fast delivery, ok food. Only they forgot about Coca Cola :(.*
- Nisam vidjela pršut na pizzi special, al nema veze, vratina je bila sasvim dovoljna!. (**Translation EN**) *I did not see the prosciutto on the pizza special, but it does not matter, the pork neck was enough!.*
- *Malo gumasto tijesto, inace OK pizza.* (**Translation EN**) *A bit rubbery dough, otherwise ok pizza.*

The MLM model augmentor generated "Treba narucivat chilly" as the correct augmentation for "Ne narucivat chilly" despite paraphrasing constraints. This may be due to the LaBSe model misclassifying texts as paraphrases of one another. Therefore, improved constraints are recommended. For Slovak, we identified cases that contained positive phrases but were labelled neutral by the authors.

- *Bol som vemi spokojný.* (**Translation EN**) *I have been very satisfied*
- *super super super .* (**Translation EN**) *Super Super Super*
- *Bola vemi príjemná a milá. .* (**Translation EN**) *She was very pleasant and nice.*
- *Vemi ústretová a ochotná .* (**Translation EN**) *Very helpful and willing*
- *Bagety,ktoré som kúpila boli perfektné...akujem .* (**Translation EN**) *Baguettes I bought were perfect ... Thank you*

In addition to classification errors, the following text processing errors have been observed: Using the *Classla* package, errors are introduced at three stages (sentence tokenisation, lemmatisation, and POS). For instance, garbled tokens are identified as nouns in the text, and improper sentence boundary detection is also detected. Typically, the user-text lacks diacritics (*narucívati* -> *naruívati*). Therefore, the processing is required to correct the spelling in order to reduce the number of failed WordNet lookups. The Bulgarian dataset consists of movie reviews with emoticons included in the text. This calls for an emoticon-aware tokenizer. *Classla* did not support the processing of nonstandard text types for Bulgarian, so standard mode was used for sentence splitting, lemma, and POS. This is a potential entry point for errors.

5.9 Revisiting research questions

We can answer our research questions after conducting the experiments and analysing the data.

5.9.1 Can the data augmentation techniques improve the performance metric?

According to our findings, using a pre-trained contextualised language encoder reduces the impact of an augmented dataset. As previously reported by Longpre et al. (2020), these transformer-based models are invariant to certain transformations such as synonym substitution. This is attributable to the close proximity of synonyms in the representation space of these encoders. Therefore, using synonyms obtained from WordNet or other sources and encoding them in these spaces does not result in a significant gain. The only way to improve performance is to generate novel linguistic structures that were not encountered during the Transformer model's pre-training.

5.9.2 What is the effect of having augmented data generated from different techniques?

We investigated three distinct data augmentation techniques in addition to three text expansion techniques. Comparing their performance reveals that training with augmented data does not outperform training with the original dataset alone. Although binary class performance has improved by a few points, this improvement is not consistent. In addition, increasing the size of the augmented data has little effect on the performance of the techniques.

5.*9.3 Can WordNet-based augmentation techniques work better with sentiment classification tasks?*

Although WordNet-based augmentation techniques appear to be more effective than MLM and CLM-based techniques, it provides no significant improvement for the downstream task. Training with lemma-based instances decreases system performance by one point for binary classification but drastically for ternary classification. Also, as Xie et al. (2020) pointed out, it is easy to improve the performance of binary sentiment classification by adding more data, but fine-grained classification has the same problem as training on the whole dataset.

5.10 Conclusion

In this chapter, we generated training examples for sentiment classification using three existing data augmentation techniques. In addition, a simple text permutation-combination technique for expanding data without additional resources was experimented upon. We trained a sentiment classifier with varying training sizes using each of the previously outlined techniques. We discovered that using augmented data with a Transformer-based encoder does not result in any significant gains.

6. CONCLUSION

Understanding the polar context of text input from a language is crucial for developing artificial systems that can comprehend human input more effectively. This is not only true for languages with abundant resources, but also for those with limited resources. The scope of application of these automatic sentiment classification systems is not limited to customer reviews but includes digital humanities and psychology, among other.

This thesis presented previous works, hypotheses, experiments, and analyses aimed at solving sentiment classification in low-resource languages using data from a variety of languages and language families, in particular the South-Slavic language family. The objective was to classify sentiments using a cross-lingual setup and resources from well-resourced languages. This study covered the six official South Slavic languages in the EU that have moderate to limited resources for sentiment analysis. In contrast to previous work, ours was designed specifically for a single-language family, focusing primarily on customer reviews. We utilised resources from the same language family, namely Russian, and compared their effects to those of English resources. Using a cross-lingual language encoder and a data augmenter, we conducted experiments to determine the efficacy of contextual language models for sentiment analysis.

Considering the issue of cross-lingual sentiment analysis in low-resource languages, this thesis investigates methods for enhancing classification performance. We address the following research questions using the data and experimental results presented in this thesis.

1. How can we select a good language model for cross-lingual sentiment analysis?

In chapter 3, we conducted experiments on over seventy-five pre-trained language models already in existence. We used three distinct tasks, namely negation, bitext, and paraphrase detection, to evaluate each model by using both existing and newly curated datasets for each task in six languages. Later, the models were fine-tuned in the target languages and re-scored to observe how their internal representations had changed. In the Multi-task setup, all three probing tasks were used for the sentiment analysis task. Using the testing datasets, the final models were re-scored. We conclude, based on the results presented in Sections 3.6, that negation can be a weak signal for selecting a language model for a subsequent sentiment analysis task.

2. What effect do language similarity and the availability of resources have on Multilingual Large Language Models?

In chapter 4, we conducted numerous experiments utilising a combination of datasets from diverse languages, language families, and scripts. We observed that languages that share vocabulary facilitate cross-lingual joint learning of sentiment tasks more effectively. Thus, we were able to conclude that the same family and distant family language datasets can be utilised for low-resource sentiment transfer.

3. Can data augmentation be utilised effectively for sentiment analysis in low-resource languages?

In chapter 5, we evaluated and proposed new techniques for data augmentation as weak supervisors and found empirically that pre-trained models do not benefit from data augmentation when used for sentiment classification. Consequently, based on the experimental findings, we conclude that data augmentation as a means to increase the dataset size for low-resource languages using pre-trained language models cannot be used to improve performance.

6.1 Contribution

Our efforts have led to numerous contributions and discoveries in the field of cross-lingual sentiment analysis. The contributions are outlined below.

- **Pre-trained language model probing**

We posed the question of selecting a single encoder for cross-lingual transfer from the large number of encoders that are publicly accessible. Using three probing tasks, we examined numerous pre-trained contextual language models learned in monolingual and multilingual settings and established a correlation between the probing tasks and sentiment analysis in low-resource languages. We demonstrated empirically that negation has a moderate correlation with the cross-lingual performance of the pre-trained contextual language model.

- **Source language**

We examined the effect of the source language(s) on the proposed multi-task model. Through our experiments, we discovered that languages within the same family with larger datasets are better at cross-lingual sentiment transfer than those with smaller datasets. Additionally, a language from a distant family can overcome typological differences to facilitate the transfer of emotion if it is trained with a high number of training examples.

- **Transfer model**

We presented a multi-task model jointly trained on multiple source and target languages and evaluated in zero-shot settings on multiple target languages. The model outperformed the fine-tuning variant that was used as a baseline and relied on a straightforward classification head. Experiments revealed that cross-lingual performance is highly dependent on the source and target language vocabularies, in our case the shared vocabulary of the text encoder.

- **Resource contributions**

Our work has made available a variety of resources for future research, including a negation dataset for South Slavic languages and a curated list of bitext and paraphrase datasets. Existing noisy neutral class Croatian sentiment data has been re-annotated and utilised in our evaluation. Existing South Slavic sentiment datasets were also presented as benchmark datasets for low-resource sentiment analysis.

6.2 Scope

As stated previously, the scope of this investigation centred on ternary sentiment classification. Although we have also conducted experiments with five-class and binary classification systems. We did not consider mixed expression as a distinct case; it was treated as neutral and handling different sentence types for sentiment is beyond the scope of this study.

Classification of sentiment can be done from three distinct perspectives. The author level captures the writer's mindset, the reader level focuses on the reader's perspective, and the text level relies solely on the text input for the sentiment. We have assumed that text is the primary means of expressing emotion in this work. Even though mainstream research focuses on text-level sentiment, there are few datasets, particularly in Slovene and Croatian, that capture reader-level sentiment. Our cross-lingual methods do not rely on parallel or comparable resources, but they do require a pre-trained model and the datasets from the source language. Performance can be enhanced further if the target language annotated data is available.

6.3 Future directions

There are numerous opportunities to improve this work, some of which are listed below.

- Due to the attention mechanism, the instance's overall sentiment class makes use of a few terms in its text. We believe that isolating these terms should be the first step, followed by the classification of sentences individually into subjective and objective classes. By completing this step, we will eliminate the bias of the objective statements that influence the final score. Future research should also consider classification tasks based on phrases and features.
- We created augmented data using WordNet synonyms. Cross-lingual Word Sense Disambiguation is an intriguing research area that can be used to generate instances with less noise. Rather than relying on lemma form, this could be combined with a re-inflection mechanism to obtain a proper inflected word form.
- The text submitted by the user contains misspellings and variants (such as missing declensions). Possessing an encoder-decoder model to correct these objects from non-standard text could improve the performance of languages with diacritics, particularly South Slavic.
- Performing these experiments on languages from language families such as Indo-Aryan can also present an interesting endeavour (Hindi, Marathi, Konkani).
- Although cross-lingual techniques rely heavily on source languages, the importance of target language datasets cannot be denied. Consequently, using reviews from Google Maps location reviews could also be considered an extension, with user ratings in the form of stars serving as the truth.
- In chapter 6, we utilised NLP tools that are inevitably susceptible to error. Utilizing tools with a lower error rate may be the simplest method of improvement.
- Several research directions in chapter 6 merit further investigation. When utilising pre-trained language models, a more sophisticated data augmentation method is required.

