Eötvös Loránd Tudományegyetem
Társadalomtudományi Kar
MESTERKÉPZÉS

Measuring media bias through word embeddings

Konzulens:
Rakovics Zsófia

Készítette:
Gelányi Péter
KJZRFX
Survey statisztika és adatanalitika szak

2024. április

Contents

Abstract

Word embeddings offer a quantitative representation of words' semantic relationships. In my thesis, I explore their potential use in studying media bias and slant. The theoretical background of my work is embedded in both the literature on media bias and word embeddings. I detail my analysis of a newly collected Hungarian online media corpus. I fit multiple word embedding models, compare their performance, and use the best one to explore the semantic relationships of specific keywords across mediums and with elements of a sentiment dictionary. My results highlight both the advantages and drawbacks of word embeddings.

Keywords: word embeddings, NLP, news, media bias, media slant, partisan coverage, news imbalance, Word2Vec, FastText, Skip-Gram

1. Introduction

Media is often referred to as the fourth estate, the expression derives from the European concept of three estates of the realm and it is an explicit acknowledgment of the media's capacity for advocacy, and ability to shape political processes. In today's digitally interconnected world, where information inundates our daily lives, the role of news in politics may be greater than ever before. With the advent of digital platforms and the proliferation of online news sources, understanding the nuances of media bias has become increasingly vital. The role of media in shaping public opinion and influencing political processes cannot be overstated. From framing issues to selecting which stories to cover, media outlets wield considerable power in shaping public discourse and perceptions. Another important consideration is the availability of raw textual data that needs to be processed by researchers. Considering the recent explosive growth of available data, researchers need scalable, automated approaches to handle large quantities of unprocessed text.

The significance of comprehending media bias extends beyond academic inquiry; it directly impacts the functioning of democratic societies. Political processes, from electoral campaigns to policymaking, are intrinsically intertwined with media narratives. A nuanced understanding of media bias is therefore indispensable for grasping the complexities of contemporary political landscapes. Without such comprehension, efforts to analyze political

phenomena would be akin to navigating a labyrinth blindfolded. Central to the discourse on media bias is the concept of agenda-setting, wherein the news media determines not only what to report but also the salience and framing of issues. The agenda-setting function of news media plays a pivotal role in shaping public perceptions and policy priorities. By highlighting certain topics while downplaying others, media outlets exert a subtle yet profound influence on public opinion and policymaking (McCombs & Valenzuela, 2020).

In light of these considerations, this paper endeavors to contribute to the burgeoning field of media studies by harnessing the analytical power of word embeddings. Through the application of advanced computational techniques, I aim to uncover latent patterns of bias embedded within media texts, by dissecting the semantic structures of news articles. The purpose of this thesis is to explore the opportunities presented by word embeddings for the study of media bias.

In the following, I will first outline the theoretical background of my thesis, starting with the literature on media bias, and then I move on to the methodological review, where I discuss the development of word embeddings and their applications. Following the theoretical background, I describe the dataset that I collected and used in my research, the basis of its selection, its size, and dimensions, and I also detail the preprocessing methods I used, and my reasoning behind them. Then I outline my analysis, describe, and interpret my results and finally, I draw my conclusions from the research process and results, which is followed by a brief discussion on potential further avenues of research.

Of all the research that I processed during my work on this thesis the following three were the most influential: Efficient Estimation of Word Representations in Vector Space by Mikolov et al. (2013), Chapter 17 - Media Capture: Empirical Evidence in the Handbook of Media Economics (Enikolopov & Petrova, 2015), and Using word embeddings to probe sentiment associations of politically loaded terms in news and opinion articles from news media outlets (Rozado & Al-Gharbi, 2022).

2. Theoretical background

2.1. The literature on media bias

2.1.1. The theoretical background of media bias

The literature on media bias includes several definitions. In general definitions of bias converge around a couple of characteristics. Namely, media bias favors either political actors or ideologies through some slant that is present in the news coverage, and this slant in the news coverage is systematic.

