Annotation and Machine Identification of Metaphors in Croatian Newspaper Articles

Anotiranje i strojno identificiranje metafore u hrvatskim novinskim člancima

Martina Ptiček, Jasminka Dobša
University of Zagreb
Sveučilište u Zagrebu
Faculty of Organization and Informatics
Fakultet organizacije i informatike
Pavlinska ul. 2, 42000, Varaždin
marpticek@student.foi.hr, jasminka.dobsa@foi.hr

Abstract. *A metaphor is an integral part of communication, but it also affects the way we think and understand the world. Identifying metaphors by using machine learning methods is a challenging task, and the first step in machine identification is the creation of annotated datasets. As part of the paper, an annotated dataset of metaphors in Croatian newspaper articles was created, taken from two newspaper portals - Index.hr and Jutarnji list. The Metaphor Identification Procedure Vrije Universiteit (MIPVU) procedure was adapted for these needs. Metaphor identification through machine learning was approached as a task of classifying text sequences, whereby large language models were used, which were trained on texts in the Croatian language as well. The paper presents the challenges and observations of annotating a metaphor. The identification model achieved an F1 score of 70.75%, and a difference in identification success was observed depending on the data source. The paper provides a comparison of classification accuracy depending on the size of the dataset, as well as guidelines for further research.*

Keywords. Metaphor, metaphor annotation, metaphor identification, large language models

1 Introduction

Since Lakoff and Johnson (1980) presented the theory of the conceptual metaphor, it has become the object of study in many sciences – from psychology, philosophy and linguistics to computational

Sažetak. *Metafora je sastavni dio komunikacije no ona utječe i na način kako razmišljamo i poimamo svijet. Identificiranje metafore metodama strojnog učenja izazovan je zadatak, a prvi korak pri strojnom identificiranju je izrada anotiranih skupova podataka. U sklopu rada izrađen je anotirani skup podataka metafore u hrvatskim novinskim člancima, preuzetim s dva novinska portala -Index.hr i Jutarnji list. Za te potrebe prilagođena je MIPVU (akronim od engl. Metaphor Identification Procedure Vrije Universiteit) procedura. Identificiranju metafore strojnim učenjem pristupilo se kao zadatku klasificiranja sekvenci teksta, pri čemu su korišteni veliki jezični modeli, trenirani i na tekstovima na hrvatskom jeziku. U radu se navode izazovi i opažanja pri anotiranju metafore. Modelom identificiranja ostvaren je F1 rezultat od 70,75% te je uočena razlika u uspješnosti identificiranja ovisno o izvoru podataka. U radu je dana usporedba točnosti klasifikacije ovisno o veličini skupa podataka, kao i smjernice za daljnja istraživanja.*

Ključne riječi. Metafora, anotiranje metafore, identificiranje metafore, veliki jezični modeli

1 Uvod

Od kada su Lakoff i Johnson (1980) predstavili teoriju konceptualne metafore, ona je postala objekt proučavanja u mnogim znanostima – od psihologije, filozofije i lingvistike, pa do računalne lingvistike i računalnih znanosti (usp. Steen et al., 2010).

Metafora je sastavni i važan dio ljudske komunikacije. Lakoff i Johnson svojom su teorijom linguistics and computer science (cf. Steen et al., 2010).

The metaphor is an integral and important part of human communication. Lakoff and Johnson demonstrated with their theory that a metaphor does not remain only in language but affects the way we think and understand the world around us. Lakoff and Johnson teach us about metaphorical concepts like ARGUMENT IS WAR (for instance, "Your claims are *indefensible*.", "*He attacked the weak points* of my argument."), where we map the more concrete, physical domain of war onto the more abstract domain of discussion, which is why we understand the discussion as a war in which someone wins and someone loses.

Lakoff and Johnson do not deal with the issue of the linguistic metaphor, but a metaphor, in addition to being conceptual, is certainly also linguistic because it is expressed by language and in language. Despite not discussing the linguistic metaphor themselves, their theory of the conceptual metaphor has also influenced how we identify the linguistic metaphor.

For instance, Steen et al. (2010), based on the MIP (*Metaphor Identification Procedure*) (Pragglejaz Group, 2007) developed the MIPVU (*Metaphor Identification Procedure Vrije Universiteit*) of tagging linguistic metaphors, which approaches metaphor identification precisely from the perspective of the conceptual metaphor (Krennmayr and Steen, 2017). Steen et al. use the MIPVU procedure to denote texts from the four registers of the BNC Baby corpus[1] – conversation, fiction, academic texts and newspaper articles. They conclude that the metaphor appears most often in academic texts (18.5%) and newspaper articles (16.4%), followed by fiction (11.9%) (cf. Krennmayr and Steen, 2017). Shutova (2017) also states that the metaphor is most common in newspaper articles, politics and a journals about language and literature.

It is precisely because of the conclusions of the previous papers that, as part of this paper, the metaphor was annotated in newspaper articles using the MIPVU procedure, adapted to the Croatian language. Also, as part of the paper, a deep learning model was developed for metaphor identification, where the problem of identification was approached as a problem of the classification of text sequences. When creating the model, contextual word embeddings, that is, large language models were used, which were trained on texts in the Croatian language, among other languages. The language models used are XLM-RoBERTa (Conneau et al., 2020), CroSloEngual (Ulčar & Robnik-Šikonja, 2020) and BERTić (Ljubešić & Lauc, 2021), and the results obtained by each of these models are compared.

pokazali kako metafora ne ostaje samo u jeziku već utječe na način kako razmišljamo i poimamo svijet oko sebe. Lakoff i Johnson uče nas o metaforičkim konceptima poput RASPRAVA JE RAT (npr. „Tvrdnje su ti *neobranjive*.“, „*Napao je slabe točke* moje argumentacije.“) gdje konkretniju, fizičku domenu rata preslikavamo na apstraktniju domenu rasprave pa tako i raspravu poimamo kao rat u kojem netko pobjeđuje a netko gubi.

Lakoff i Johnson ne bave se pitanjem lingvističke metafore, no metafora je, osim konceptualna, svakako i lingvistička jer se izražava jezikom i u jeziku. Unatoč tome što sami ne govore o lingvističkoj metafori, njihova je teorija konceptualne metafore utjecala i na to kako identificiramo lingvističku metaforu.

Tako Steen et al. (2010), na temelju MIP (akronim od engl. *Metaphor Identification Procedure*) (Pragglejaz Group, 2007) razvijaju MIPVU (akronim od engl. *Metaphor Identification Procedure Vrije Universiteit*) proceduru označavanja lingvističke metafore, koja identificiranju metafore pristupa upravo iz rakursa konceptualne metafore (Krennmayr i Steen, 2017). MIPVU procedurom Steen et al. označavaju tekstove iz četiri registra BNC Baby korpusa[1] – konverzacija, fikcija, akademski tekstovi i novinski članci. Zaključuju kako se metafora najčešće pojavljuje u akademskim tekstovima (18,5%) i novinskim člancima (16,4%) te nakon toga u fikciji (11,9%) (usp Krennmayr i Steen, 2017). Shutova (2017) također navodi kako je metafora najčešća u novinskim člancima, politici i časopisu o jeziku i književnosti.

Upravo zbog zaključaka prethodnih radova, u sklopu ovog rada metafora je anotirana u novinskim člancima i to korištenjem MIPVU procedure, prilagođene hrvatskom jeziku. Također, u sklopu rada izrađen je model dubokog učenja za identificiranje metafore, gdje se problemu identificiranja pristupilo kao problemu klasifikacije sekvenci teksta. Pri izradi modela korišteni su kontekstualni vektorski prikazi riječi (engl. *contextual word embeddings*), odnosno veliki jezični modeli, koji su, između ostalih jezika, trenirani i na tekstovima na hrvatskom jeziku. Jezični modeli koji se koriste su XLM-RoBERTa (Conneau et al., 2020), CroSloEngual (Ulčar & Robnik-Šikonja, 2020) i BERTić (Ljubešić & Lauc, 2021), te se uspoređuju rezultati dobiveni svakim od navedenih modela.

Prema našem saznanju, ovo je prvi rad koji koristi MIPVU proceduru te izrađuje anotirani skup podataka metafore u hrvatskim novinskim člancima, kao i prvi rad koji koristi velike jezične modele u identificiranju metafore u hrvatskim novinskim člancima.

[1] The corpus is available at http://www.natcorp.ox.ac.uk/corpus/babyinfo.html

[1] Korpus je dostupan na poveznici http://www.natcorp.ox.ac.uk/corpus/babyinfo.html

To our knowledge, this is the first paper that uses the MIPVU procedure and creates an annotated dataset of metaphors in Croatian newspaper articles, as well as the first paper that uses large language models to identify metaphors in Croatian newspaper articles.

The rest of the paper is organized as follows - in the section on related work, an overview of relevant metaphor annotating procedures and papers that approach metaphor identification using large language models is provided. An overview of the collected data and the annotation procedure is provided in the *Data* section, while the *Classification Model Settings* chapter provides an overview of the metaphor identification system using large language models. The results of the research are presented in the *Results* section, and the conclusion provides a summary of the research, as well as suggestions and guidelines for further research.