This may be the result of difficult-to-classify instances encountered by the neural network.

- Instead of treating data points as mere data points, we would like to assign each of the augmented instances with informational value. Low-information-value instances provide no benefit and should therefore be filtered out of the training set. In contrast, a training instance with a higher information value should be utilised.
- Recent advancements in meta-learning algorithms (Xia et al., 2021) have also presented opportunities for future research. Few-shot and zero-shot performance enhancements have been demonstrated in multilingual sentiment classification (Sun et al., 2021). The application of these algorithms to settings comparable to ours could be a focus of future research.
- The negation capabilities contained within the other layers of the pre-trained language models, which were not investigated in this study, require additional research.
- Our cross-lingual experiment could be expanded to focus more on the sensitivity of different domain datasets from different source languages.
- With sample importance (T. B. Johnson & Guestrin, 2018; Katharopoulos & Fleuret, 2018), machine translation from source to target language could also be investigated as an augmentation technique.
- During the error analysis, we discovered a large number of examples with noisy labels. We believe that confident learning (Northcutt et al., 2017, 2021) can be used to remove noise from labelled instances and train robust sentiment classification systems.

BIBLIOGRAPHY

1. Adi, Y., Kermany, E., Belinkov, Y., Lavi, O., & Goldberg, Y. (2017). Fine-grained Analysis of Sentence Embeddings Using Auxiliary Prediction Tasks. *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. https://openreview.net/forum?id=BJh6Ztuxl
2. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. J. (2011, June). Sentiment analysis of twitter data. In *Proceedings of the workshop on language in social media*, (pp. 30-38).
3. Agić, Ž., Ljubešić, N., & Tadić, M. Towards Sentiment Analysis of Financial Texts in Croatian. *Bull market*, 143(45), 69.
4. Agić, Ž., & Merkler, D. (2012, December). Rule-based sentiment analysis in narrow domain: Detecting sentiment in daily horoscopes using sentiscope. In *Proceedings of the 2nd Workshop on Sentiment Analysis where AI meets Psychology* (pp. 115-124).
5. Ali, A. E., Stratmann, T. C., Park, S., Schöning, J., Heuten, W., & Boll, S. C. J. (2018). Measuring, Understanding, and Classifying News Media Sympathy on Twitter after Crisis Events. In R. L. Mandryk, M. Hancock, M. Perry, & A. L. Cox (Eds.), *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, Montreal, QC, Canada, April 21-26, 2018* (p. 556). ACM. https://doi.org/10.1145/3173574.3174130
6. Anaby-Tavor, A., Carmeli, B., Goldbraich, E., Kantor, A., Kour, G., Shlomov, S., Tepper, N., & Zwerdling, N. (2020). Do Not Have Enough Data? Deep Learning to the Rescue! *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, 7383–7390. https://ojs.aaai.org/index.php/AAAI/article/view/6233
7. R., B. A., Joshi, A., & Bhattacharyya, P. (2012). Cross-Lingual Sentiment Analysis for Indian Languages using Linked WordNets. In M. Kay & C. Boitet (Eds.), *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Posters, 8-15 December 2012, Mumbai, India* (pp. 73–82). Indian Institute of Technology Bombay. https://aclanthology.org/C12-2008/
8. Artetxe, M., Labaka, G., & Agirre, E. (2017). Learning bilingual word embeddings with (almost) no bilingual data. In R. Barzilay & M.-Y. Kan (Eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers* (pp. 451–462). Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-1042
9. Aue, A., & Gamon, M. (2005, September). Customizing sentiment classifiers to new domains: A case study. In *Proceedings of recent advances in natural language processing (RANLP)* (Vol. 1, No. 3.1, pp. 2-1).
10. Balamurali, A. R., Joshi, A., & Bhattacharyya, P. (2012, December). Cross-lingual sentiment analysis for Indian languages using linked wordnets. *In Proceedings of COLING 2012: Posters* (pp. 73-82).
11. Balahur, A., & Turchi, M. (2012, July). Multilingual sentiment analysis using machine translation?. In *Proceedings of the 3rd workshop in computational approaches to subjectivity and sentiment analysis* (pp. 52-60).
12. Banea, C., Mihalcea, R., & Wiebe, J. (2008). A Bootstrapping Method for Building Subjectivity Lexicons for Languages with Scarce Resources. *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26*

May - 1 June 2008, Marrakech, Morocco. http://www.lrec-conf.org/proceedings/lrec2008/summaries/700.html

13. Banea, C., Mihalcea, R., Wiebe, J., & Hassan, S. (2008, October). Multilingual subjectivity analysis using machine translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (pp. 127-135).
14. Bartusiak, R., Augustyniak, L., Kajdanowicz, T., & Kazienko, P. (2015, September). Sentiment analysis for polish using transfer learning approach. In *2015 Second European Network Intelligence Conference* (pp. 53-59). IEEE.
15. Basiri, M. E., Nemati, S., Abdar, M., Cambria, E., & Acharya, U. R. (2021). ABCDM: An Attention-based Bidirectional CNN-RNN Deep Model for sentiment analysis. *Future Generation Computer Systems*, *115*, 279–294. https://doi.org/10.1016/j.future.2020.08.005
16. Bender. (2019, September 14). The #BenderRule: On Naming the Languages We Study and Why It Matters. Retrieved July 2, 2022, from https://thegradient.pub/the-benderrule-on-naming-the-languages-we-study-and-why-it-matters/
17. Blitzer, J., Dredze, M., & Pereira, F. (2007). Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. In J. A. Carroll, A. van den Bosch, & A. Zaenen (Eds.), *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic*. The Association for Computational Linguistics. https://aclanthology.org/P07-1056/
18. Bollegala, D., Weir, D. J., & Carroll, J. A. (2011). Using Multiple Sources to Construct a Sentiment Sensitive Thesaurus for Cross-Domain Sentiment Classification. In D. Lin, Y. Matsumoto, & R. Mihalcea (Eds.), *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA* (pp. 132–141). The Association for Computer Linguistics. https://aclanthology.org/P11-1014/
19. Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In L. Màrquez, C. Callison-Burch, J. Su, D. Pighin, & Y. Marton (Eds.), *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015* (pp. 632–642). The Association for Computational Linguistics. https://doi.org/10.18653/v1/d15-1075
20. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.-F. Balcan, & H.-T. Lin (Eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html
21. Bučar, J. (2017). Manually sentiment annotated slovenian news corpus SentiNews 1.0. Retrieved from http://hdl.handle.net/11356/1110 (Slovenian language resource repository CLARIN.SI)

22. Bučar, J., Povh, J., & Žnidaršič, M. (2016). Sentiment classification of the Slovenian news texts. In *Proceedings of the 9th International Conference on Computer Recognition Systems CORES 2015* (pp. 777-787). Springer, Cham.
23. Calvo, R. A., & D'Mello, S. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on affective computing*, *1*(1), 18-37.
24. Cambria, E., Speer, R., Havasi, C., & Hussain, A. (2010, November). Senticnet: A publicly available semantic resource for opinion mining. In *2010 AAAI fall symposium series*.
25. Çano, E., & Bojar, O. (2019). Sentiment Analysis of Czech Texts: An Algorithmic Survey. In *Proceedings of the 11th International Conference on Agents and Artificial Intelligence, ICAART 2019, Volume 2, Prague, Czech Republic, February 19-21*, 2019 (pp. 973–979). SciTePress.
26. Cer, D. M., Diab, M. T., Agirre, E., Lopez-Gazpio, I., & Specia, L. (2017). SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In S. Bethard, M. Carpuat, M. Apidianaki, S. M. Mohammad, D. M. Cer, & D. Jurgens (Eds.), *Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval@ACL 2017, Vancouver, Canada, August 3-4, 2017* (pp. 1–14). Association for Computational Linguistics. https://doi.org/10.18653/v1/S17-2001
27. Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., St. John, R., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Strope, B., & Kurzweil, R. (2018). Universal Sentence Encoder for English. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 169–174. https://doi.org/10.18653/v1/D18-2029
28. Chen, Q., Li, W., Lei, Y., Liu, X., & He, Y. (2015). Learning to Adapt Credible Knowledge in Cross-lingual Sentiment Analysis. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, 419–429. https://doi.org/10.3115/v1/p15-1041
29. Chen, T., Xu, R., He, Y., & Wang, X. (2017). Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN. *Expert Systems with Applications*, *72*, 221-230.
30. Chen, X., Hassan, A., Hassan, H., Wang, W., & Cardie, C. (2019, July). Multi-Source Cross-Lingual Model Transfer: Learning What to Share. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3098-3112).
31. Chen, X., Sun, Y., Athiwaratkun, B., Cardie, C., & Weinberger, K. (2018). Adversarial deep averaging networks for cross-lingual sentiment classification. *Transactions of the Association for Computational Linguistics*, 6, (pp. 557-570).
32. Chetviorkin, I., & Loukachevitch, N. (2013). Evaluating sentiment analysis systems in Russian. *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, (pp. 12–17).
33. Cheung, T.-H., & Yeung, D.-Y. (2021). MODALS: Modality-agnostic Automated Data Augmentation in the Latent Space. *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. https://openreview.net/forum?id=XjYgR6gbCEc