„Partisan media bias, henceforth PMB for short, is a political or ideological slanting of the news in a way that favors, criticizes, emphasizes or ignores certain political actors, policies, events or topics." (Shultziner & Stukalin, 2021)

„All of the accounts are based on the same set of underlying facts. Yet by selective omission, choice of words, and varying credibility ascribed to the primary source, each conveys a radically different impression of what happened. The choice to slant information in this way is what we will mean in this paper by media bias." (Gentzkow & Shapiro, 2010)

„Bias can be defined as any systematic slant favoring one candidate or ideology over another." (Waldman & Devitt, 1998)

Measuring media bias is useful for several research areas, primarily political science, sociology, economics, and media studies. The motivation for systematically quantifying the degree of bias (through any given method) is twofold. First of all, it provides an objective method of comparison across time, and between media organizations. Secondly, it is necessary to investigate the potential relation of media bias with other relevant phenomena, such as incumbency in different positions of power (Gentzkow et al., 2015), or state-facilitated transfers (Szeidl & Szucs, 2021).

Another consideration before I review the different methods of measuring media bias is the limitations of the task and the limitation of the current literature's scope. Much of the literature concerns print media, this is likely in large part due to the easier accessibility of print news compared to other forms of news, such as television or radio, another such limitation is

the overrepresentation of U.S. news media in prior studies. Media bias is inherently qualitative and subjective. Tim Groeling in his review of the literature defines two main obstacles to the measurement of media bias, these are the problem of subjectivity and the problem of the unobserved population (2013). Although his work predates a significant portion of the studies that I review here I find that the obstacles he describes are rooted in theory, and are generally applicable even in the case of more up-to-date studies.

The problem of subjectivity concerns the fact that perceptions of bias are subjective and dependent on context such as the perceived origin of the news. This problem may be circumvented by automated methods that do not have prior preconceptions about what constitutes bias (Gentzkow & Shapiro, 2010). The problem of the unobserved population relates to editorial choices. If these choices are framed in terms of the language of statistics, then we would consider all potential news that could have made it into the coverage of the given medium as the population and the news that made it into the coverage as the sample. The problem is that the population is not available for researchers, only samples. One way to avoid this pitfall is to measure the bias of mediums relative to each other, not making claims about favoritism but rather about a medium favoring a given actor more or less, than other mediums.

The literature contains a wide and diverse set of methods. The selection of the methods has to reflect researchers' prior knowledge regarding the given media environment, which is under investigation, and their expectations regarding the type of media bias. Previously several typologies of media bias have been outlined.

„Bias can take a number of forms. A media outlet can be selective in what issues it covers (issue bias), what aspects of the issues it includes or excludes (facts bias), how the facts are presented (framing bias), and how it is commented (ideological stand bias). Distinguishing these different forms of bias is useful since determinants and effects are different." (Prat & Strömberg, 2013)

Filtering ad distortion as defined by Gentzkow and Shapiro can be loosely equated with the above categories (Gentzkow, Shapiro, et al., 2015). Where facts bias and framing bias relate to distortion and issue bias and ideological bias relate to filtering.

Partisan selection bias and presentation bias as defined by Tim Groeling concerns the news way information is filtered before it's included in news reports.

„Building on my definition of partisan media bias, the definition of partisan selection bias is choosing news stories that present a significantly distorted sample of reality that systematically and disproportionately favors one party over the other." (Groeling, 2013)

„I define partisan presentation bias as composing news stories in a manner that presents a significantly distorted view of reality, which systematically and disproportionately favors one party over the other." (Groeling, 2013)

Selection bias and description bias are also the terms used by Doron Shultziner and Yelena Stukalin in their study regarding the mechanisms of partisan media bias (2021). Naturally, there is considerable but not exact overlap among these categorizations.

2.1.2. Empirical methods for the measurement of media bias

Of the myriad ways to measure media bias, the simplest is to rely on the self-reporting of media organizations. This approach is contingent on the norms of the given media environment. Matthew Gentzkow and his fellow researchers exploit the fact that traditionally a large portion of U.S. newspapers (around half in the period under investigation 1869-1928) had explicit party affiliations, to investigate the effect of incumbency in executive offices on the press (2015). In cases where such an explicit affiliation is not present on the level of individual mediums, previous studies have exploited the presence of content that contains explicit positions regarding politically contentious subjects. Examples include the investigation of newspaper editorials (D. E. Ho & Quinn, 2007; D. Ho & Quinn, 2008). Similar studies survey newspaper endorsements (Ansolabehere et al., 2006) and also ballot propositions (Puglisi & Snyder, 2015).