2 Related work

The first step in metaphor identification using machine learning methods is the creation of annotated datasets. Below is a brief overview of previous procedures in metaphor annotation that are relevant for this paper, as well as metaphor identification methods using neural networks and large language models.

2.1 Annotating Metaphors

For annotating linguistic metaphors, the currently most accepted procedures are MIP (Pragglejaz Group, 2007) and MIPVU (Steen et al., 2010), which is an upgrade of MIP. Both methods tag the metaphor on a single lexical unit, which most often coincides with a single word, but sometimes compounds and phrases also represent a lexical unit.

MIP has three steps in which it determines whether a word is used metaphorically and does not distinguish between types of metaphors. MIPVU, on the other hand, introduces a distinction between "regular" ("*MRW*"[2]), direct ("*MRW, direct*") and implicit ("*MRW, implicit*") metaphors. However, as we find in Krennmayr and Steen (2017), it was shown that direct and implicit metaphors are represented in an extremely small percentage of 0.2% of the entire corpus.

The MIPVU procedure was used to annotate the well-known VUA[3] dataset, which is used in a large number (if not in all) of papers on metaphor identification in the English language, and 186,695 words were annotated in it.

Ostatak rada organiziran je na sljedeći način - u poglavlju povezani radovi dan je pregled relevantnih procedura označavanja metafore te radova koji identificiranju metafore pristupaju korištenjem velikih jezičnih modela. Pregled prikupljenih podataka i procedure anotiranja dan je u poglavlju *Podaci*, dok je u poglavlju *Postavke modela za klasifikaciju* dan pregled sustava identificiranja metafore korištenjem velikih jezičnih modela. Rezultati istraživanja prikazani su u poglavlju *Rezultati* te je u zaključku dan sažetak istraživanja kao i prijedlozi i smjernice za daljnje istraživanje.

2 Povezani radovi

Prvi korak u identificiranju metafore metodama strojnog učenja izrada je anotiranih skupova podataka. U nastavku je dan kratki pregled za ovaj rad relevantnih prethodnih procedura u anotiranja metafore, kao i metoda identificiranja metafore korištenjem neuronskih mreža i velikih jezičnih modela.

2.1 Anotiranje metafore

Za anotiranje lingvističke metafore, trenutno najprihvaćenije procedure su MIP (Pragglejaz Group, 2007) i MIPVU (Steen et al., 2010), koja je nadogradnja na MIP. Obje metode metaforu označavaju na pojedinoj leksičkoj jedinici, što se najčešće poklapa s jednom riječi, no ponekad i složenice i fraze predstavljaju leksičku jedinicu.

MIP ima tri koraka u kojima određuje koristi li se riječ metaforički te ne razlikuje vrste metafore. MIPVU pak uvodi razlikovanje između „obične“ metafore (oznaka „*MRW*“[2]), direktne (oznaka „*MRW, direct*“) i implicitne (oznaka „*MRW, implicit*“) metafore. No, kao što nalazimo kod (Krennmayr i Steen (2017), pokazalo se da su direktna i implicitna metafora zastupljene u izrazito malom postotku od 0.2% cijelog korpusa.

MIPVU procedurom označen je dobro poznati VUA[3] skup podataka koji se koristi u velikom broju radova (ako ne i u svim) identificiranja metafore u engleskom jeziku, te je u njemu anotirano 186.695 riječi.

2.2 Identificiranje metafore velikim jezičnim modelima

Prema našim saznanjima, prvi rad koji koristi kontekstualni vektorski prikaz riječi je Gao et al. (2018), gdje autori/ce koriste ELMo (Peters et al., 2018). Identificiranju metafore u radu se pristupa

[2] MRW is an acronym for *Metaphor Related Word*
[3] The dataset is available at http://www.vismet.org/metcor/documentation/home.html
[2] MRW je akronim od engl. *Metaphor Related Word*
[3] Skup podataka dostupan je na poveznici http://www.vismet.org/metcor/documentation/home.html

2.2 Metaphor identification by using large language models

To the best of our knowledge, the first paper using contextual word embeddings is Gao et al. (2018), in which the authors use ELMo (Peters et al., 2018). Metaphor identification in the paper is approached as tasks of (1) Sequence labeling, where each word is tagged either as a metaphor or as a literal use; (2) classification by which only the verb in the sentence is tagged as a metaphor or literal use. In both cases, the BiLSTM network is used, and a layer of the attention mechanism was added to the classification model. Using ELMo when tagging text sequences results in an F1 of 70.4% versus an F1 of 61.7% without ELMo.

Mao et al. (2019) base their paper on the hypothesis that the use of linguistic theories of the metaphor - MIP and selectional preferences (Katz & Fodor, 1963; Wilks, 1975, 1978) - will improve the success of identifying metaphors. They use GloVe (Pennington et al., 2014) and ELMo and create two architectures, where MIP is implemented in the first and selectional preferences in the second. The MIP-based model approaches metaphor classification based on the difference between the literal meaning of the word and the contextual one. For the representation of the contextual meaning, the model uses BiLSTM, while for the representation of the literal meaning, it uses GloVe and ELMo, and in this architecture the model achieves an F1 score of 74.0% for the entire VUA data set, and 70.8% for VUA verbs.

Su et al. (2020) approach metaphor identification as a reading comprehension problem and use BERT in their paper (Devlin et al., 2019) embedding layer, while they use a transformers neural network as the basis of their model (Vaswani et al., 2017). In the training process, they use the RoBERTA (Liu et al., 2019) word embeddings, with the aim of fine-tuning the transformers layers. They achieved an F1 score of 80.4% for VUA verbs and 76.9% for all words in the VUA dataset.

Gong et al. (2020) also use the RoBERTA model and approach identification as a task of tagging text sequences. The model is fed linguistic properties (POS, distribution of words in topics, concreteness of words, WordNet, VerbNet and properties from the corpus) while a Feed forward network is used for classification. The best F1 score of 77.1% was achieved for VUA verbs.

Chen et al. (2020), using the BERT fine-tuning method, performed three different experiments – correcting spelling errors in the TOEFL data set, using data from another domain in the training phase, and learning from another type of figurative language. In learning from another type of figurative language, they compose sets of phrases, with the idea that phrases are often metaphors, and in this experiment they achieved the best F1 score of 77.5% kao zadacima (1) označavanja sekvenci teksta (engl. *Sequence labeling*), gdje se svaka riječ označava ili kao metafora ili kao doslovno korištenje; (2) klasifikacije kojom se samo glagol u rečenici označava kao metafora ili doslovno korištenje. U oba slučaja koristi se BiLSTM mreža te je u klasifikacijskom modelu dodan sloj mehanizma pažnje (engl. *Attention*). Korištenjem ELMo-a pri označavanja sekvenci teksta ostvaruje se F1 od 70,4% naspram F1 61,7% bez ELMo-a.

Mao et al. (2019) svoj rad temelje na hipotezi kako će korištenje lingvističkih teorija metafore - MIP i selekcijskih preferenci (Katz & Fodor, 1963; Wilks, 1975, 1978) - poboljšati uspješnost identificiranja metafore, te koriste GloVe (Pennington et al., 2014) i ELMo. Izrađuju dvije arhitekture, gdje je u prvoj implementiran MIP a u drugoj selekcijske preference. Model temeljen na MIP-u pristupa klasificiranju metafore na osnovi razlike između doslovnog značenja riječi i kontekstualnog. Za reprezentaciju kontekstualnog značenja model koriste BiLSTM dok za reprezentaciju doslovnog značenja koriste GloVe i ELMo te u toj arhitekturi ostvaruju F1 vrijednost 74,0% za cijeli VUA skup podataka, odnosno 70,8% za VUA glagole.

Su et al. (2020) identificiranju metafore pristupaju kao problemu čitanja s razumijevanjem te u radu koriste BERT (Devlin et al., 2019) u sloju za ugradnju, dok kao osnovu svog modela koriste transformers neuronsku mrežu (Vaswani et al., 2017). U procesu treniranja koriste RoBERTa (Liu et al., 2019) vektorski prikaz riječi, s ciljem finog podešavanja transformers slojeva. Ostvaruju F1 mjeru 80,4% za VUA glagole i 76,9% za sve riječi u VUA skupu podataka.

Gong et al. (2020) također koriste RoBERTa model te identificiranju pristupaju kao zadatku označavanja sekvenci teksta. Modelu predaju lingvistička svojstva (POS, distribucija riječi u temama, konkretnost riječ, WordNet, VerbNet te svojstva iz korpusa) dok se za klasifikaciju koristi aciklička mreža (engl. *Feed forward network*). Najbolji F1 rezultat od 77,1% ostvaren je za VUA glagole.

Chen et al. (2020) u svom radu metodom finog podešavanja BERT-a izvršavaju tri različita eksperimenta – ispravljanje grešaka u pravopisu TOEFL skupa podataka, korištenje podataka iz druge domenu u fazi treniranja te učenje iz druge vrste figurativnog jezika. U učenju iz druge vrste figurativnog jezika sastavljaju skupove fraza, s idejom kako su fraze često metafore te u tom eksperimentu ostvaruju najbolje rezultate od 77,5% za VUA glagole i 73,4% za VUA sve vrste riječi.