34. Chi, E. A., Hewitt, J., & Manning, C. D. (2020). Finding Universal Grammatical Relations in Multilingual BERT. In D. Jurafsky, J. Chai, N. Schluter, & J. R. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020* (pp. 5564–5577). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.493
35. Chidambaram, M., Yang, Y., Cer, D., Yuan, S., Sung, Y.-H., Strope, B., & Kurzweil, R. (2019). Learning Cross-Lingual Sentence Representations via a Multi-task Dual-Encoder Model. In I. Augenstein, S. Gella, S. Ruder, K. Kann, B. Can, J. Welbl, A. Conneau, X. Ren, & M. Rei (Eds.), *Proceedings of the 4th Workshop on Representation Learning for NLP, RepL4NLP@ACL 2019, Florence, Italy, August 2, 2019* (pp. 250–259). Association for Computational Linguistics. https://doi.org/10.18653/v1/w19-4330
36. Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020). ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. https://openreview.net/forum?id=r1xMH1BtvB
37. Clercq, O. D., Lefever, E., Jacobs, G., Carpels, T., & Hoste, V. (2017). Towards an integrated pipeline for aspect-based sentiment analysis in various domains. In A. Balahur, S. M. Mohammad, & E. van der Goot (Eds.), *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA@EMNLP 2017, Copenhagen, Denmark, September 8, 2017* (pp. 136–142). Association for Computational Linguistics. https://doi.org/10.18653/v1/w17-5218
38. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(ARTICLE), (pp. 2493-2537).
39. Conneau, A., Kruszewski, G., Lample, G., Barrault, L., & Baroni, M. (2018). What you can cram into a single \$&!#* vector: Probing sentence embeddings for linguistic properties. In I. Gurevych & Y. Miyao (Eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers* (pp. 2126–2136). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1198
40. Conneau, A., Wu, S., Li, H., Zettlemoyer, L., & Stoyanov, V. (2020, July). Emerging Cross-lingual Structure in Pretrained Language Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 6022-6034).
41. Cotterell, R., & Heigold, G. (2017). Cross-lingual Character-Level Neural Morphological Tagging. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, (pp. 748–759). doi:10.18653/v1/D17-1078
42. Crystal, D. (2009). *A Dictionary of Linguistics and Phonetics* (Vol. 18). Wiley-Blackwell.
43. Cui, H., Mittal, V. O., & Datar, M. (2006). Comparative Experiments on Sentiment Classification for Online Product Reviews. *Proceedings, The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference, July 16-20, 2006, Boston, Massachusetts, USA*, 1265–1270. http://www.aaai.org/Library/AAAI/2006/aaai06-198.php
44. Dadvar, M., Hauff, C., & de Jong, F. (2011, February). Scope of negation detection in sentiment analysis. In *Proceedings of the Dutch-Belgian Information Retrieval Workshop (DIR 2011)* (pp. 16-20). University of Amsterdam.

45. Dalloux, C., Claveau, V., & Grabar, N. (2019). Speculation and Negation detection in French biomedical corpora. In R. Mitkov & G. Angelova (Eds.), *Proceedings of the International Conference on Recent Advances in Natural Language Processing, RANLP 2019, Varna, Bulgaria, September 2-4, 2019* (pp. 223–232). INCOMA Ltd. https://doi.org/10.26615/978-954-452-056-4_026
46. Das, A., & Bandyopadhyay, S. (2011, June). Dr Sentiment knows everything!. In *Proceedings of the ACL-HLT 2011 System Demonstrations* (pp. 50-55).
47. Das, A., & Sarkar, S. (2020). A survey of the model transfer approaches to cross-lingual dependency parsing. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(5), (pp. 1-60).
48. Dashtipour, K., Poria, S., Hussain, A., Cambria, E., Hawalah, A. Y. A., Gelbukh, A. F., & Zhou, Q. (2016). Multilingual Sentiment Analysis: State of the Art and Independent Comparison of Techniques. *Cogn. Comput.*, *8*(4), 757–771. https://doi.org/10.1007/s12559-016-9415-7
49. Del Barrio, E., Cuesta-Albertos, J. A., & Matrán, C. (2018). An optimal transportation approach for assessing almost stochastic order. In *The Mathematics of the Uncertain* (pp. 33-44). Springer, Cham.
50. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers),* (pp. 4171–4186). doi:10.18653/v1/N19-1423
51. Dong, D., Wu, H., He, W., Yu, D., & Wang, H. (2015). Multi-task learning for multiple language translation. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, (pp. 1723–1732).
52. Dong, L., Wei, F., Tan, C., Tang, D., Zhou, M., & Xu, K. (2014). Adaptive Recursive Neural Network for Target-dependent Twitter Sentiment Classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Volume 2: Short Papers), 49–54. https://doi.org/10.3115/v1/P14-2009
53. Dror, R., Baumer, G., Shlomov, S., & Reichart, R. (2018). The Hitchhiker’s Guide to Testing Statistical Significance in Natural Language Processing. In I. Gurevych & Y. Miyao (Eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers* (pp. 1383–1392). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1128
54. Dror, R., Shlomov, S., & Reichart, R. (2019, July). Deep dominance-how to properly compare deep neural models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 2773-2785).
55. Duh, K., Fujino, A., & Nagata, M. (2011). Is Machine Translation Ripe for Cross-Lingual Sentiment Classification? *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA - Short Papers*, 429–433. https://aclanthology.org/P11-2075/
56. Edunov, S., Ott, M., Auli, M., & Grangier, D. (2018). Understanding Back-Translation at Scale. In E. Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language*

Processing, Brussels, Belgium, October 31 - November 4, 2018 (pp. 489–500). Association for Computational Linguistics. https://doi.org/10.18653/v1/d18-1045
57. Erjavec, T., & Fiser, D. (2006). Building Slovene WordNet. In N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk, & D. Tapias (Eds.), *Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC 2006, Genoa, Italy, May 22-28, 2006* (pp. 1678–1683). European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2006/summaries/150.html
58. Esuli, A., & Sebastiani, F. (2006, May). Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*.
59. Ettinger, A., Elgohary, A., Phillips, C., & Resnik, P. (2018). Assessing Composition in Sentence Vector Representations. In E. M. Bender, L. Derczynski, & P. Isabelle (Eds.), *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018* (pp. 1790–1801). Association for Computational Linguistics. https://aclanthology.org/C18-1152/
60. Evkoski, B., Pelicon, A., Mozetič, I., Ljubešić, N., & Kralj Novak, P. (2021). Slovenian Twitter dataset 2018-2020 1.0. http://hdl.handle.net/11356/1423
61. Fei, H., & Li, P. (2020). Cross-lingual unsupervised sentiment classification with multi-view transfer learning. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (pp. 5759–5771).
62. Feng, Y., & Wan, X. (2019, November). Towards a unified end-to-end approach for fully unsupervised cross-lingual sentiment analysis. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)* (pp. 1035-1044).
63. Friedman, M. (1937). The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *Journal of the American Statistical Association*, *32*(200), 675–701. https://doi.org/10.1080/01621459.1937.10503522
64. Galeshchuk, S., Qiu, J., & Jourdan, J. (2019). Sentiment Analysis for Multilingual Corpora. *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, (pp. 120–125). doi:10.18653/v1/W19-3717
65. Gamon, M. (2004). Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In *COLING 2004: Proceedings of the 20th international conference on computational linguistics* (pp. 841-847).
66. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., ... & Lempitsky, V. (2016). Domain-adversarial training of neural networks. *The journal of machine learning research*, *17*(1), 2096-2030.
67. Garg, S., & Ramakrishnan, G. (2020). BAE: BERT-based Adversarial Examples for Text Classification. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020* (pp. 6174–6181). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.498
68. Garimella, K., Morales, G. D. F., Gionis, A., & Mathioudakis, M. (2018). Quantifying controversy on social media. *ACM Transactions on Social Computing*, *1*(1), 1-27.
69. Georgieva-Trifonova, T., Stefanova, M., & Kalchev, S. (2018). Customer Feedback Text Analysis for Online Stores Reviews in Bulgarian. *IAENG International Journal of Computer Science*, 45(4), (pp. 560–568).

70. Glavaš, G., Korenčić, D., & Šnajder, J. (2013). Aspect-oriented opinion mining from user reviews in Croatian. *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, (pp. 18–23).
71. Glavas, G., Litschko, R., Ruder, S., & Vulic, I. (2019). How to (Properly) Evaluate Cross-Lingual Word Embeddings: On Strong Baselines, Comparative Analyses, and Some Misconceptions. In A. Korhonen, D. R. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers* (pp. 710–721). Association for Computational Linguistics. https://doi.org/10.18653/v1/p19-1070
72. Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N project report, Stanford, 1*(12), 2009.
73. Golubev, A., & Loukachevitch, N. (2020). Transfer Learning for Improving Results on Russian Sentiment Datasets, *Computational Linguistics and Intellectual Technologies* (pp. 268-2771).
74. Golubev, A., & Loukachevitch, N. (2020). Improving Results on Russian Sentiment Datasets. In *Artificial Intelligence and Natural Language* (pp. 109–121).
75. Golubović, J., & Gooskens, C. (2015). Mutual intelligibility between West and South Slavic languages. *Russian Linguistics*, *39*(3), 351–373. https://doi.org/10.1007/s11185-015-9150-9
76. Grancharova, M., & Dalianis, H. (2021). Applying and sharing pre-trained BERT-models for named entity recognition and classification in Swedish electronic patient records. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)* (pp. 231-239).
77. Gupta, A., Rallabandi, S. K., & Black, A. W. (2021, April). Task-specific pre-training and cross lingual transfer for sentiment analysis in Dravidian code-switched languages. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages* (pp. 73-79).
78. Habernal, I., & Brychcín, T. (2013, September). Semantic spaces for sentiment analysis. In *International Conference on Text, Speech and Dialogue* (pp. 484-491). Springer, Berlin, Heidelberg.
79. Habernal, I., Ptáček, T., & Steinberger, J. (2013). Sentiment analysis in czech social media using supervised machine learning. *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, (pp. 65–74).
80. Hadsell, R., Chopra, S., & LeCun, Y. (2006). Dimensionality Reduction by Learning an Invariant Mapping. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA*, 1735–1742. https://doi.org/10.1109/CVPR.2006.100
81. Hagège, C. (2011). Linguistically Motivated Negation Processing: an Application for the Detection of Risk Indicators in Unstructured Discharge Summaries. *Polibits*, *43*, 101–106. https://doi.org/10.17562/pb-43-14
82. Haniewicz, K., Rutkowski, W., Adamczyk, M., & Kaczmarek, M. (2013). Towards the lexicon-based sentiment analysis of polish texts: Polarity lexicon. *International Conference on Computational Collective Intelligence*, (pp. 286–295).
83. Hassan, A., & Mahmood, A. (2017, April). Deep learning approach for sentiment analysis of short texts. In *2017 3rd international conference on control, automation and robotics (ICCAR)* (pp. 705-710). IEEE.