Visual analysis is another popular approach to the measurement of bias. The advantage of it is that it investigates a different aspect of media reporting. At the same time, the disadvantage is the need for qualitative evaluation on the researcher's part, making it less scalable. Most of these studies examine newspaper pictures, but television reports are also

present (Banning & Coleman, 2009; Barrett & Barrington, 2005; Grabe & Bucy, 2009; Hehman et al., 2012; Kepplinger, 1982; Moriarty & Garramone, 1986; Waldman & Devitt, 1998).

A particular set of studies exploits the fact that newspapers rely on outside organizations, experts, and professionals for expert opinion and content. This necessitates some method of selection, in cases where the political or ideological leaning of these organizations, experts, and professionals is either given or can be confidently evaluated in their selection, platforming can be used to estimate the leaning of a given medium. Examples include comparing the airtime afforded to politicians from different parties (Durante & Knight, 2012), comparing citations of thinktanks by news organizations to their citations by members of Congress (Groseclose & Milyo, 2005), similarly comparing the citations of public intellectuals (Gans & Leigh, 2012), and finally comparisons on the reporting of pollsters (Groeling, 2008).

Another set of studies considers the ability of news media to set the political agenda, by investigating the subjects of articles and news reports systematically. David Niven has researched the reporting on congressional party switches (2003). Other examples include analysis on the subjects of the 2005 Iraq war (Aday, 2010), domestic presidential travel in the U.S. (Barrett & Peake, 2007), the internal conflict of political parties (Baum & Groeling, 2009), and also domestic social issues (Covert & Washburn, 2007).

Text mining approaches have also been adopted for the study of media bias; they are in general a good fit for this particular research question which often involves the processing of large amounts of raw text data. Most of the early text mining-based studies used lexical approaches (Lowry, 2008), particularly sentiment analysis (Eshbaugh-Soha, 2010). Gentzkow and Shapiro used a more sophisticated approach by deriving weights for the vocabulary of news coverage from congressional speeches based on party affiliation (2010)

The categories that I outline are not perfect, they can overlap, and do not account for all unique approaches to the empirical study of media bias, for example, experimental methods (Butler & Schofield, 2010). This is especially true for novel, contemporary approaches such as the ones that employ advanced NLP methods, such as aiding neural text classification methods with second-order information (Chen et al., 2020), deep learning methods (Hamborg,

2020), and other mixed approaches that also include elements of NLP (Hamborg et al., 2019; Kim & Johnson, 2022; Spinde et al., 2020).

2.2. Numeric representation of corpora

2.2.1. Frequency-based methods

Natural Language Processing (NLP) is a branch of artificial intelligence focusing on the interaction between computers and human language. It encompasses a variety of techniques and algorithms aimed at enabling computers to understand, interpret, and generate human language in a way that is both meaningful and contextually relevant. NLP draws from linguistics, computer science, and machine learning to process and analyze large volumes of natural language data. The applications of NLP are vast and diverse, spanning multiple industries and domains, it is applicable both in a research context as well as within the private sector (Khurana et al., 2023). The simultaneous explosion of available raw text data and the speedy development of NLP and text mining techniques have ensured their widespread dissemination.

For all NLP and text mining tasks, the conversion of text data, as in words to a format, is interpretable for algorithms, namely a numeric representation. There are a number of different approaches for the conversion of text data to a numeric form, these either represent the documents of the corpus as a matrix or as a vector of a matrix that represents the whole of the corpus. The simplest of these methods are One-hot encoding and the bag of word approaches.