Dankers et al.(2020) predlažu arhitekturu koja inkorporira širi diskurs, tako što koriste prozor konteksta od 2*k* + 1 rečenica prije i poslije rečenice koja je u fokusu identificiranja metafore. U radu eksperimentiraju s dvije različite arhitekture – prva u for VUA verbs and 73.4% for VUA all types of words.

Dankers et al. (2020) propose an architecture that incorporates the wider discourse, by using a context window of 2*k* + 1 sentences before and after the sentence that is the focus of identifying the metaphor. They experimented with two different architectures – in the first, GloVe and ELMo are in the embedding layer, and BiLSTM in the encoding layer, and in the second the BERT model is in the embedding and the encoding layer. The best F1 score of 71.5% was achieved on the architecture with BERT and hierarchical attention, on the VUA dataset.

Lin et al. (2021) approach metaphor identification with the Contrastive Learning approach designed in order for the model to learn the difference between the literal and metaphorical use of a word, based on their distance in the vector space representation. It uses the RoBERTA model for the encoder and contextual display of words, and achieves an F1 score of 79% for VUA all types of words and 75.6% for VUA verbs.

3 Data

Since previous papers on metaphor identification have shown that a large percentage of metaphors are represented in newspaper articles (Krennmayr and Steen, 2017; Shutova, 2017), for the purposes of this paper, articles from two popular Croatian news portals, Index.hr and Jutarnji list, were collected. [4]

The portals were searched by the keywords *migranti* (engl. migrants), *premijer* (engl. prime minister), *predsjednik* (engl. president), *vlada* (engl. government), *Europska unija* (engl. European Union), *korona* (engl. corona). Furthermore, the criterion was that the articles were published between 1/1/2010 and 12/31/2021, while the articles themselves were chosen randomly, with 30 articles were downloaded from each newspaper portal.

In the preparation of texts for tagging, the articles were tokenized to the sentence level, using the NLTK library (Bird et al., 2009), and 1,265 sentences from Index.hr and 742 sentences from Jutarnji list were obtained, i.e. a total of 2,007 sentences.

In order to determine whether there is a difference in the classification result depending on the size of the set, assuming that there is, a comparison of the accuracy of the model was made as part of the paper, depending on the size of the dataset, and the results of the comparison are presented in the *Results* chapter.

kojoj je u sloju ugradnje GloVe i ELMo, te u sloju enkodiranja BiLSTM, te druga u kojoj je u sloju ugradnje i enkodiranja BERT model. Najbolji F1 rezultat od 71,5%, ostvaren je na arhitekturi s BERT-om i hijerarhijskom pažnjom, na VUA skupu podataka.

Lin et al. (2021) identificiranju metafore pristupaju kontrastnim učenjem (engl. *Contrastive learning*), kako bi model naučio razliku između doslovnog i metaforičkog korištenja neke riječi, na temelju njihove udaljenosti u prikazu u vektorskom prostoru. Za enkoder i kontekstualni prikaz riječi koristi RoBERTa model, te ostvaruju rezultat od F1 79% za VUA sve vrste riječi te 75,6% za VUA glagole.

3 Podaci

Budući da su prethodni radovi identificiranja metafore pokazali kako je metafora u velikom postotku zastupljena u novinskim člancima (Krennmayr i Steen, 2017; Shutova, 2017) za potrebe ovog rada prikupljeni su članci s dva popularna hrvatska novinska portala, Index.hr i Jutarnji list. [4]

Portali su pretraživani po ključnim riječima migranti, premijer, predsjednik, vlada, Europska unija, korona. Nadalje je kriterij bio da su članci objavljeni između 1.1.2010. i 31.12.2021., dok su sami članci birani nasumičnim odabirom te je sa svakog novinskog portala preuzeto 30 članka.

U pripremi tekstova za označavanje, članci su tokenizirani na razinu rečenice, pri čemu je korištena NLTK biblioteka (Bird et al., 2009), te je dobiveno 1.265 rečenica s Index.hr i 742 rečenice s Jutarnjeg lista, odnosno ukupno 2.007 rečenica.

Kako bi se utvrdilo postoji li razlika u rezultatu klasifikacije ovisno o veličini skupa, uz pretpostavku kako postoji, u sklopu rada napravljena je i usporedba točnosti modela, ovisno o veličini skupa podataka te su rezultati usporedbe prikazani u poglavlju *Rezultati*.

3.1 Procedura označavanja metafore

Kao osnova za anotiranje metafore uzeta je MIPVU procedura, koja je nadograđena i prilagođena za hrvatski jezik. Naime, MIPVU metoda je primarno rađena za engleski jezik, te samim time ne uzima u obzir specifičnosti drugih jezika (usp. Bogetić et al., 2019; Nacey et al., 2019).

Bogetić et al. (2019) primjenjuju MIPVU proceduru na srpski jezik te uočavaju kako su potrebne prilagodbe za anotiranje metafore u srpskom jeziku. Te su prilagodbe, zbog strukture jezika, primjenjive i na hrvatski jezik. Tako autorice

[4] According to Similarweb (accessed on 02/23/2022), Index.hr is the third and Jutarnji list the fifth most visited portal in Croatia.

[4] Prema Similarweb (pristupljeno 23.02.2022.), Index.hr je treći a Jutarnji list peti najposjećeniji portal u Hrvatskoj.

3.1 Metaphor Annotation Procedure

The MIPVU procedure, which was upgraded and adapted for the Croatian language, was used as the basis for annotating the metaphors. Namely, the MIPVU method was primarily developed for the English language, and therefore does not take into account the specificities of other languages (cf. Bogetić et al., 2019; Nacey et al., 2019).

Bogetić et al. (2019) apply the MIPVU procedure to the Serbian language and observe that adjustments are needed for annotating metaphors in the Serbian language. Due to the structure of the language, these adjustments are also applicable to the Croatian language. For instance, the authors state that when determining a lexical unit, (1) reflexive verbs (e.g. "smijem se" in Croatian and Serbian, meaning "I'm laughing"), (2) common expressions such as "u stvari" ("in fact"), "u redu" ("alright"), "bez obzira" ("regardless") and (3) compounds like "srednja škola" ("secondary school") should be considered one lexical unit, while (4) a preposition followed by a negative pronoun is divided into two units (e.g. "ni s kim" ("with no one") is divided into "ni s" and „kim").

In addition to determining the lexical unit, Bogetić et al. particularly draw attention to the metaphor expressed in oblique cases, which is a specificity of the Serbian but also the Croatian language, and for the purposes of this paper the MIPVU method has been expanded with instructions for tagging the metaphor expressed in oblique cases:

- For each word that is inflected with an oblique case and stands alone, without a preposition, determine which case it is and what the meaning is in the context.
- Determine whether the meaning of the case in the context is different from the basic meaning, but can be understood in comparison with it. If yes, mark the word as a metaphor expressed by case (*MRW, inflectional*)

An important step in the MIPVU procedure is determining whether a word that could potentially be a metaphor has a more basic, simpler meaning. When determining this, it is necessary to consult a monolingual dictionary, for which the *Veliki rječnik hrvatskoga standardnog jezika* (Great Dictionary of the Croatian Standard Language) was chosen (Jojić and Nakić, 2015). If, after checking this dictionary, there is uncertainty as to whether there is a more basic meaning, an additional check was made on the *Hrvatski jezični portal* (Croatian Language Portal). [5]

In addition to the MIPVU procedure and its adaptation, the annotators were instructed to try to determine the conventionality of the metaphor, on a scale from 1 to 5, where 1 is a completely new and 5

navode, a što se preuzima za potrebe ovog rada, kako je pri određivanju leksičke jedinice potrebno (1) povratne glagole (npr. „smijem se"), (2) uobičajene izraze poput „u stvari", „u redu", „bez obzira" i (3) složenice poput „srednja škola" smatrati jednom leksičkom jedinicom, dok se (4) prijedlog praćen negativnom zamjenicom dijeli na dvije jedinice (npr. „ni s kim" se dijeli na „ni s" i „kim").

Osim određivanja leksičke jedinice, Bogetić et al. posebno skreću pažnju na metaforu izraženu kosim padežima, što je specifičnost srpskog ali i hrvatskog jezika, te je i za potrebe ovog rada MIPVU metoda proširena uputama za označavanje metafore izražene kosim padežom:

- Za svaku riječ koja je deklinirana kosim padežom te stoji samostalno, bez prijedloga, utvrdite o kojem se padežu radi te koje je značenje u kontekstu.
- Utvrdite da li značenje padeža u kontekstu različito od osnovnog značenja, ali se može razumjeti u usporedbi s njim. Ukoliko da, označite riječ kao metaforu izraženu padežom (*MRW, inflectional*)

Važan korak u MIPVU proceduri je određivanje da li za riječ koja bi potencijalno mogla biti metafora, postoji osnovnije, jednostavnije značenje. Pri određivanju je potrebno konzultirati jednojezični rječnik, za što je odabran Veliki rječnik hrvatskoga standardnog jezika (Jojić i Nakić, 2015). Ukoliko bi nakon provjere u navedenom rječniku postojala nesigurnost ima li riječ osnovnije značenje, dodatno je vršena provjera na Hrvatskom jezičnom portalu[5].