84. Hedderich, M. A., Lange, L., Adel, H., Strötgen, J., & Klakow, D. (2021). A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tür, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, & Y. Zhou (Eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021* (pp. 2545–2568). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.201
85. Hewitt, J., & Manning, C. D. (2019). A Structural Probe for Finding Syntax in Word Representations. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)* (pp. 4129–4138). Association for Computational Linguistics. https://doi.org/10.18653/v1/n19-1419
86. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780.
87. Hollander, M., & Wolfe, D. A. (1973). Nonparametric statistical methods John Wiley. *New York. pp*.
88. Horn, L. R., & Wansing, H. (2022). Negation. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Winter 2022). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2022/entries/negation/
89. Hristova, G. (2021). Text Analytics in Bulgarian: An Overview and Future Directions. *Cybernetics and Information Technologies*, *21*(3), (pp. 3–23).
90. Huang, E. H., Socher, R., Manning, C. D., & Ng, A. Y. (2012, July). Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 873-882).
91. Huang, Q., Chen, R., Zheng, X., & Dong, Z. (2017, August). Deep sentiment representation based on CNN and LSTM. In *2017 international conference on green informatics (ICGI)* (pp. 30-33). IEEE.
92. Huang, E. H., Socher, R., Manning, C. D., & Ng, A. Y. (2012). Improving word representations via global context and multiple word prototypes. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (pp. 873–882).
93. itertools combinations. (2022). *{itertools} combinations*. https://docs.python.org/3/library/itertools.html#itertools.combinations
94. itertools permutations. (2022). *{itertools} permutations*. https://docs.python.org/3/library/itertools.html#itertools.permutations
95. Jakopović, H., & Mikelić Preradović, N. (2016). Identifikacija online imidža organizacija temeljem analize sentimenata korisnički generiranog sadržaja na hrvatskim portalima. *Medijska Istraživanja: Znanstveno-Stručni Časopis Za Novinarstvo i Medije*, *22*(2), (pp. 63–82).
96. Jiang, H., He, P., Chen, W., Liu, X., Gao, J., & Zhao, T. (2020). SMART: Robust and Efficient Fine-Tuning for Pre-trained Natural Language Models through Principled Regularized Optimization. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (pp. 2177–2190).
97. Johnson, M., Schuster, M., Le, Q. v, Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., & others. (2017). Google's multilingual

neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, *5*, (pp. 339–351).

98. Johnson, T. B., & Guestrin, C. (2018). Training Deep Models Faster with Robust, Approximate Importance Sampling. In S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada* (pp. 7276–7286). https://proceedings.neurips.cc/paper/2018/hash/967990de5b3eac7b87d49a13c6834978-Abstract.html
99. Kadunc, K., & Robnik-Šikonja, M. (2017). Opinion corpus of Slovene web commentaries KKS 1.001. http://hdl.handle.net/11356/1115
100. Kanayama, H., Nasukawa, T., & Watanabe, H. (2004). Deeper sentiment analysis using machine translation technology. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics* (pp. 494-500).
101. Kandula, H., & Min, B. (2021). Improving Cross-Lingual Sentiment Analysis via Conditional Language Adversarial Nets. *Proceedings of the Third Workshop on Computational Typology and Multilingual NLP*, (pp. 32–37).
102. Kapukaranov, B., & Nakov, P. (2015). Fine-grained sentiment analysis for movie reviews in Bulgarian. *Proceedings of the International Conference Recent Advances in Natural Language Processing*, (pp. 266–274).
103. Katharopoulos, A., & Fleuret, F. (2018). Not All Samples Are Created Equal: Deep Learning with Importance Sampling. In J. G. Dy & A. Krause (Eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018* (Vol. 80, pp. 2530–2539). PMLR. http://proceedings.mlr.press/v80/katharopoulos18a.html
104. Keung, P., Lu, Y., Szarvas, G., & Smith, N. A. (2020). The Multilingual Amazon Reviews Corpus. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.
105. Kim, S.-M., & Hovy, E. (2004). Determining the Sentiment of Opinions. *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, (pp. 1367–1373). https://aclanthology.org/C04-1200
106. Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (pp. 1746–1751). https://doi.org/10.3115/v1/D14-1181
107. Klouda, I. K., & Langr. (2019). *Lukáš; Daniel Vaˇsata, Ing:Title: Product review sentiment analysis in the Czech language*. [Bachelor's thesis, Czech Technical University in Prague]. https://dspace.cvut.cz/bitstream/handle/10467/83127/F8-BP-2019-Langr-Lukas-thesis.pdf
108. Kobayashi, S. (2018). Contextual Augmentation: Data Augmentation by Words with Paradigmatic Relations. In M. A. Walker, H. Ji, & A. Stent (Eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)* (pp. 452–457). Association for Computational Linguistics. https://doi.org/10.18653/v1/n18-2072
109. Kocoń, J., Zaśko-Zielińska, M., & Miłkowski, P. (2019, September). Multi-level analysis and recognition of the text sentiment on the example of consumer opinions. In

Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019) (pp. 559-567).
110. Koeva, S., Genov, A., & Totkov, G. (2004). Towards bulgarian wordnet. *Romanian Journal of Information Science and Technology*, *7*(1-2), 45-60.
111. Krchnavy, R., & Simko, M. (2017). Sentiment analysis of social network posts in Slovak language. In M. Bieliková & M. Simko (Eds.), *12th International Workshop on Semantic and Social Media Adaptation and Personalization, SMAP 2017, Bratislava, Slovakia, July 9-10, 2017* (pp. 20–25). IEEE. https://doi.org/10.1109/SMAP.2017.8022661
112. Kulshreshtha, S., García, J. L. R., & Chang, C.-Y. (2020). Cross-lingual Alignment Methods for Multilingual BERT: A Comparative Study. In T. Cohn, Y. He, & Y. Liu (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020: Vol. EMNLP 2020* (pp. 933–942). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.83
113. Kuznetsova, E. S., Loukachevitch, N. v, & Chetviorkin, I. I. (2013). Testing rules for a sentiment analysis system. *Proceedings of International Conference Dialog*, *2*, (pp. 71–80).
114. Lazarova, G., & Koychev, I. (2015). Semi-supervised multi-view sentiment analysis. In *Computational Collective Intelligence* (pp. 181–190). Springer.
115. Li, D., Zhang, Y., Peng, H., Chen, L., Brockett, C., Sun, M.-T., & Dolan, B. (2021). Contextualized Perturbation for Textual Adversarial Attack. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tür, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, & Y. Zhou (Eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021* (pp. 5053–5069). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.400
116. Li, L., Ma, R., Guo, Q., Xue, X., & Qiu, X. (2020). BERT-ATTACK: Adversarial Attack Against BERT Using BERT. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020* (pp. 6193–6202). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.500
117. Li, Z., Zhang, Y., Wei, Y., Wu, Y., & Yang, Q. (2017). End-to-End Adversarial Memory Network for Cross-domain Sentiment Classification. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, (pp. 2237–2243)
118. Lin, Y.H., Chen, C.Y., Lee, J., Li, Z., Zhang, Y., Xia, M., Rijhwani, S., He, J., Zhang, Z., Ma, X., Anastasopoulos, A., Littell, P., & Neubig, G. (2019). Choosing Transfer Languages for Cross-Lingual Learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3125–3135). Association for Computational Linguistics.
119. Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, *5*(1), 1–184. https://doi.org/10.2200/S00416ED1V01Y201204HLT016
120. Liu, B. (2020). *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge university press.

121. Liu, B., & others. (2010). Sentiment analysis and subjectivity. *Handbook of Natural Language Processing*, *2*(2010), 627–666.
122. Liu, N. F., Gardner, M., Belinkov, Y., Peters, M. E., & Smith, N. A. (2019). Linguistic Knowledge and Transferability of Contextual Representations. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)* (pp. 1073–1094). Association for Computational Linguistics. https://doi.org/10.18653/v1/n19-1112
123. Liu, P., Qiu, X., & Huang, X. (2017). Adversarial Multi-task Learning for Text Classification. In R. Barzilay & M.-Y. Kan (Eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers* (pp. 1–10). Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-1001
124. Liu, R., Shi, Y., Ji, C., & Jia, M. (2019). A Survey of Sentiment Analysis Based on Transfer Learning. *IEEE Access*, *7*, 85401–85412. https://doi.org/10.1109/ACCESS.2019.2925059
125. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, *abs/1907.11692*. http://arxiv.org/abs/1907.11692
126. Ljubešić, N., Fišer, D., Erjavec, T., & Šulc, A. (2021). Offensive language dataset of Croatian, English and Slovenian comments FRENK 1.1. http://hdl.handle.net/11356/1462
127. Lohar, P., Afli, H., & Way, A. (2017). Maintaining Sentiment Polarity of Translated User Generated Content. *The Prague Bulletin of Mathematical Linguistics*, *108*(1), (pp. 73–84).
128. Lohar, P., Afli, H., & Way, A. (2018). Balancing Translation Quality and Sentiment Preservation. *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas*, (pp. 81–88).
129. Lohar, P., Popović, M., & Way, A. (2019). Building English-to-Serbian Machine Translation System for IMDb Movie Reviews. *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, (pp. 105–113).
130. Long, M., Cao, Z., Wang, J., & Jordan, M. I. (2018, December). Conditional adversarial domain adaptation. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, (pp. 1647-1657).
131. Lula, P., & Wójcik, K. (2011). Sentiment analysis of consumer opinions written in Polish. *Economics and Management*, *16*(1), (pp. 1286-1291).
132. MacCartney, B., & Manning, C. D. (2008). Modeling Semantic Containment and Exclusion in Natural Language Inference. In D. Scott & H. Uszkoreit (Eds.), *COLING 2008, 22nd International Conference on Computational Linguistics, Proceedings of the Conference, 18-22 August 2008, Manchester, UK* (pp. 521–528). https://aclanthology.org/C08-1066/
133. Machová, K., Mikula, M., Gao, X., & Mach, M. (2020). Lexicon-based Sentiment Analysis Using the Particle Swarm Optimization. *Electronics*, *9*(8), 1317.
134. Marasovic, A., & Frank, A. (2018). SRL4ORL: Improving Opinion Role Labeling Using Multi-Task Learning with Semantic Role Labeling. In M. A. Walker, H. Ji, & A. Stent (Eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,*

NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers) (pp. 583–594). Association for Computational Linguistics. https://doi.org/10.18653/v1/n18-1054
135. Martinc, M., Montariol, S., Pivovarova, L., & Zosa, E. (2022). Effectiveness of Data Augmentation and Pretraining for Improving Neural Headline Generation in Low-Resource Settings. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022* (pp. 3561–3570). European Language Resources Association. https://aclanthology.org/2022.lrec-1.381
136. McDonald, R. T., Hannan, K., Neylon, T., Wells, M., & Reynar, J. C. (2007). Structured Models for Fine-to-Coarse Sentiment Analysis. In J. A. Carroll, A. van den Bosch, & A. Zaenen (Eds.), *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic*. The Association for Computational Linguistics. https://aclanthology.org/P07-1055/
137. McDonald, R., Petrov, S., & Hall, K. (2011). Multi-Source Transfer of Delexicalized Dependency Parsers. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, (pp. 62–72).
138. Mejova, Y., & Srinivasan, P. (2011). Exploring Feature Definition and Selection for Sentiment Classifiers. In L. A. Adamic, R. Baeza-Yates, & S. Counts (Eds.), *Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011*. The AAAI Press. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2808
139. Mejova, Y., Zhang, A. X., Diakopoulos, N., & Castillo, C. (2014). Controversy and Sentiment in Online News. *CoRR*, *abs/1409.8152*. http://arxiv.org/abs/1409.8152
140. Meng, J., Long, Y., Yu, Y., Zhao, D., & Liu, S. (2019). Cross-domain text sentiment analysis based on CNN_FT method. *Information*, *10*(5), 162. https://doi.org/10.3390/info10050162
141. Merchant, A., Rahimtoroghi, E., Pavlick, E., & Tenney, I. (2020, November). What Happens To BERT Embeddings During Fine-tuning?. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP* (pp. 33-44).
142. Mihalcea, R., Banea, C., & Wiebe, J. (2007). Learning Multilingual Subjective Language via Cross-Lingual Projections. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics* (pp. 976–983). Association for Computational Linguistics.
143. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In Y. Bengio & Y. LeCun (Eds.), *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*. http://arxiv.org/abs/1301.3781
144. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In C. J. C. Burges, L. Bottou, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States* (pp. 3111–3119).

https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html
145. Miller, G. A. (1995). WordNet: A Lexical Database for English. *Commun. ACM*, *38*(11), 39–41. https://doi.org/10.1145/219717.219748
146. Mintz, M., Bills, S., Snow, R., & Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. In K.-Y. Su, J. Su, & J. Wiebe (Eds.), *ACL 2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2-7 August 2009, Singapore* (pp. 1003–1011). The Association for Computer Linguistics. https://aclanthology.org/P09-1113/
147. Mohammad, S. M. (2016). A Practical Guide to Sentiment Annotation: Challenges and Solutions. In A. Balahur, E. V. der Goot, P. Vossen, & A. Montoyo (Eds.), *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA@NAACL-HLT 2016, June 16, 2016, San Diego, California, USA* (pp. 174–179). The Association for Computer Linguistics. https://doi.org/10.18653/v1/w16-0429
148. Mohammad, S. M., Salameh, M., & Kiritchenko, S. (2016). How Translation Alters Sentiment. *J. Artif. Intell. Res.*, *55*, 95–130. https://doi.org/10.1613/jair.4787
149. Moraes, R., Valiati, J. F., & GaviãO Neto, W. P. (2013). Document-Level Sentiment Classification: An Empirical Comparison between SVM and ANN. *Expert Syst. Appl.*, *40*(2), (pp. 621–633). https://doi.org/10.1016/j.eswa.2012.07.059
150. Mozetič, I., Grčar, M., & Smailović, J. (2016). Multilingual Twitter Sentiment Classification: The Role of Human Annotators. *PLoS ONE*, *11*(5), e0155036.
151. Mrsic, L., Kopal, R., & Klepac, G. (2017). Analyzing Slavic Textual Sentiment Using Deep Convolutional Neural Networks. In A. K. Sangaiah, A. Abraham, P. Siarry, & M. Sheng (Eds.), *Intelligent Decision Support Systems for Sustainable Computing - Paradigms and Applications* (Vol. 705, pp. 207–224). Springer. https://doi.org/10.1007/978-3-319-53153-3_11
152. Mullen, T., & Collier, N. (2004). Sentiment analysis using support vector machines with diverse information sources. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, (pp. 412–418).
153. Munezero, M., Montero, C. S., Sutinen, E., & Pajunen, J. (2014). Are They Different? Affect, Feeling, Emotion, Sentiment, and Opinion Detection in Text. *IEEE Trans. Affect. Comput.*, *5*(2), 101–111. https://doi.org/10.1109/TAFFC.2014.2317187
154. Nakov, P. and, Rosenthal, S. and, Kozareva, Z. and, Stoyanov, V. and, Ritter, A. and, & Wilson, T. (2013). SemEval-2013 Task 2: Sentiment Analysis in Twitter, 2nd Joint onf. Lexical and Computational Semantics (* SEM), 7th Int. *Workshop SemEval, Atlanta, Une*, (pp. 14–15).
155. Nandi, R., Maiya, G., Kamath, P., & Shekhar, S. (2021, February). An empirical evaluation of word embedding models for subjectivity analysis tasks. In *2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)* (pp. 1-5). IEEE.
156. Nawaz, R., Thompson, P., & Ananiadou, S. (2013). Negated bio-events: analysis and identification. *BMC Bioinform.*, *14*, 14. https://doi.org/10.1186/1471-2105-14-14
157. Nielsen, F. Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *Workshop on'Making Sense of Microposts: Big Things Come in Small Packages*, (pp. 93–98).

158. Noreen, E. W. (1989). Computer intensive methods for hypothesis testing: An introduction. *Wiley, New York*, *19*, 21.
159. Northcutt, C., Jiang, L., & Chuang, I. (2021). Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research, 70,* 1373-1411.
160. Northcutt, C. G., Wu, T., & Chuang, I. L. (2017). Learning with Confident Examples: Rank Pruning for Robust Classification with Noisy Labels. In G. Elidan, K. Kersting, & A. Ihler (Eds.), *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence, UAI 2017, Sydney, Australia, August 11-15, 2017*. AUAI Press. http://auai.org/uai2017/proceedings/papers/35.pdf
161. Osenova, P., & Simov, K. I. (2012, May). The Political Speech Corpus of Bulgarian. In *LREC* (Vol. 2012, pp. 1744-1747).
162. Paltoglou, G., & Thelwall, M. (2012). Twitter, MySpace, Digg: Unsupervised sentiment analysis in social media. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *3*(4), 1-19.
163. Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, *2*(1–2), 1-135.
164. Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, (pp. 79–86).
165. Paulus, R., Socher, R., & Manning, C. D. (2014). Global Belief Recursive Neural Networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada* (pp. 2888–2896). https://proceedings.neurips.cc/paper/2014/hash/1415db70fe9ddb119e23e9b2808cde38-Abstract.html
166. Pecar, S., Šimko, M., & Bielikova, M. (2019, August). Improving sentiment classification in Slovak language. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing* (pp. 114-119).
167. Pelicon, A., Pranjić, M., Miljković, D., Škrlj, B., & Pollak, S. (2020). Sentiment Annotated Dataset of Croatian News. http://hdl.handle.net/11356/1342
168. Pelicon, A., Pranjić, M., Miljković, D., Škrlj, B., & Pollak, S. (2020). Zero-shot learning for cross-lingual news sentiment classification. *Applied Sciences*, *10*(17), 5993.
169. Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global Vectors for Word Representation. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL* (pp. 1532–1543). ACL. https://doi.org/10.3115/v1/d14-1162
170. Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep Contextualized Word Representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, (pp. 2227–2237).
171. Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P. S. H., Bakhtin, A., Wu, Y., & Miller, A. H. (2019). Language Models as Knowledge Bases? In K. Inui, J. Jiang, V.