One-hot is a vector of bits where the only legal combination is a single value of 1 and all other values are 0 regardless of order. Suppose we envision a document from a corpus, where the document is composed of n tokens and our corpus has a total of m unique tokens. In that case, we can represent this document as an ***m*** x ***n*** matrix, where each row represents a single token of the document, and the column length equals the number of tokens. Therefore, in the case of an example where our corpus is composed of a single document, the one-hot representation of that document would be an identity matrix.

Table 1: example one-hot encoding of the phrase: *„The sun shines bright."*

the	sun	shines	bright
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

Another popular way of representing text data numerically is to use a Document Feature Matrix DFM or Document Term Matrix DTM, the two terms are equivalent with the exception that DTM specifically refers to cases, where our tokens are simple terms instead of other options for example n-grams, or subword level units. DFM is a bag of words approach, these approaches get their name from the fact that they do not preserve the order of the words of a document, hence turning documents into a bag of words. Suppose that we have a corpus of ***m*** documents and the vocabulary of our corpus is composed of ***n*** unique words, in that case, the DFM form of that corpus would be an ***m*** x ***n*** matrix, where each document is represented by a row vector, while words are represented by a column vector, each value of the matrix is either a dummy encoding showing whether or not the given word is present or a number showing the number of occurrences of the word in the given document. In the latter case, the row vectors relating to each document are equivalent to simply summing the columns of the document's One-hot encoding matrix. Table 2 shows an example of a DFM based on a corpus composed of the following short sentences: „The sun shines bright.", „I like this weather.", „The weather is nice".

Table 2: Example of DFM representation

	the	sun	shines	bright	i	like	this	weather	is	nice
doc1	1	1	1	1	0	0	0	0	0	0
doc2	0	0	0	0	1	1	1	1	0	0
doc3	1	0	0	0	0	0	0	1	1	1

Unlike One-hot encoding, a DFM does not take the order of words into account, word order can be introduced into a DFM through the use of n-grams, where n consecutive units of text, in this particular case words are joined together (in all possible combinations) to form a token. As an example a n_gram, where ***n*** = 2 is called a bigram, the bigrams of the sentence: *„The sun shines bright."* would be the following: the_sun, sun_shines, shines_bright. Both One-hot encoding and DFM representation share two important problems. The first is that in both cases the resulting matrixes are very sparse, this can be mitigated by removing certain words or weighing them based on predefined criteria such as term frequency, and inverse document frequency (tf-idf). The second problem is that these representations of words, don't convey relevant information about their relations in essence they are just indices in a vocabulary.

2.2.2. Word embeddings - Word2vec

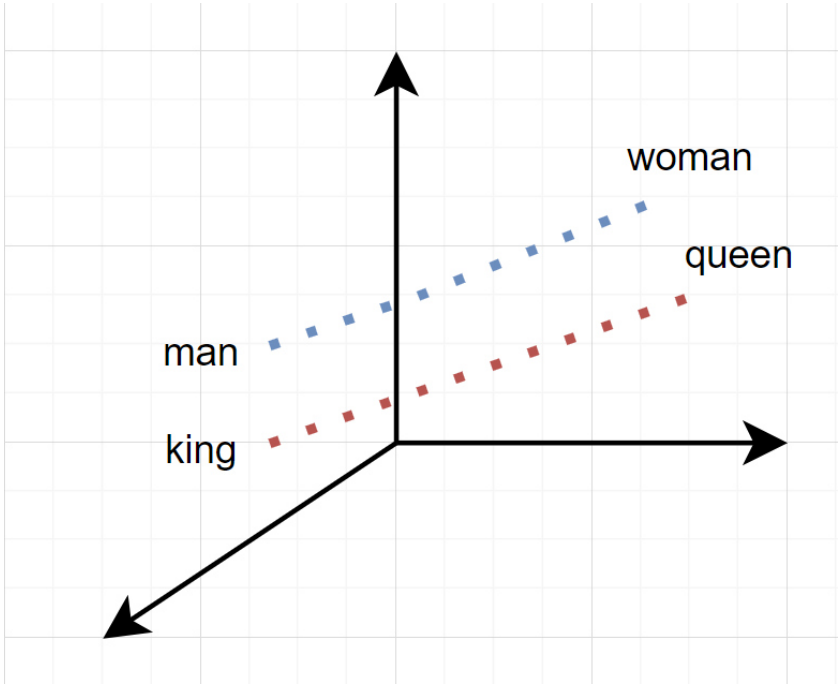
The novelty of word embedding methods is that word representations produced in such a way contain important information about words' syntactic/morphological and semantic similarities. In essence, we want to place words in a vector space, where each word within our vocabulary has a corresponding vector, and the vectors of „similar" words are similar to each other. Embeddings can represent words as vectors. the similarity/distance of these vectors indicates their semantic and or morphological similarity. Words embedded in this vector space can have multiple degrees of similarity (Mikolov et al., 2013). Furthermore, relations of terms can be reflected by simple algebraic operations, this was shown first by Tomas Mikolov et al.