Dodatno na MIPVU proceduru i njenu prilagodbu, anotatorice su upućene da pokušaju odrediti konvencionalnost metafore, na skali od 1 do 5, gdje je 1 potpuno nova a 5 potpuno konvencionalna metafora. Ovaj je korak dodan budući da kod Shutove (2017) nalazimo kako je upravo konvencionalnost bila predmet neslaganja među anotatorima/cama, budući da su neke metafore u potpunosti konvencionalne i uobičajene u jeziku pa ih neki anotatori nisu niti smatrali metaforama.

Budući da je za anotiranje metafore potrebno razumjeti i teorijske okvire poput konceptualne metafore, odabrane su anotatorice koje imaju formalno obrazovanje iz područja humanističkih znanosti.

3.2 Označeni skup podataka

Metaforu su anotirale dvije osobe, korištenjem tzv. *community* verzije alata Label Studio[6].

Cohen-Kappa slaganje među anotatoricama na razini rečenice je 82.79% za tekstove s Index.hr, dok je za tekstove s Jutarnjeg lista 90.08%. Slaganje među anotatoricama na razini riječi je nešto manje

[5] Available at https://hjp.znanje.hr

[5] Dostupno na https://hjp.znanje.hr

[6] https://labelstud.io/

a completely conventional metaphor. This step was added because in Shutova (2017) we find that it was precisely conventionality that was the subject of disagreement among the annotators, since some metaphors are completely conventional and common in the language, so some annotators did not even consider them to be metaphors.

Considering that it is necessary to understand theoretical frameworks such as the conceptual metaphor to annotate a metaphor, annotators who have formal education in the field of humanities were selected.

3.2 Annotated dataset

The metaphor was annotated by two people, using the so-called community version of the Label Studio tool. [6]

Cohen-Kappa agreement among annotators at the sentence level is 82.79% for texts from Index.hr, while for texts from Jutarnji list it is 90.08%. The agreement between the annotators at the word level is slightly less, the Cohen-Kappa for Index.hr is 67.69%, while for Jutarnji list it is 68.08%. Although the agreement here is less, it is still satisfactory. Namely, as shown by Reidsma and Carletta (2008), apart from the fact that an agreement between 67% and 80% is considered tolerable in computational linguistics, modern machine learning methods can accept less agreement if the disagreement acts as random noise in the data.

Table 1. Dataset overview. The metaphor column shows the number of sentences in which both annotators marked a metaphor and the percentage of the metaphor in the total number of sentences.

	No. of Sentences	**Anotator A**	**Anotator B**	**Metaphor**
Index.hr	1.265	515	576	492 38.89%
Jutarnji list	742	417	435	408 54.99%
Total	2,007	932	1011	900 44.84%
Dataset train test split				
Train	1,605			720 44.86%
Test	402			180 44.78%

It is interesting here that there are sentences for which both annotators agree that they contain a metaphor, but there is a disagreement about which word is the metaphor (e.g. "Posljednjih dana već su *stizali signali* da je "slučaj Marić" mnogo ozbiljniji nego što se u javnosti mogao dobiti dojam i da će odrediti daljnje odnose koalicijskih partnera. ", engl. "In recent days, signs are showing (in Croatian,

pa je tako Cohen-Kappa za Index.hr 67.69% dok je za Jutarnji list 68.08%. Unatoč tome što je slaganje ovdje manje, i dalje je zadovoljavajuće. Naime, kao što su pokazali Reidsma i Carletta (2008), osim što se slaganje između 67% i 80% smatra podnošljivim u računarskoj lingvistici, suvremene metode strojnog učenja mogu prihvatiti manje slaganje, ukoliko neslaganje djeluje kao slučajni šum u podacima.

Tablica 1. Pregled skupa podataka. U stupcu metafora prikazan je broj rečenica u kojima su obje anotatorice označile metaforu te postotak metafore u ukupnom broju rečenica.

	Broj rečenica	**Anotatorica A**	**Anotatorica B**	**Metafora**
Index.hr	1.265	515	576	492 38.89%
Jutarnji list	742	417	435	408 54.99%
Ukupno	2.007	932	1.011	900 44.84%
Podjela skupa na podskupove za treniranje i testiranje				
Treniranje	1.605			720 44.86%
Testiranje	402			180 44.78%

Ovdje je zanimljivo kako postoje rečenice za koje se obje anotatorice slažu kako sadrže metaforu, no postoji neslaganje koja je riječ metafora (npr. „Posljednjih dana već su *stizali signali* da je "slučaj Marić" mnogo ozbiljniji nego što se u javnosti mogao dobiti dojam i da će odrediti daljnje odnose koalicijskih partnera.“ – anotatorica B kao metaforu označava samo *stizali*; „Naposljetku se po medijima *potegnulo pitanje* promjene samoga teksta prisege, a i upitno je postalo što činiti sa svim hrvatskim zakonima, pravnim aktima i propisima koji su pisani na isti način: u muškome rodu.“ – anotatorica B označava samo *potegnulo*; „Sasvim je sigurno da će se na *tragediji* ovih ljudi i dalje *skupljati jeftini* politički *bodovi*, a nedostatak pravog političkog programa i plana *zakrivati* populističkom retorikom.“, anotatorica A označava *skupljati, jeftini, bodovi, zakrivati* dok anotatorica B označava *tragediji, skupljati, zakrivati*).

Potrebno je imati na umu kako su označavanja u lingvistici, odnosno označavanja u tekstu kompleksni zadaci te anotatori/ce često odabiru različite oznake za istu stvar (usp. Reidsma & Carletta, 2008), čak i kada se radi o znatno jednostavnijim označavanjima od metafore.

U finalni skup podataka za identificiranje metafore, kao pozitivni primjeri uzete su one rečenice za koje se obje anotatorice slažu da sadrže metaforu. Sve ostale rečenice uzete su kao negativni primjeri, što uključuje rečenice za koje se obje

[6] https://labelstud.io/

literally "*signs are arriving*") that the "Marić case" is much more serious than the public could have known and that it will determine the future relations of the coalition partners." – annotator B marked only "*arriving*" as a metaphor; "Naposljetku se po medijima *potegnulo pitanje* promjene samoga teksta prisege, a i upitno je postalo što činiti sa svim hrvatskim zakonima, pravnim aktima i propisima koji su pisani na isti način: u muškome rodu. ", engl. "Ultimately, the issue of changing the text of the oath itself was discussed in the media (in Croatian, literally "*question was pulled*", and it became questionable what to do with all Croatian laws, legal acts and regulations that are written in the same way: in the masculine gender." - annotator B marked only "*pulled*"; "Sasvim je sigurno da će se na *tragediji* ovih ljudi i dalje *skupljati jeftini* politički *bodovi*, a nedostatak pravog političkog programa i plana *zakrivati* populističkom retorikom. ", engl. "It is quite certain that *cheap* political *points* will continue to be scored (in Croatian, literally "*collected*") on the *tragedy* of these people, and the lack of a real political program and plan will be *covered up* with populist rhetoric.", annotator A annotated *collected*, *cheap*, *points*, *cover up*, while annotator B annotated *tragedy, collected, cover up*).

It is necessary to keep in mind that annotation in linguistics, i.e. annotation in the text, is a complex tasks, and annotators often choose different tags for the same thing (cf. Reidsma & Carletta, 2008), even when it comes to much simpler annotations than annotating a metaphor.

The final dataset for metaphor identification as positive examples included only those sentences where metaphor was identified by both annotators. The rest of the sentences were labeled as negative examples and they include both the sentences where none of the annotators identified metaphor and the sentences where metaphor was identified by only one of the annotators (altogether 143 sentences). The final dataset contains 900 positive examples (492 Index.hr and 408 Jutarnji list) out of the total number of 2.007 sentences.

From the dataset, 20% of the examples were separated into the test set, using a stratification method to ensure an equal proportion of positive and negative examples in both the training and test sets. This means that the training set contains 1605 sentences, of which 720 are positive examples, while the testing set consists of 402 sentences, of which 180 are positive examples.

In the pre-processing of the data, all letters are converted to lowercase and all special characters are deleted.

The prepared dataset is publicly available at GitHub.[7]

anotatorice slažu kako ne sadrže metaforu ali i one gdje jedna anotatorica smatra da postoji metafora, no druga se ne slaže (takvih je rečenica 143). Finalni skup podataka sadrži 900 pozitivnih primjera (492 Index.hr te 408 Jutarnji list), od ukupno 2.007 rečenica.

Iz skupa podataka, 20% primjera je odvojeno u skup za testiranje, pri čemu je korištena metoda stratifikacije kako bi se osigurao jednak omjer pozitivnih i negativnih primjera i u skupu za treniranje i u skupu za testiranje. Tako skup za treniranje sadrži 1605 rečenica, od čega 720 pozitivnih primjera, dok se skup za testiranje sastoji od 402 rečenice, od čega je 180 pozitivnih primjera.