Ng, & X. Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019* (pp. 2463–2473). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1250

172. Picard, R. W., & Healey, J. (1997). Affective wearables. *Personal Technologies, 1*(4), 231–240.
173. Pirnat, Ž. (2015). Genesis of the Genitive of Negation in Balto-Slavic and Its Evidence in Contemporary Slovenian. *Slovenski jezik/Slovene Linguistic Studies*, *10*.
174. Polanyi, L., & Zaenen, A. (2006). Contextual valence shifters. In *Computing attitude and affect in text: Theory and applications* (pp. 1–10). Springer.
175. Prettenhofer, P., & Stein, B. (2010). Cross-Language Text Classification Using Structural Correspondence Learning. In J. Hajic, S. Carberry, & S. Clark (Eds.), *ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden* (pp. 1118–1127). The Association for Computer Linguistics. https://aclanthology.org/P10-1114/
176. Pribán, P., & Steinberger, J. (2022). Czech Dataset for Cross-lingual Subjectivity Classification. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022* (pp. 1381–1391). European Language Resources Association. https://aclanthology.org/2022.lrec-1.148
177. Przepiórkowski, A. (2000). Long distance genitive of negation in Polish. *Journal of Slavic linguistics*, 119-158.
178. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.
179. Raffaelli, I., Tadic, M., Bekavac, B., & Agic, Ž. (2008). Building croatian wordnet. In *Proceedings of GWC* (pp. 349-360).
180. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.*, *21*, 140:1-140:67. http://jmlr.org/papers/v21/20-074.html
181. Read, J. (2005). Using Emoticons to Reduce Dependency in Machine Learning Techniques for Sentiment Classification. In K. Knight, H. T. Ng, & K. Oflazer (Eds.), *ACL 2005, 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 25-30 June 2005, University of Michigan, USA* (pp. 43–48). The Association for Computer Linguistics. https://aclanthology.org/P05-2008/
182. Read, J., & Carroll, J. A. (2009). Weakly supervised techniques for domain-independent sentiment classification. In M. Jiang & B. Yu (Eds.), *Proceedings of the 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion, TSA '09, Hong Kong, SAR, China, November 6, 2009* (pp. 45–52). ACM. https://doi.org/10.1145/1651461.1651470
183. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019* (pp. 3980–3990). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

184. Reimers, N., & Gurevych, I. (2020). Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020* (pp. 4512–4525). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.365
185. Ren, S., Deng, Y., He, K., & Che, W. (2019). Generating Natural Language Adversarial Examples through Probability Weighted Word Saliency. In A. Korhonen, D. R. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers* (pp. 1085–1097). Association for Computational Linguistics. https://doi.org/10.18653/v1/p19-1103
186. Riloff, E., & Wiebe, J. (2003). Learning extraction patterns for subjective expressions. *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, (pp. 105–112).
187. Robnik-Šikonja, M., Reba, K., & Mozetič, I. (2021). Cross-lingual transfer of sentiment classifiers. Slovenščina 2.0: Empirical. *Applied and Interdisciplinary Research*, *9*(1), (pp. 1–25). https://doi.org/10.4312/slo2.0.2021.1.1-25
188. Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, *65*(6), 386.
189. Rotim, L., & Šnajder, J. (2017). Comparison of short-text sentiment analysis methods for croatian. *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, (pp. 69–75).
190. Rybiński, K. (2017). Sentiment analysis of Polish politicians. *Konrad Niklewicz*, 162.
191. Saif, H., Fernandez, M., He, Y., & Alani, H. Evaluation Datasets for Twitter Sentiment Analysis. *Emotion and Sentiment in Social and Expressive Media*, 9.
192. Samanta, S., & Mehta, S. (2017). Towards Crafting Text Adversarial Samples. *CoRR, abs/1707.02812*. http://arxiv.org/abs/1707.02812
193. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv Preprint ArXiv:1910.01108*.
194. Scherrer, Y. (2020). TaPaCo: A Corpus of Sentential Paraphrases for 73 Languages. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020* (pp. 6868–6873). European Language Resources Association. https://aclanthology.org/2020.lrec-1.848/
195. Sennrich, R., Haddow, B., & Birch, A. (2016). Improving Neural Machine Translation Models with Monolingual Data. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. https://doi.org/10.18653/v1/p16-1009
196. Shamma, D. A., Kennedy, L., & Churchill, E. F. (2009). Tweet the debates: understanding community annotation of uncollected sources. *Proceedings of the First SIGMM Workshop on Social Media,* (pp. 3–10).
197. Sido, J., Pražák, O., Přibáň, P., Pašek, J., Seják, M., & Konopík, M. (2021, September). Czert–Czech BERT-like Model for Language Representation. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)* (pp. 1326-1338).

198. Simard, P. Y., LeCun, Y., Denker, J. S., & Victorri, B. (2012). Transformation Invariance in Pattern Recognition - Tangent Distance and Tangent Propagation. In G. Montavon, G. B. Orr, & K.-R. Müller (Eds.), *Neural Networks: Tricks of the Trade - Second Edition* (Vol. 7700, pp. 235–269). Springer. https://doi.org/10.1007/978-3-642-35289-8_17
199. Smailović, J., Kranjc, J., Grčar, M., Žnidaršič, M., & Mozetič, I. (2015). Monitoring the Twitter sentiment during the Bulgarian elections. *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, (pp. 1–10).
200. Smetanin, S., & Komarov, M. (2021). Deep transfer learning baselines for sentiment analysis in Russian. *Information Processing & Management*, *58*(3), 102484. https://doi.org/https://doi.org/10.1016/j.ipm.2020.102484
201. Socher, R., Huval, B., Manning, C. D., & Ng, A. Y. (2012). Semantic compositionality through recursive matrix-vector spaces. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, (pp. 1201–1211).
202. Socher, R., Lin, C. C., Manning, C., & Ng, A. Y. (2011). Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the 28th international conference on machine learning (ICML-11)* (pp. 129-136).
203. Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, (pp. 1631–1642).
204. Søgaard, A., & Goldberg, Y. (2016). Deep multi-task learning with low level tasks supervised at lower layers. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers*. https://doi.org/10.18653/v1/p16-2038
205. Steinberger, J., Ebrahim, M., Ehrmann, M., Hurriyetoglu, A., Kabadjov, M., Lenkova, P., Steinberger, R., Tanev, H., Vázquez, S., & Zavarella, V. (2012). Creating sentiment dictionaries via triangulation. *Decision Support Systems*, *53*(4), (pp. 689–694).
206. Stone, P. J., & Hunt, E. B. (1963). A computer approach to content analysis: studies using the general inquirer system. *Proceedings of the May 21-23, 1963, Spring Joint Computer Conference*, (pp. 241–256).
207. Straka, M., Náplava, J., Straková, J., & Samuel, D. (2021, September). RobeCzech: Czech RoBERTa, a monolingual contextualized language representation model. In *International Conference on Text, Speech, and Dialogue* (pp. 197-209). Springer, Cham.
208. Strapparava, C., & Özbal, G. (2010, August). The color of emotions in texts. In *Proceedings of the 2nd Workshop on Cognitive Aspects of the Lexicon* (pp. 28-32).
209. Strapparava, C., & Valitutti, A. (2004, May). Wordnet affect: an affective extension of wordnet. In *Lrec* (Vol. 4, No. 1083-1086, p. 40).
210. Su, P., Li, G., Wu, C., & Vijay-Shanker, K. (2019). Using distant supervision to augment manually annotated data for relation extraction. *PloS one*, *14*(7), e0216913.
211. Sun, P., Ouyang, Y., Zhang, W., & Dai, X. (2021). MEDA: Meta-Learning with Data Augmentation for Few-Shot Text Classification. In Z.-H. Zhou (Ed.), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021* (pp. 3929–3935). ijcai.org. https://doi.org/10.24963/ijcai.2021/541

212. Sussex, R., & Cubberley, P. (2006). *The slavic languages*. Cambridge University Press.
213. Tang, D., Qin, B., & Liu, T. (2015). Document Modeling with Gated Recurrent Neural Network for Sentiment Classification. In L. Màrquez, C. Callison-Burch, J. Su, D. Pighin, & Y. Marton (Eds.), *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015* (pp. 1422–1432). The Association for Computational Linguistics. https://doi.org/10.18653/v1/d15-1167
214. Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., Kim, N., Durme, B. V., Bowman, S. R., Das, D., & Pavlick, E. (2019). What do you learn from context? Probing for sentence structure in contextualized word representations. *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. https://openreview.net/forum?id=SJzSgnRcKX
215. Tesfagergish, S. G., Kapočiūtė-Dzikienė, J., & Damaševičius, R. (2022). Zero-Shot Emotion Detection for Semi-Supervised Sentiment Analysis Using Sentence Transformers and Ensemble Learning. *Applied Sciences*, *12*(17), 8662.
216. Thakkar, G., Preradovic, N. M., & Tadic, M. (2021, April). Multi-task Learning for Cross-Lingual Sentiment Analysis. In *CLEOPATRA@ WWW* (pp. 76-84).
217. Thoits, P. A. (1989). The sociology of emotions. *Annual Review of Sociology*, 317–342.
218. Thongtan, T., & Phienthrakul, T. (2019). Sentiment Classification Using Document Embeddings Trained with Cosine Similarity. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, (pp. 407–414).
219. Townsend, C. E., & Janda, L. A. (1996). *Common and comparative Slavic: phonology and inflection: with special attention to Russian, Polish, Czech, Serbo-Croatian, Bulgarian*. Slavica Pub.
220. Turney, P. D. (2002). Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, 417–424. https://doi.org/10.3115/1073083.1073153
221. Ulcar, M., & Robnik-Sikonja, M. (2020). FinEst BERT and CroSloEngual BERT - Less Is More in Multilingual Models. In P. Sojka, I. Kopecek, K. Pala, & A. Horák (Eds.), *Text, Speech, and Dialogue - 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8-11, 2020, Proceedings* (Vol. 12284, pp. 104–111). Springer. https://doi.org/10.1007/978-3-030-58323-1_11
222. Ulmer, D., Hardmeier, C., & Frellsen, J. (2022). deep-significance-Easy and Meaningful Statistical Significance Testing in the Age of Neural Networks. *ArXiv Preprint ArXiv:2204.06815*.
223. van Aken, B., Winter, B., Löser, A., & Gers, F. A. (2019). How Does BERT Answer Questions?: A Layer-Wise Analysis of Transformer Representations. In W. Zhu, D. Tao, X. Cheng, P. Cui, E. A. Rundensteiner, D. Carmel, Q. He, & J. X. Yu (Eds.), *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019* (pp. 1823–1832). ACM. https://doi.org/10.1145/3357384.3358028
224. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, & R. Garnett (Eds.),

Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA (pp. 5998–6008). https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html
225. Veselovská, K. (2012). Sentence-level sentiment analysis in Czech. In D. D. Burdescu, R. Akerkar, & C. Badica (Eds.), *2nd International Conference on Web Intelligence, Mining and Semantics, WIMS '12, Craiova, Romania, June 6-8, 2012* (p. 65:1-65:4). ACM. https://doi.org/10.1145/2254129.2254208
226. Volkova, S., Wilson, T., & Yarowsky, D. (2013). Exploring Sentiment in Social Media: Bootstrapping Subjectivity Clues from Multilingual Twitter Streams. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 2: Short Papers*, 505–510. https://aclanthology.org/P13-2090/
227. Vulić, I., Glavaš, G., Reichart, R., & Korhonen, A. (2019, November). Do We Really Need Fully Unsupervised Cross-Lingual Embeddings?. 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)* (pp. 4407-4418).
228. Vulic, I., & Moens, M. F. (2013). Cross-lingual semantic similarity of words as the similarity of their semantic word responses. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013)* (pp. 106-116). ACL; East Stroudsburg, PA.
229. Vysusilova, P., & Straka, M. (2021). Sentiment Analysis (Czech Model). http://hdl.handle.net/11234/1-4601
230. Wallat, J., Singh, J., & Anand, A. (2020, November). BERTnesia: Investigating the capture and forgetting of knowledge in BERT. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP* (pp. 174-183).
231. Wan, X. (2009). Co-Training for Cross-Lingual Sentiment Classification. *Proceedings of the Joint Conference of the 47th Annual Meeting of the {ACL} and the 4th International Joint Conference on Natural Language Processing of the {AFNLP},* (pp. 235–243).
232. Wan, X. (2013, August). Co-regression for cross-language review rating prediction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 526-531).
233. Wang, H., Henderson, J., & Merlo, P. (2021, June). Multi-Adversarial Learning for Cross-Lingual Word Embeddings. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 463-472).
234. Wang, J., Lan, C., Liu, C., Ouyang, Y., Qin, T., Lu, W., ... & Yu, P. (2022). Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*.
235. Wang, J., Yu, L.-C., Lai, K. R., & Zhang, X. (2016). Dimensional Sentiment Analysis Using a Regional {CNN}-{LSTM} Model. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers),* (pp. 225–230).

236. Wang, X., Jiang, W., & Luo, Z. (2016). Combination of Convolutional and Recurrent Neural Network for Sentiment Analysis of Short Texts. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, (pp. 2428–2437).
237. Wang, X., Liu, Y., Sun, C., Wang, B., & Wang, X. (2015). Predicting Polarities of Tweets by Composing Word Embeddings with Long Short-Term Memory. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, (pp. 1343–1353).
238. Wang, Z., Ng, P., Ma, X., Nallapati, R., & Xiang, B. (2019). Multi-passage BERT: A Globally Normalized BERT Model for Open-domain Question Answering. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, (pp. 5878–5882).
239. Wawer, A., & Sobiczewska, J. (2019). Predicting Sentiment of Polish Language Short Texts. *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, (pp. 1321–1327).
240. Wei, J. W., & Zou, K. (2019). EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019* (pp. 6381–6387). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1670
241. Wiebe, J., & Mihalcea, R. (2006). Word Sense and Subjectivity. In N. Calzolari, C. Cardie, & P. Isabelle (Eds.), *ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July 2006*. The Association for Computer Linguistics. https://doi.org/10.3115/1220175.1220309
242. Williams, A., Nangia, N., & Bowman, S. R. (2018). A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In M. A. Walker, H. Ji, & A. Stent (Eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)* (pp. 1112–1122). Association for Computational Linguistics. https://doi.org/10.18653/v1/n18-1101
243. Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, (pp. 347–354).
244. Wilson, T., Wiebe, J., & Hoffmann, P. (2009). Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational Linguistics, 35*(3), (pp. 399–433).
245. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., & Rush, A. (2020). Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the*

2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (pp. 38–45). Association for Computational Linguistics.
246. Wu, S., & Dredze, M. (2020, July). Are All Languages Created Equal in Multilingual BERT?. In *Proceedings of the 5th Workshop on Representation Learning for NLP* (pp. 120-130).
247. Xia, M., Zheng, G., Mukherjee, S., Shokouhi, M., Neubig, G., & Hassan, A. (2021, June). MetaXL: Meta Representation Transformation for Low-resource Cross-lingual Learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 499-511).
248. Xie, Q., Dai, Z., Hovy, E. H., Luong, T., & Le, Q. (2020). Unsupervised Data Augmentation for Consistency Training. In H. Larochelle, M. Ranzato, R. Hadsell, M.-F. Balcan, & H.-T. Lin (Eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. https://proceedings.neurips.cc/paper/2020/hash/44feb0096faa8326192570788b38c1d1-Abstract.html
249. Xu, J., Xu, R., Ding, Y., Wang, X., & Kit, C. (2010). Cross Lingual Opinion Analysis via Transfer Learning. *Australian Journal of Intelligent Information Processing Systems*, *11*(2), 28.
250. Yang, Y., Malaviya, C., Fernandez, J., Swayamdipta, S., Bras, R. L., Wang, J.-P., Bhagavatula, C., Choi, Y., & Downey, D. (2020). G-DAug: Generative Data Augmentation for Commonsense Reasoning. In T. Cohn, Y. He, & Y. Liu (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020: Vol. EMNLP 2020* (pp. 1008–1025). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.90
251. Yang, Z., Dai, Z., Yang, Y., Carbonell, J. G., Salakhutdinov, R., & Le, Q. V. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada* (pp. 5754–5764). https://proceedings.neurips.cc/paper/2019/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html
252. Yeh, A. (2000). More accurate tests for the statistical significance of result differences. In *COLING 2000 Volume 2: The 18th International Conference on Computational Linguistics*.
253. Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, & Veselin Stoyanov (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR, abs/1907.11692.*
254. Yoo, J. Y., & Qi, Y. (2021). Towards Improving Adversarial Training of NLP Models. In M.-F. Moens, X. Huang, L. Specia, & S. W. Yih (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021* (pp. 945–956). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-emnlp.81
255. Yuan, Z., Wu, S., Wu, F., Liu, J., & Huang, Y. (2018). Domain attention model for multi-domain sentiment classification. *Knowl. Based Syst.*, *155*, 1–10. https://doi.org/10.1016/j.knosys.2018.05.004

256. Zafra, S. M. J., Díaz, N. P. C., Taboada, M., & Martín-Valdivia, M. T. (2021). Negation detection for sentiment analysis: A case study in Spanish. *Nat. Lang. Eng.*, *27*(2), 225–248. https://doi.org/10.1017/S1351324920000376
257. Zaśko-Zielińska, M., Piasecki, M., & Szpakowicz, S. (2015). A large wordnet-based sentiment lexicon for Polish. *Proceedings of the International Conference Recent Advances in Natural Language Processing*, (pp. 721–730).
258. Zhang, X., & LeCun, Y. (2015). Text Understanding from Scratch. *CoRR*, *abs/1502.01710*. http://arxiv.org/abs/1502.01710
259. Zhang, X., Zhao, J. J., & LeCun, Y. (2015). Character-level Convolutional Networks for Text Classification. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada* (pp. 649–657). https://proceedings.neurips.cc/paper/2015/hash/250cf8b51c773f3f8dc8b4be867a9a02-Abstract.html
260. Zhao, H., Lu, Z., & Poupart, P. (2015). Self-Adaptive Hierarchical Sentence Model. In Q. Yang & M. J. Wooldridge (Eds.), *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015* (pp. 4069–4076). AAAI Press. http://ijcai.org/Abstract/15/571
261. Zhenzhou Wu and Sean Saito. 2017. HiNet: Hierarchical Classification with Neural Network. In *the workshop of the International Conference on Learning Representations.*
262. Zhou, X., Wan, X., & Xiao, J. (2016). Attention-based LSTM Network for Cross-Lingual Sentiment Classification. In J. Su, X. Carreras, & K. Duh (Eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016* (pp. 247–256). The Association for Computational Linguistics. https://doi.org/10.18653/v1/d16-1024
263. Zhou, X., Wan, X., & Xiao, J. (2016). Cross-Lingual Sentiment Classification with Bilingual Document Representation Learning. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. https://doi.org/10.18653/v1/p16-1133
264. Žitnik, S. (2019). Slovene corpus for aspect-based sentiment analysis - SentiCoref 1.0. http://hdl.handle.net/11356/1285
265. Zuo, X., Chen, Y., Liu, K., & Zhao, J. (2020, December). KnowDis: Knowledge Enhanced Data Augmentation for Event Causality Detection via Distant Supervision. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 1544-1550).