with a now well-known example, in their paper, it is shown that subtracting the vector representation of „men" from the vector representation of „king" and adding the vector representation of „woman" yields a vector to which the closest vector representation within the vocabulary belongs to the term queen.

„king" – „man" + „woman" = „queen"*

We can also depict this relationship within the vector space with a simple illustration.

Figure 1.: Assumed relation of the words: king, man, woman, and queen in a well-fitted vector space (only illustration)



Source: personally edited.

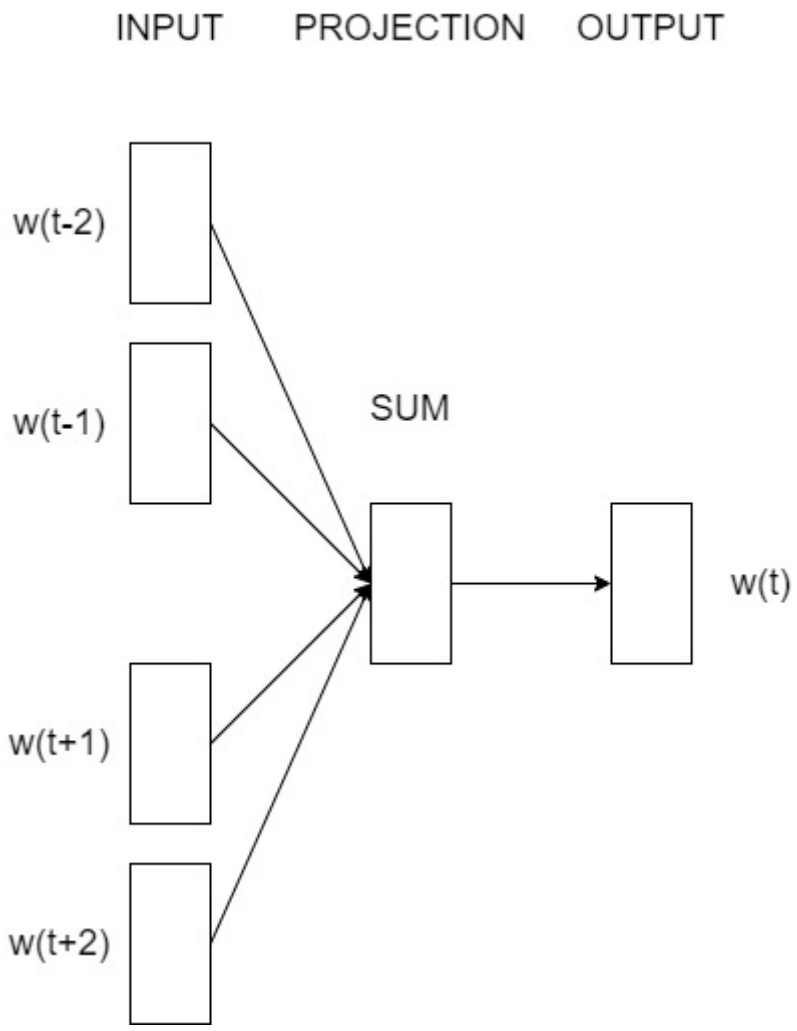
Before detailing specific word embedding techniques, it is important to mention a few precursor techniques that primarily originate from the field of Distributional Semantic Models. Before word embeddings, a number of models have been proposed for the continuous representations of words. Latent Semantic Analysis also known as Latent Semantic Indexing (LSA) alongside Latent Dirichlet Allocation (LDA) uses documents as context to capture semantic relatedness (Blei et al., 2003; Dumais, 2004). Techniques that also warrant mentioning include Self Organizing Maps (SOM) (Ponmalai & Kamath, 2019), and Simple

Recurrent Networks (SRN) (Elman, 1991). The latter of these two can be considered as a precursor to neural language models.