U pred-procesiranju podataka, sva su slova pretvorena u mala te su izbrisani svi posebni znakovi.

Pripremljeni skup podataka javno je dostupan na GitHubu.[7]

3.3 Opažanja pri označavanju metafore

Označavanje metafore kompleksan je i dugotrajan zadatak (za anotiranje je svakoj anotatorici bilo potrebno oko 70 sati rada). MIPVU procedura, na kojoj se temelji procedura primijenjena u ovom radu, na prvo čitanje djeluje jednostavna no potrebno je stalno vraćanje na upute kako bi se utvrdilo o kakvoj se metafori radi.

Nakon nekog vremena provedenog u anotiranju obje anotatorice primjećuju kako praćenjem uputa i anotiranjem, sve počinje djelovati kao metafora. Slična opažanja nalazimo i kod (Krennmayr i Steen, 2017) koji navode kako, ipak, unatoč tome što MIPVU procedura ostavlja dojam kako bi sve mogla biti metafora, u konačnici je metafora zastupljena u samo 13,6% leksičkih jedinica u VUA skupu.

Anotatorice navode i kako je teško biti konzistentan pri određivanju stupnja konvencionalnosti metafore ali i pri samom određivanju da li se neka riječ koristi metaforički, čak i u slučajevima kada se koristi u istom kontekstu. To zahtjeva stalno vraćanje na prethodno označene rečenice, što ponekad rezultira promjenom mišljenja o prethodnoj anotaciji.

U procesu anotiranja primiječeno je kako novinski tekstovi sadrže najviše metafora koje su u potpunosti ili skoro u potpunosti konvencionalne, odnosno označene su s 4 ili 5 na skali konvencionalnosti. Anotatorica A od ukupno 1.193 metafore (u 932 rečenice), njih 96% označava s 4 ili 5, dok Anotatorica B od ukupno 1.234 metafore (u 1.011 rečenica), njih 87% označava s 4 ili 5. Anotatorice primjećuju kako su neke metafore toliko konvencionalne kako je upitno da li je uopće ispravno smatrati ih metaforama (npr. „Njegov je

[7] https://github.com/pticek/Identification-of-Metaphors-Cro-Newspaper.

[7] https://github.com/pticek/Identification-of-Metaphors-Cro-Newspaper

3.3 Observations on metaphor annotation

Annotating metaphors is a complex and time-consuming task (each annotator needed about 70 hours of work to annotate). The MIPVU procedure, on which the procedure applied in this paper is based, seems simple at first, but the annotator must constantly return to the instructions in order to determine what kind of metaphor it is.

After some time spent annotating, both annotators notice that after a while of following the instructions and annotating, everything starts to seem as a metaphor. We find similar observations in (Krennmayr and Steen, 2017), who state that, despite the fact that the MIPVU procedure gives the impression that everything could be a metaphor, ultimately the metaphor is represented in only 13.6% of the lexical items in the VUA set.

The annotators state that it is difficult to be consistent when determining the degree of conventionality of a metaphor, but also when determining whether a word is used metaphorically, even in cases where it is used in the same context. This requires constantly returning to previously annotated sentences, which sometimes results in a change of mind about the previous annotation.

In the annotation process, it was noticed that the newspaper texts contain the most metaphors that are completely or almost completely conventional, that is, they are marked with 4 or 5 on the conventionality scale. Annotator A, out of a total of 1,193 metaphors (in 932 sentences), marked 96% with a 4 or 5, while Annotator B, out of a total of 1,234 metaphors (in 1,011 sentences), marked 87% with a 4 or 5. The annotators note that some metaphors are so conventional that it is questionable whether they should even be considered metaphors (e.g. "Njegov je jedini cilj zaštititi Inu kao kompaniju od nacionalnog interesa - kažu nam *izvori*.", engl. "His only goal is to protect INA as a company of national interest - *sources* tell us.")

4 Classification model settings

Metaphor identification was approached as a problem of text sequence classification, at the sentence level. In the classification, large language models were used, trained on datasets in the Croatian language (BERTić, CroSloEngual and XLM-RoBERTa base) using the Huggingface library (Wolf et al., 2020).

The same library was used for tokenization, by calling the AutoTokenizer classes when using the BERTić and CroSloEngual models, or by calling the XLMRobertaTokenizer class in the model that uses the XLM-RoBERTa language model.

The Huggingface classes AutoModelForSequenceClassification were used for classification with the BERTić and CroSloEngual jedini cilj zaštititi Inu kao kompaniju od nacionalnog interesa - kažu nam *izvori*.“)

4 Postavke modela za klasifikaciju

Identificiranju metafore pristupilo se kao problemu klasifikacije sekvenci teksta, i to na razini rečenice. Pri klasifikaciji su korišteni veliki jezični modeli, trenirani i na skupovima podataka na hrvatskom jeziku, (BERTić, CroSloEngual i XLM-RoBERTa base) i to korištenjem Huggingface biblioteke (Wolf et al., 2020).

Ista je biblioteka korištena za tokenizaciju, pozivanjem klasa AutoTokenizer pri korištenju BERTić i CroSloEngual modela, odnosno pozivanjem klase XLMRobertaTokenizer u modelu koji koristi XLM-RoBERTa jezični model.

Za klasifikaciju su korištene Huggingface klase AutoModelForSequenceClassification pri klasifikaciji BERTić i CroSloEngual modelom, odnosno XLMRobertaForSequenceClassification za klasifikaciju XLM-RoBERTa modelom.

Za svaki je model rađeno fino podešavanje (engl. *fine-tuning*), te su tako napravljena četiri eksperimenta sa svakim modelom i to (1) broj epoha 3, stopa učenja 3e-5; (2) broj epoha 3, stopa učenja 2e-5; (3) broje epoha 5, stopa učenja 3e-5; (4) broj epoha 5, stopa učenja 2-e5. Veličina serije (engl. *batch size*) je uvijek 32 te su rezultati klasifikacije bilježeni za svaku epohu testiranja te je kao konačan rezultat uzet najbolji rezultat za epohu. Modeli su izvršavani na Google Colab okruženju, korištenjem GPU-a.

5 Rezultati

Rezultati klasifikacije na pozitivnim primjerima prikazani su u tablici 2, gdje možemo vidjeti kako je najbolji F1 rezultat od 70,75% ostvaren s BERTić modelom, pri testu s 5 epoha i stopi učenja 3e-5. Preciznost je u ovom slučaju 70,95% dok je odziv 70,56%. U tablici 3 prikazan je cijeli izvještaj klasificiranja za BERTić model, pa tako vidimo kako je točnost 73,88%, dok je ponderirani F1 rezultat 73,87%. Najbolji je rezultat ostvaren u trećoj epohi, nakon čega F1 na pozitivnim primjerima pada na 70,25% u četvrtoj epohi pa na 69% u petoj epohi.

Sljedeći najbolji F1 rezultat je 67,60% i ostvaren je s XLM-RoBERT-a modelom, u testu s 5 epoha i stopom učenja 2e-5.

Sva tri modela pravilno su klasificirala 80 rečenica, odnosno 44,4% od svih pozitivnih primjera, dok su sva tri pogrešno klasificirala 33 iste rečenice, što čini 18.3%.

models, and the XLMRobertaForSequenceClassification for classification with the XLM-RoBERTA model.

Fine-tuning was done for each model, and thus four experiments were made with each model, namely (1) number of epochs 3, learning rate 3e-5; (2) number of epochs 3, learning rate 2e-5; (3) number of epochs 5, learning rate 3e-5; (4) number of epochs 5, learning rate 2-e5. The batch size is always 32 and the classification results are recorded for each testing epoch, while the best result in an epoch is regarded as the best result of the experiment. The models were executed on the Google Colab environment, using GPU.

5 Results

The classification results on positive examples are shown in Table 2, where we can see that the best F1 result of 70.75% was achieved with the BERTić model, with a test with 5 epochs and a learning rate of 3e-5. In this case, the precision is 70.95%, while the recall is 70.56%. Table 3 shows the entire classification report for the BERTić model, and we can see that the accuracy is 73.88%, while the weighted F1 result is 73.87%. The best result was achieved in the third epoch, after which F1 on positive examples drops to 70.25% in the fourth epoch and to 69% in the fifth epoch.

The next best F1 result is 67.60% and was achieved with the XLM-RoBERT model, in a test with 5 epochs and a learning rate of 2e-5.

All three models correctly classified 80 sentences, or 44.4% of all positive examples, while all three incorrectly classified 33 of the same sentences, which is 18.3%.

Table 2. Model comparison – precision, recall and F1 measure for positive examples. The best F1 result of 70.75% was achieved with the BERTić model, in a test with 5 epochs and a learning rate of 3e-5. Bold letters indicate the best F1 results for each of the language models used.