APPENDIX A

Language Models

Model	Multilinguality
LaBSE	109
T-Systems-onsite/german-roberta-sentence-transformer-v2	50+ languages2 German English
allenai-specter	English
average-word-embeddings-glove.6B.300d	English
average-word-embeddings-glove.840B.300d	English
average-word-embeddings-komninos	English
average-word-embeddings-levy-dependency	English
bert-base-nli-cls-token	English
bert-base-nli-max-tokens	English
bert-base-nli-mean-tokens	English
bert-base-nli-stsb-mean-tokens	English
bert-base-nli-stsb-wkpooling	English
bert-base-nli-wkpooling	English
bert-base-wikipedia-sections-mean-tokens	English
bert-large-nli-cls-token	English
bert-large-nli-max-tokens	English
bert-large-nli-mean-tokens	English
bert-large-nli-stsb-mean-tokens	English
clip-ViT-B-32-multilingual-v1	50+
distilbert-base-nli-max-tokens	English
distilbert-base-nli-mean-tokens	English
distilbert-base-nli-stsb-mean-tokens	English
distilbert-base-nli-stsb-quora-ranking	English
distilbert-base-nli-stsb-wkpooling	English
distilbert-base-nli-wkpooling	English
distilbert-multilingual-nli-stsb-quora-ranking	50
distilroberta-base-msmarco-v1	English
distilroberta-base-msmarco-v2	English
distilroberta-base-paraphrase-v1	English
distiluse-base-multilingual-cased	15
distiluse-base-multilingual-cased-v1	15

facebook-dpr-ctx_encoder-multiset-base	English
facebook-dpr-ctx_encoder-single-nq-base	English
facebook-dpr-question_encoder-multiset-base	English
facebook-dpr-question_encoder-single-nq-base	English
msmarco-MiniLM-L-12-v3	English
msmarco-MiniLM-L-6-v3	English
msmarco-distilbert-base-dot-prod-v3	English
msmarco-distilbert-base-v2	English
msmarco-distilbert-base-v3	English
msmarco-distilbert-multilingual-en-de-v2-tmp-lng-aligned	English German
msmarco-distilbert-multilingual-en-de-v2-tmp-trained-scratch	English German
msmarco-distilroberta-base-v2	English
msmarco-roberta-base-ance-fristp	English
msmarco-roberta-base-v2	English
msmarco-roberta-base-v3	English
nli-bert-base	English
nli-bert-base-cls-pooling	English
nli-bert-base-max-pooling	English
nli-bert-large	English
nli-bert-large-cls-pooling	English
nli-bert-large-max-pooling	English
nli-distilbert-base	English
nli-distilbert-base-max-pooling	English
nli-roberta-base	100
nli-roberta-large	100
nq-distilbert-base-v1	English
paraphrase-distilroberta-base-v1	English
paraphrase-xlm-r-multilingual-v1	50+
quora-distilbert-base	English
quora-distilbert-multilingual	50+
stsb-bert-base	English
stsb-bert-large	English
stsb-distilbert-base	English
stsb-roberta-base	100
stsb-roberta-large	100
stsb-xlm-r-multilingual	50+
xlm-r-100langs-bert-base-nli-mean-tokens	100

xlm-r-100langs-bert-base-nli-stsb-mean-tokens	100
xlm-r-base-en-ko-nli-ststb	100 /English, Korean
xlm-r-bert-base-nli-mean-tokens	100
xlm-r-bert-base-nli-stsb-mean-tokens	100
xlm-r-distilroberta-base-paraphrase-v1	50+
xlm-r-large-en-ko-nli-ststb	100 /English, Korean
bert-base-multilingual-cased	104
xlm-roberta-base	100
CroSloEngual BERT	Croatian, Slovenian, and English

Table A.1 List of models used for probing.

APPENDIX B

Language	
Bulgarian	|Bulgarian + English
	Czech + Bulgarian
Croatian	Croatian + Czech
	Croatian + Polish,
	Croatian + Bulgarian
	Croatian + English
Czech	
English	Czech + English
Polish	Bulgarian (Latin) + Polish
	Russian (Latin) + Polish
Russian	Bulgarian + Russian
	Croatian + Russian
Slovak	Slovak + English
	Croatian + Slovene
	Slovak + Polish + Slovak
	Bulgarian + Slovak
	Bulgarian (Latin) + Slovene + Slovak
	Russian (Latin) + Slovene + Slovak
Slovene	Czech + Slovene
	Slovene + English
	Russian (Latin) + Slovene
	Croatian + Slovene + Slovak + Bulgarian
	Polish + Slovene
	Bulgarian + Slovene
	Bulgarian (Latin) + Slovene + Slovak

	Slovene + Russian
	Russian (Latin) + Slovene + Slovak
Bulgarian (Latin)	Bulgarian (Latin) + English
	Bulgarian (Latin) + Polish
	Bulgarian (Latin) + Croatian
Russian (Latin)	Russian (Latin) + English
	Russian (Latin) + Polish
	Russian (Latin) + Czech

Table B.1 List of statistically significant language combinations for each language.

APPENDIX C
Error Examples

Croatian

- Descriptive reviews which describe a lot events that have happened
  - Osvrt na aferu quot;lana gratis palainkaquot;. Zelim reci da mi se vlasnik restorana zaista i javio, ispricao i ponudio jednu komplet narudzbu na njegov racun sto je vrijedno hvale. Svakome se dogodi to je razumljivo ali moj komentar je bio upucen iskljucivo timu pauza.hr u smislu da ne drze stvari pod kontrolom i da se kockaju sa povjerenjem svo-
    jih korisnika. Ako vidite da generalno ugostitelji na nekim stvarima kiksaju ee(a kod mene je to slucaj sa quot;gratis palacinkomquot;) onda im treba na to ukazivati preventivno, ako treba i svakodnevno radi ocuvanja vjerodostojnosti vaseg portala od kojeg vi zivite jer,
    budimo realni, ljudi ne vole kad im se zezas sa hranom to bi vi gospodo iz Pauza.hr trebali dobro znati. Da ne duljim. Pozdrav.
    - Review of the "fake free pancake" affair. I want to say that the owner of the
    restaurant really contacted me, apologized and offered me a complete order on his account, which is worthy of praise. It happens to everyone, it’s understandable, but my comment was addressed exclusively to the pauza.hr team in the sense that they don’t have things under control and are gambling with the trust of their users. If you see that in general caterers screw up on some things more often (and in my case this is the case with "gratis pancakes") then you should point this out to them preventively, if necessary and daily in order to preserve the credibility of your portal from which you live because, let’s be realistic, people don’t like it when you mess with their food - you guys from Pauza.hr should know that well. Not to make it any longer. Greetings.
    - Original positive predicted negative

Bulgarian

- （**Original**）Не е лошо филмчето! Ne e losho filmcheto!
  - ！Not a bad movie!

- This comment is treated as positive by the user and neutral by the classifier.

BIOGRAPHY OF THE AUTHOR

Gaurish Pandurang Thakkar earned his bachelor's degree in computer engineering from Goa University in India in 2011. In 2015, he received his MA in Computer Science from Goa University in India. From 2011 to 2019, he worked in the IT industry as an application developer and data science engineer. He is currently employed as an Early-stage researcher for the CLEOPATRA project at the Faculty of Humanities and Social Sciences, University of Zagreb. His broad research interests include natural language processing and the creation of dataset resources. He had 12 research papers published in international journals and conference proceedings. He has supervised four graduate theses.

LIST OF PUBLISHED WORKS

N. Mikelic Preradovic G. Thakkar, I. S. (2021). LEARNERSOURCING IN HUMANITIES AND SOCIAL SCIENCES. *14th Annual International Conference of Education, Research and Innovation*, 861–866.

Alves, D., Thakkar, G., & Tadic, M. (2021). Building and Evaluating Universal Named-Entity Recognition English corpus. *CLEOPATRA@ WWW*, 2–16.

Alves, D., Thakkar, G., Amaral, G., Kuculo, T., & Tadić, M. (2021). *Building Multilingual Corpora for a Complex Named Entity Recognition and Classification Hierarchy using Wikipedia and DBpedia*. 11.

Thakkar, G., Preradović, N. M., & Tadić, M. (2021). Negation Detection Using NooJ. *2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO)*, 237–241.

Gottschalk, S., Kacupaj, E., Abdollahi, S., Alves, D., Amaral, G., Koutsiana, E., Kuculo, T., Major, D., Mello, C., Cheema, G. S., & others. (2021). OEKG: The Open Event Knowledge Graph. *CLEOPATRA@ WWW*, 61–75.

Sarajlic, J., Thakkar, G., Alves, D., & Preradovic, N. M. (2021). Quotations, Coreference Resolution, and Sentiment Annotations in Croatian NewsArticles: An Exploratory Study. *Proceedings of the Conference on Digital Curation Technologies (Qurator 2021), Berlin, Germany, February 8th - to - 12th, 2021, 2836*.

Thakkar, G., Preradovic, N. M., & Tadic, M. (2021). Multi-task Learning for Cross-Lingual Sentiment Analysis. In *Proceedings of the 2nd International Workshop on Cross-lingual Event-centric Open Analytics co-located with the 30th The Web Conference {(WWW} 2021), Ljubljana, Slovenia, April 12, 2021 (online event due to {COVID-19} outbreak)* (Vol. 2829, pp. 76–84). CEUR-WS.org. http://ceur-ws.org/Vol-2829/short1.pdf

Thakkar, G., & Pinnis, M. (2020). Pretraining and Fine-Tuning Strategies for Sentiment Analysis of Latvian Tweets. *Human Language Technologies -- The Baltic Perspective - Proceedings of the Ninth International Conference Baltic HLT 2020*, 55–61.

Alves, D., Thakkar, G., Amaral, G., Kuculo, T., & Tadić, M. (2021). *Building Multilingual Corpora for a Complex Named Entity Recognition and Classification Hierarchy using Wikipedia and DBpedia*. 11.

Alves, D., Kuculo, T., Amaral, G., Thakkar, G., & Tadic, M. (2020). UNER: Universal Named-Entity Recognition Framework. *CEUR Workshop Proceedings, CLEOPATRA Workshop 2020 Co-Located with ESWC*.

Alves, D., Thakkar, G., & Tadić, M. (2020, May). Evaluating Language Tools for Fifteen EU-official Under-resourced Languages. *Proceedings of the 12th Language Resources and Evaluation Conference*, 1866–1873. Retrieved from https://aclanthology.org/2020.lrec-1.230

Alves, D., Thakkar, G., & Tadić, M. (2020, May). Natural Language Processing Chains Inside a Cross-lingual Event-Centric Knowledge Pipeline for European Union Under-resourced Languages. *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, 153–158. Retrieved from https://aclanthology.org/2020.sltu-1.21