Word2vec by Mikolov et al. was a significant development in the field of word embeddings (2013). In their paper, they reflect on two different models of NNLM the Feedforwards Neural Net Language Model (Bengio et al., 2000) and the RNNLM Recurrent Neural Net Language Model (Bengio & LeCun, 2007). An NNLM model is composed of input, projection, hidden, and output layers. ***N*** represents the number of previous words that are One-hot encoded, meaning we have ***N*** different vectors, and ***V*** denotes the size of the vocabulary, ergo the length of our ***N*** vectors. This input layer is then projected to the projection layer ***P***. Most of the computational cost comes from the interaction of the ***P*** projection layer and the ***H*** hidden layer. The ***H*** hidden layer is used to compute a probability distribution over all ***V*** words of the vocabulary. Mikolov et al. used hierarchical softmax to reduce the computational cost of calculating probability distributions over ***V*** words (Mikolov et al., 2011; Mnih & Hinton, 2008; Mohammed & Umaashankar, 2018; Morin & Bengio, 2005), therefore most of the computational cost in the model comes from the interaction of ***P*** and ***H***. RNNLM models do not have a projection layer, they get their name from the recurrent matrix that connects the hidden layer to itself.

Mikolov et al. suggest two simpler models, than previous neural network-based models CBOW as in Continuous Bag-of-Words and Continuous Skip-Gram. Continuous Bag-of-Words disposes of the hidden layer in neural network-based word embeddings thereby making the computational cost of word representation in vector space significantly lighter. The projection layer of CBOW is shared for all words as it is shown in Figure 2. meaning that all of the surrounding context words get projected into the same position. This in turn means that the particular order of these words does not affect the embedding. Unlike the standard bag of words, model CBOW uses a continuously distributed representation of the current word context, thus the name Continuous Bag-of-Words. Another important detail that differentiates CBOW from NNLM is the fact that the window around the ***t***-th word also includes future words and not just the previous words. Figure 2 shows the structure of CBOW it predicts the ***t***-th word based on an arbitrarily sized window (in this particular case 4).

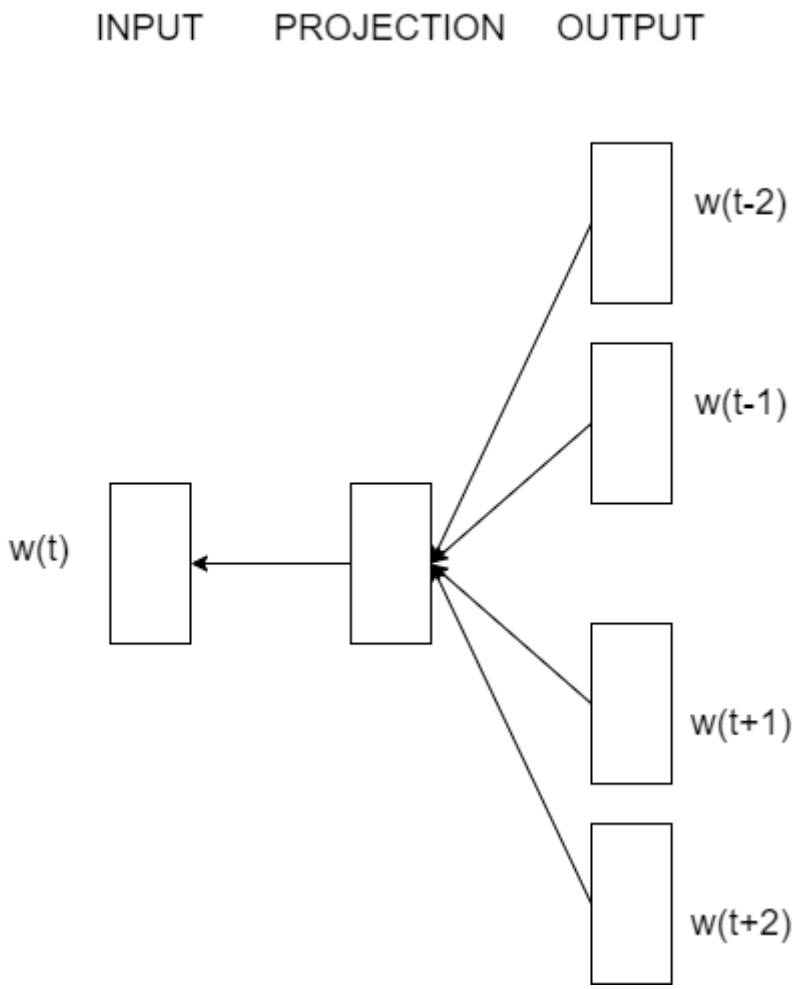
Figure 2.: Structure of the CBOW model.