	3 ep, 2e-5	**3 ep., 3e-5**	**5 ep., 2e-5**	**5 ep., 3e-5**
BERTić				
Precision	73.68%	65.41%	68.00%	70.95%
Recall	54.44%	67.22%	66.11%	70.56%
F1	62.62%	66.30%	67.04%	**70.75%**
CroSloEngual				
Precision	51.61%	65.17%	64.29%	70.95%
Recall	88.89%	64.44%	70.00%	58.33%
F1	65.31%	64.80%	**67.02%**	64.02%
XLM-RoBERTa				

Tablica 2. Usporedba modela – preciznost, odziv i F1 mjera za pozitivne primjere. Najbolji F1 rezultat od 70.75% ostvaren je s BERTić modelom, pri testu s 5 epoha i stopom učenja 3e-5. Masnim slovima označeni su najbolji F1 rezultati za svaki od korištenih jezičnih modela.

	3 ep, 2e-5	**3 ep., 3e-5**	**5 ep., 2e-5**	**5 ep., 3e-5**
BERTić				
Preciznost	73,68%	65,41%	68,00%	70,95%
Odziv	54,44%	67,22%	66,11%	70,56%
F1	62,62%	66,30%	67,04%	**70,75%**
CroSloEngual				
Preciznost	51,61%	65,17%	64,29%	70.95%
Odziv	88,89%	64,44%	70,00%	58,33%
F1	65,31%	64,80%	**67,02%**	64,02%
XLM-RoBERTa				
Preciznost	61,50%	55,97%	67,98%	65,26%
Odziv	68,33%	83,33%	67,22%	68,89%
F1	64,74%	66,96%	**67,60%**	67,03%

Tablica 3. Izvještaj klasificiranja za BERTić model, kojim je, u usporedbi s ostalim modelima, ostvaren najbolji rezultat.

	Preciznost	**Odziv**	**F1**
Negativni	76,23%	76,58%	76,40%
Pozitivni	70,95%	70,56%	70,75%
Točnost			73,88%
Makro prosjek	73,59%	73,57%	73,58%
Ponderirani prosjek	73,87%	73,88%	73,87%

S obzirom na ovo opažanje, napravljena su tri dodatna klasificiranja korištenjem BERTić modela (5 epoha, stopa učenja 3e-5), i to (1) zasebna klasifikacija tekstova s Index.hr odnosno Jutarnjeg lista, (2) skup za treniranje su rečenice s Index.hr a za testiranje Jutarnji list i (3) skup za treniranje su rečenice s Jutarnjeg lista a za testiranje s Index.hr.

Kao što možemo vidjeti u tablici 4, rezultati za klasifikaciju rečenica s Jutarnjeg lista su, za pozitivne primjere, značajno bolji nego za Index.hr – za Jutarnji list F1 mjera za je 67,36% dok je za Index.hr 55,62%. S druge pak strane, rezultati za klasifikaciju negativnih primjera znatno su bolji za Index.hr nego za Jutarnji - F1 42,59% za Jutarnji list, naspram 77,74% za Index.hr.

U tablici 5 prikazani su rezultati testova odvajanjem jednog izvora podataka za treniranje a drugog za testiranje. Kao što možemo vidjeti, rezultati klasifikacije su znatno bolji kada se za treniranje koristi Index.hr a za testiranje Jutarnji list

Precision	61.50%	55.97%	67.98%	65.26%
Recall	68.33%	83.33%	67.22%	68.89%
F1	64.74%	66.96%	**67.60%**	67.03%

By analyzing the misclassified sentences, we can see that there are more of these from Index.hr than from Jutarnji list. In the test dataset there are 101 positive examples from Index.hr (56%) and 79 from Jutarnji list (44%), while BERTić incorrectly classified 41 sentences from Index.hr (66%) and 21 from Jutarnji list (34%). As shown in Table 1, the texts from Jutarnji list contain a higher percentage of metaphors in the total number of sentences from that portal, but in general there are more examples from Index.hr.

Table 3. Classification report for the BERTić model, which, in comparison with other models, achieved the best result.

	Precision	**Recall**	**F1-score**
Negative	76,23%	76.58%	76.40%
Positive	70.95%	70.56%	70.75%
Accuracy			73.88%
Macro avg	73,59%	73.57%	73.58%
Weighted avg	73.87%	73.88%	73.87%

Considering this observation, three additional classifications were made using the BERTić model (5 epochs, learning rate 3e-5), namely (1) a separate classification of texts from Index.hr and Jutarnji list, (2) a training set consisting of sentences with Index.hr and for testing Jutarnji list and (3) a training set that contains sentences from Jutarnji list and for testing from Index.hr.

As we can see in Table 4, the results for the classification of sentences from Jutarnji list are, for positive examples, significantly better than for Index.hr - for Jutarnji list the F1 measure is 67.36%, while for Index.hr it is 55.62% . On the other hand, the results for the classification of negative examples are significantly better for Index.hr than for Jutarnji - F1 42.59% for Jutarnji list, compared to 77.74% for Index.hr.

Table 5 shows the test results achieved by separating one data source for training and another for testing. As we can see, the classification results are significantly better when Index.hr is used for training and Jutarnji list is used for testing (F1 for positive examples 69.92%), rather than vice versa (F1 for positive examples 58.71%).

If we compare the sentences that were misclassified in the first test with the data set of 2007 sentences from both sources with those that were misclassified in the tests where the set of one source was used for training or testing, we come to an interesting observation in the case when the sentences (F1 za pozitivne primjere 69,92%), nego li obratno (F1 za pozitivne primjere 58,71%).

Usporedimo li rečenice koje su pogrešno klasificirane u prvom testu sa skupom podataka od 2007 rečenice iz oba izvora, s onima koje su pogrešno klasificirane u testovima gdje je skup jednog izvora korišten za treniranje odnosno za testiranje, dolazimo do zanimljivog opažanja u slučaju kada se rečenice s Index.hr koriste za testiranje – 21 rečenica od 41 koje su u prvom testu pogrešno klasificirane, u ovom su slučaju ispravno klasificirane. S druge strane, kada je Index.hr korišten za treniranje a Jutarnji list za testiranje, od 21 rečenice s Jutarnjeg lista koje su u prvom testu pogrešno klasificirane, u ovom slučaju je njih 18 pogrešno klasificirano.

Ovdje je potrebno svakako imati na umu kako svaki portal tj novinari portala imaju svoj stil pisanja te bi bilo potrebno napraviti dodatnu lingvističku analizu samih primjera kako bi se pokušao pronaći uzorak u njima. Podrobnija bi nam lingvistička analiza tekstova potencijalno mogla pokazati i ograničenja samih velikih jezičnih modela, pri identificiranju metafore, što je van opsega ovog rada te se predlaže za buduće istraživanje.

Budući da su u skupu podataka pozitivni primjeri samo one rečenice gdje se obje anotatorice slažu da postoji metafora u rečenici, analizom rezultata testiranja provjereno je koliko je negativnih primjera BERTić klasificirao kao pozitivne, a da se poklapaju s primjerima koje je samo jedna od anotatorica označila kao metaforu. U skupu za testiranje nalazi se 28 negativnih primjera koji su negativni jer ih je samo jedna od anotatorica označila kao rečenice s metaforom te je od tog broja 11 rečenica klasificirano kao pozitivno. Svakako bi u narednim istraživanjima bilo zanimljivo vidjeti kakvi bi rezultati bili kada bi se kao pozitivni primjeri uključile i one rečenice koje je samo jedna anotatorica označila kao rečenice s metaforom.

Tablica 4. Rezultati zasebne klasifikacije rečenica s Jutarnjeg lista i Index.hr Tablica prikazuje F1, odziv i preciznost za pozitivne, odnosno negativne primjere.

	Preciznost	**Odziv**	**F1**
	Pozitivni primjeri		
Jutarnji list	59,26%	78,05%	67,37%
Index.hr	66,20%	47,95%	55,62%
	Negativni primjeri		
Jutarnji list	56,10%	34,33%	42,59%
Index.hr	71,98%	84,51%	77,74%

Tablica 5. Rezultati klasifikacija podjelom skupova podataka na skup za testiranje i testiranje prema izvoru podatka.

from Index.hr are used for testing - 21 sentences out of 41 that were incorrectly classified in the first test were correctly classified in this case. On the other hand, when Index.hr was used for training and Jutarnji list for testing, out of 21 sentences from Jutarnji list that were misclassified in the first test, 18 of them were misclassified in this case.

Here it is necessary to keep in mind that each portal, i.e. portal journalist, has its own writing style, and it would be necessary to do an additional linguistic analysis of the examples themselves in order to try to find a pattern in them. A more detailed linguistic analysis of the texts could potentially show us the limitations of the large language models themselves when identifying metaphors, which is beyond the scope of this paper and is suggested for future research.

Since the positive examples in this dataset are only the sentences in which metaphor was identified by both annotators, an additional analysis of the testing results was done in order to check how many negative examples were classified as positive by BERTić which at the same time were labeled as metaphor by only one annotator. In the test set there are 28 negative examples that were labeled as negative because only one of the annotators has identified metaphor in them and out of this pool 11 sentences were classifies as positive. Therefore, it would certainly be interesting to further explore this issue in future research and conduct experiments on a dataset that would include as positive, also the sentences where metaphor was identified by only one of the annotators.

Table 4. Results of separate classification of sentences from Jutarnji list and Index.hr The table shows precision, recall and F1 for positive and negative examples.

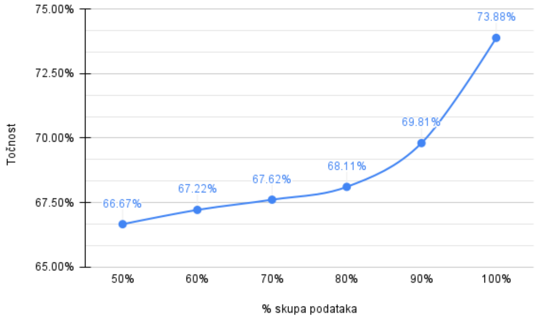
	Precision	Recall	F1
	Positive examples		
Jutarnji list	59,26%	78,05%	67,37%
Index.hr	66,20%	47,95%	55,62%
	Negative examples		
Jutarnji list	56,10%	34,33%	42,59%
Index.hr	71,98%	84,51%	77,74%

Table 5. Results of classification achieved by dividing data sets into a test set and testing according to data source.

	Precision	Recall	F1
Train set Index.hr, test set Jutarnji list			
Positive	73,51%	66,66%	69,92%
Negative	63,44%	70,65%	66,85%
Train set Jutarnji list, test set Index.hr			
Positive	52,57%	66,46%	58,71%
Negative	74,34%	61,84%	67,51%

	Preciznost	Odziv	F1
Treniranje Index.hr, testiranje Jutarnji list			
Pozitivni	73,51%	66,66%	69,92%
Negativni	63,44%	70,65%	66,85%
Treniranje Jutarnji list, testiranje Index.hr			
Pozitivni	52,57%	66,46%	58,71%
Negativni	74,34%	61,84%	67,51%

Kako bi se utvrdilo koliki je rast performansi modela ovisno o veličini skupa podataka, s BERTić modelom (5 epoha, 3e-5) napravljeni su testovi s 50%, 60%, 70%, 80% i 90% skupa, gdje je u svim skupovima omjer rečenica s Index.hr i Jutarnjeg lista bio isti kao u cijelom skupu podataka. Kao što možemo vidjeti na slici 1, točnost raste što je skup podataka veći te su najbolji rezultati ostvareni u 3 epohi, nakon čega rezultati padaju. Također možemo uočiti kako postoji veliki skok u točnosti između skupova podataka sa 90% i 100% primjera, što svakako ukazuje kako je potrebno napraviti dodatna istraživanja s većim skupovima podataka.



Slika 1. Usporedba točnosti modela ovisno o veličini skupa podataka.

6 Zaključak

U radu je izložena prilagodba MIPVU procedure za anotiranje metafore u hrvatskim novinskim člancima, kao i opažanja o anotiranju. Pokazalo se kako je, budući da je metafora kompleksan fenomen, anotiranje metafore zadatak koji zahtjeva mnogo vremena, ali i razumijevanje teorijskih okvira na kojima počivaju procedure anotiranja,.

Identificiranje metafore na razini rečenice modelom klasifikacije sekvenci teksta i korištenjem velikih jezičnih modela dalo je zadovoljavajuće rezultate. Najbolji rezultat ostvaren je korištenjem BERTić modela - F1 rezultat 70,75%. Analizom pogrešno klasificiranih pozitivnih primjera, uočeno je kako je model bolje klasificirao primjere s Jutarnjeg lista nego s Index.hr portala. Svakako je potrebna podrobnija lingvistička analiza pogrešno klasificiranih primjera, kako bi se u njima pokušao

In order to determine the increase of the model's performance depending on the size of the dataset, tests were made with the BERTić model (5 epochs, 3e-5) with 50%, 60%, 70%, 80% and 90% of the set, where in all sets, the ratio of sentences from Index.hr and Jutarnji list was the same as in the entire data set. As we can see in Figure 1, the accuracy increases as the dataset is larger and the best results are achieved in 3 epochs, after which the results drop. We can also see that there is a large jump in accuracy between the datasets with 90% and 100% examples, which certainly indicates that additional research with larger datasets is required.

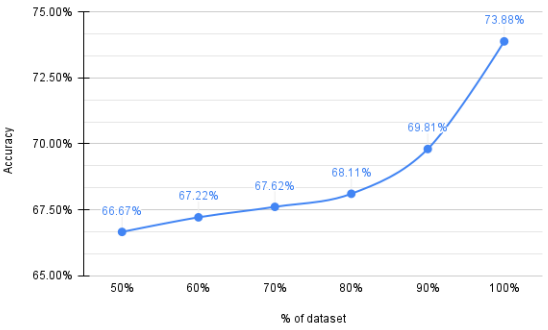


Figure 1. Comparison of model accuracy depending on dataset size.

6 Conclusion

The paper presents the adaptation of the MIPVU procedure for annotating metaphors in Croatian newspaper articles, as well as observations about the annotation. It has been shown that, since the metaphor is a complex phenomenon, annotating a metaphor is a task that requires a lot of time, but also an understanding of the theoretical frameworks on which annotation procedures are based.

Identifying metaphors at the sentence level with the text sequence classification model and by using large language models provided satisfactory results. The best result was achieved by using the BERTić model - an F1 result 70.75%. By analyzing the misclassified positive examples, it was observed that the model better classified the examples from Jutarnji list than those from the Index.hr portal. A more detailed linguistic analysis of misclassified examples is certainly required in order to try to find a pattern in them, which would potentially indicate the limitations of large language models.

A comparison of model results depending on the size of the dataset showed that the accuracy of the model increases with the size of the dataset, and a larger dataset will be prepared in future research.

Bearing in mind that metaphor is expressed on the level of a word (or a few words), the limitation of the model presented in this paper is that metaphor is identifyed on the level of a sentence. In further pronaći uzorak, a koji bi potencijalno ukazao i na ograničenja velikih jezičnih modela.

Usporedba rezultata modela ovisno o veličini skupa podataka pokazala je kako s veličinom skupa raste i točnost modela te će se u budućim istraživanjima pripremiti veći skup podataka.

Identificiranje metafore na razini rečenice je ograničenje modela predstavljenog u ovom radu, budući da se metafora izražava na razini riječi (ili nekoliko riječi) u rečenici. U daljnjim će se istraživanjima pristupiti i izradi modela koji identificiraju metaforu na razini riječi (tokena), odnosno leksičke jedinice.

Također, za naredna je istraživanja potrebno izraditi skup podataka koji sadrži akademske tekstove, posebice o književnosti i jeziku, budući da su prethodna istraživanja pokazala veliku zastupljenost metafore u tim tekstovima.

Reference

Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. O'Reilly Media, Inc.

Bogetić, K., Broćić, A., & Rasulić, K. (2019). Linguistic metaphor identification in Serbian. In S. Nacey, A. G. Dorst, T. Krennmayr, & W. G. Reijnierse (Eds.), Metaphor Identification in Multiple Languages (pp. 203–226). John Benjamins. https://doi.org/10.1075/celcr.22.10bog

Chen, X., Leong, C. W., Flor, M., & Klebanov, B. B. (2020). Go Figure! Multi-task transformer-based architecture for metaphor detection using idioms: ETS team in 2020 metaphor shared task. Proceedings of the Second Workshop on Figurative Language Processing. Association for Computational Linguistics, 235–243. https://doi.org/10.18653/v1/2020.figlang-1.32

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

Dankers, V., Malhotra, K., Kudva, G., Medentsiy, V., & Shutova, E. (2020). Being neighbourly: Neural metaphor identification in discourse. Proceedings of the Second Workshop on Figurative Language Processing. Association

research, the creation of models that identify metaphors at the level of words (tokens), or lexical units, will be approached.

Also, for subsequent research, it is necessary to create a dataset that contains academic texts, especially on literature and language, since previous research has shown a high prevalence of metaphors in these texts.

References

Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. O'Reilly Media, Inc.

Bogetić, K., Bročić, A., & Rasulić, K. (2019). Linguistic metaphor identification in Serbian. In S. Nacey, A. G. Dorst, T. Krennmayr, & W. G. Reijnierse (Eds.), Metaphor Identification in Multiple Languages (pp. 203–226). John Benjamins. https://doi.org/10.1075/celcr.22.10bog

Chen, X., Leong, C. W., Flor, M., & Klebanov, B. B. (2020). Go Figure! Multi-task transformer-based architecture for metaphor detection using idioms: ETS team in 2020 metaphor shared task. Proceedings of the Second Workshop on Figurative Language Processing. Association for Computational Linguistics, 235–243. https://doi.org/10.18653/v1/2020.figlang-1.32

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

Dankers, V., Malhotra, K., Kudva, G., Medentsiy, V., & Shutova, E. (2020). Being neighbourly: Neural metaphor identification in discourse. Proceedings of the Second Workshop on Figurative Language Processing. Association for Computational Linguistics, 227–234. https://doi.org/10.18653/v1/2020.figlang-1.31

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT 2019, 4171–4186. https://doi.org/10.48550/arXiv.1810.04805

Gao, G., Choi, E., Choi, Y., & Zettlemoyer, L. (2018). Neural Metaphor Detection in Context. Proceedings of the 2018 Conference on Empirical Methods in Natural Language

for Computational Linguistics, 227–234. https://doi.org/10.18653/v1/2020.figlang-1.31

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT 2019, 4171–4186. https://doi.org/10.48550/arXiv.1810.04805

Gao, G., Choi, E., Choi, Y., & Zettlemoyer, L. (2018). Neural Metaphor Detection in Context. Proceedings of the 2018 Conference on Empirical Methods in Natural Language

Processing, 607–613. https://doi.org/10.18653/v1/D18-1060

Gong, H., Gupta, K., Jain, A., & Bhat, S. (2020). IlliniMet: Illinois System for Metaphor Detection with Contextual and Linguistic Information. Proceedings of the Second Workshop on Figurative Language Processing. Association for Computational Linguistics, 146–153. https://doi.org/10.18653/v1/2020.figlang-1.21

Jojić, L., & Nakić, A. (Eds.). (2015). Veliki rječnik hrvatskoga standardnog jezika. Školska knjiga.

Katz, J. J., & Fodor, J. A. (1963). The Structure of a Semantic Theory. Language (Linguistic Society of America), 39(2), 170–210. https://doi.org/10.2307/411200

Krennmayr, T., & Steen, G. (2017). VU Amsterdam Metaphor Corpus. In Handbook of Linguistic Annotation, by Nancy Ide and James Pustejovsky, 1053–1071. Springer.

Lakoff, G., & Johnson, M. (1980). Metaphors We Live By. University of Chicago Press.

Lin, Z., Ma, Q., Yan, J., & Chen, J. (2021). CATE: A Contrastive Pre-trained Model for Metaphor Detection with Semi-supervised Learning. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, 3888–3898. https://doi.org/10.18653/v1/2021.emnlp-main.316

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. https://doi.org/10.48550/arXiv.1907.11692

Ljubešić, N., & Lauc, D. (2021). BERTić - The Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian. Proceedings of the 8th BSNLP Workshop on Balto-Slavic Natural Language Processing, 37–42.

Mao, R., Lin, C., & Guerin, F. (2019). End-to-End Sequential Metaphor Identification Inspired by

Processing, 607–613. https://doi.org/10.18653/v1/D18-1060

Gong, H., Gupta, K., Jain, A., & Bhat, S. (2020). IlliniMet: Illinois System for Metaphor Detection with Contextual and Linguistic Information. Proceedings of the Second Workshop on Figurative Language Processing. Association for Computational Linguistics, 146–153. https://doi.org/10.18653/v1/2020.figlang-1.21

Jojić, L., & Nakić, A. (Eds.). (2015). Veliki rječnik hrvatskoga standardnog jezika. Školska knjiga.

Katz, J. J., & Fodor, J. A. (1963). The Structure of a Semantic Theory. Language (Linguistic Society of America), 39(2), 170–210. https://doi.org/10.2307/411200

Krennmayr, T., & Steen, G. (2017). VU Amsterdam Metaphor Corpus. In Handbook of Linguistic Annotation, by Nancy Ide and James Pustejovsky, 1053–1071. Springer.

Lakoff, G., & Johnson, M. (1980). Metaphors We Live By. University of Chicago Press.

Lin, Z., Ma, Q., Yan, J., & Chen, J. (2021). CATE: A Contrastive Pre-trained Model for Metaphor Detection with Semi-supervised Learning. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, 3888–3898. https://doi.org/10.18653/v1/2021.emnlp-main.316

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. https://doi.org/10.48550/arXiv.1907.11692

Ljubešić, N., & Lauc, D. (2021). BERTić - The Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian. Proceedings of the 8th BSNLP Workshop on Balto-Slavic Natural Language Processing, 37–42.

Mao, R., Lin, C., & Guerin, F. (2019). End-to-End Sequential Metaphor Identification Inspired by Linguistic Theories. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 3888–3898. https://doi.org/10.18653/v1/P19-1378

Nacey, S., Dorst, A. G., Krennmayr, T., & Reijnierse, W. G. (2019). Metaphor Identification in Multiple Languages. John Benjamins. https://doi.org/10.1075/celcr.22

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global Vectors for Word Representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language

Linguistic Theories. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 3888–3898. https://doi.org/10.18653/v1/P19-1378

Nacey, S., Dorst, A. G., Krennmayr, T., & Reijnierse, W. G. (2019). Metaphor Identification in Multiple Languages. John Benjamins. https://doi.org/10.1075/celcr.22

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global Vectors for Word Representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1532–1543. https://doi.org/10.3115/v1/D14-1162

Peters, M. E., 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, 1, 2227–2237. https://doi.org/10.18653/v1/N18-1202

Pragglejaz Group. (2007). MIP: A Method for Identifying Metaphorically Used Words in Discourse. Metaphor and Symbol, 1, 1–39.

Rei, M., Bulat, L., Kiela, D., & Shutova, E. (2017). Grasping the Finer Point: A Supervised Similarity Network for Metaphor Detection. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 1537–1546. https://doi.org/10.18653/v1/D17-1162

Reidsma, D., & Carletta, J. (2008). Reliability Measurement without Limits. Computational Linguistics, 34(3), 319–326. https://doi.org/10.1162/coli.2008.34.3.319

Shutova, E. (2017). Annotation of Linguistic and Conceptual Metaphor. In Handbook of Linguistic Annotation, by Nancy Ide and James Pustejovsky, 1073-1100 (pp. 1073–1100). Springer. https://doi.org/10.1007/978-94-024-0881-2_40

Steen, G. J., Dorst, A. G., Herrmann, J. B., Kaal, A. A., Krennmayr, T., & Pasma, T. (2010). A Method for Linguistic Metaphor Identification. John Benjamins. https://doi.org/10.1075/celcr.14

Su, C., Fukumoto, F., Huang, X., Li, J., Wang, R., & Chen, Z. (2020). DeepMet: A Reading Comprehension Paradigm for Token-level Metaphor Detection. Proceedings of the Second Workshop on Figurative Language Processing. Association for Computational Linguistics, 30–39. https://doi.org/10.18653/v1/2020.figlang-1.4

Processing (EMNLP), 1532–1543. https://doi.org/10.3115/v1/D14-1162

Peters, M. E., 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, 1, 2227–2237. https://doi.org/10.18653/v1/N18-1202

Pragglejaz Group. (2007). MIP: A Method for Identifying Metaphorically Used Words in Discourse. Metaphor and Symbol, 1, 1–39.

Rei, M., Bulat, L., Kiela, D., & Shutova, E. (2017). Grasping the Finer Point: A Supervised Similarity Network for Metaphor Detection. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 1537–1546. https://doi.org/10.18653/v1/D17-1162

Reidsma, D., & Carletta, J. (2008). Reliability Measurement without Limits. Computational Linguistics, 34(3), 319–326. https://doi.org/10.1162/coli.2008.34.3.319

Shutova, E. (2017). Annotation of Linguistic and Conceptual Metaphor. In Handbook of Linguistic Annotation, by Nancy Ide and James Pustejovsky, 1073-1100 (pp. 1073–1100). Springer. https://doi.org/10.1007/978-94-024-0881-2_40

Steen, G. J., Dorst, A. G., Herrmann, J. B., Kaal, A. A., Krennmayr, T., & Pasma, T. (2010). A Method for Linguistic Metaphor Identification. John Benjamins. https://doi.org/10.1075/celcr.14

Su, C., Fukumoto, F., Huang, X., Li, J., Wang, R., & Chen, Z. (2020). DeepMet: A Reading Comprehension Paradigm for Token-level Metaphor Detection. Proceedings of the Second Workshop on Figurative Language Processing. Association for Computational Linguistics, 30–39. https://doi.org/10.18653/v1/2020.figlang-1.4

Ulčar, M., & Robnik-Šikonja, M. (2020). FinEst BERT and CroSloEngual BERT. In Lecture Notes in Computer Science (pp. 104–111). https://doi.org/10.1007/978-3-030-58323-1_11

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. Proceedings of the 31st Annual Conferenceon Neural Information Processing Systems, 6000–6010.

Wilks, Y. (1975). A preferential, pattern-seeking, Semantics for natural language inference. Artificial Intelligence , 6(1), 53–74. https://doi.org/10.1016/0004-3702(75)90016-8

Ulčar, M., & Robnik-Šikonja, M. (2020). FinEst BERT and CroSloEngual BERT. In Lecture Notes in Computer Science (pp. 104–111). https://doi.org/10.1007/978-3-030-58323-1_11

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. Proceedings of the 31st Annual Conferenceon Neural Information Processing Systems, 6000–6010.

Wilks, Y. (1975). A preferential, pattern-seeking, Semantics for natural language inference. Artificial Intelligence , 6(1), 53–74. https://doi.org/10.1016/0004-3702(75)90016-8

Wilks, Y. (1978). Making preferences more active. Artificial Intelligence, 11(3), 197–223. https://doi.org/10.1016/0004-3702(78)90001-2

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., … Rush, A. (2020). Transformers: State-of-the-Art Natural Language Processing. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Wilks, Y. (1978). Making preferences more active. Artificial Intelligence, 11(3), 197–223. https://doi.org/10.1016/0004-3702(78)90001-2

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., … Rush, A. (2020). Transformers: State-of-the-Art Natural Language Processing. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6