Source: personally edited, based on (Mikolov et al., 2013)

The Skip-Gram model is very similar to CBOW, but it is reversed, while CBOW creates word vectors by trying to predict a target word based on a surrounding word with 2 2-layer neural network, Skip-Gram predicts surrounding words based on the target word.

Figure 3.: Structure of the Skip-Gram model.



Source: personally edited, based on (Mikolov et al., 2013)

The models presented by Mikolov et al. contain a trade-off, the removal of the hidden layer from the structure, this results in simpler models that are not able to represent data as precisely as neural network-based models but are a lot less computationally costly therefore making it possible to train vector representations of words on large corpora much more efficiently. For an in-depth description of CBOW's and Skip-Gram's parameter learning see (Rong, 2016).

2.2.3. Word embeddings - GloVe

In 2014 Jeffrey Pennington, Richard Socher, and Christopher D. Manning presented GloVe as in Global Vectors for Word Representation (Pennington et al., 2014). In their work, they divide

the approaches to the creation of word vectors into two general categories. Matrix factorization methods such as Latent Semantic Analysis (Dumais, 2004), and local context window methods. The previously discussed Skip-gram and CBOW belong in the latter category since they create representations of words within vector space based on a predefined ***n*** context window alone. They note that although the context window-based approaches perform better on the word analogy tasks devised by Mikolov et al. (2013) they do not efficiently leverage the statistical information that is present in the corpus.

GloVe aims to improve on this shortcoming by capturing the corpus statistics directly, hence the name Global Vectors. GloVe uses an FCM Feature Cooccurrence Matrix. In an FCM denoted by ***X,*** the value of X_{ij} denotes the number of times that the ***j***-th word occurs in the context of the ***i***-th word, while X_i is the number of times that any word appears in the context of the i-th word.

$$Xi = \sum_{k} Xik$$

P_{ij} is then the probability that the ***j***-th word occurs within the context of the ***i***-th word.

$$\mathrm{P_{ij}} = \mathrm{P(j|i)} = \frac{X_{ij}}{X_i}$$

Let's take a simple example of an FCM, where we consider the context to be the two adjacent words, an FCM like that, for the following sentences: „I like this weather", and „The weather is nice" is shown in Table 3, since a word is not considered to be in its own context, therefore the diagonal of the matrix contains only zero values.

Table 3.: Example for a Feature cooccurrence table

	i	like	this	weather	the	is	nice
i	0	1	0	0	0	0	0
like	1	0	1	0	0	0	0

this	0	1	0	1	0	0	0
weather	0	0	1	0	1	1	0
the	0	0	0	1	0	0	0
is	0	0	0	1	0	0	1
nice	0	0	0	0	0	1	0

Pennington et al. suggest that instead of cooccurrence probabilities the difference in cooccurrence probabilities should be the basis for calculating word vectors (***w***), the writers give an intuitive example based on the relation of three words: ice, steam, and water. The words' ice and steam are both related to water, and neither is related to random words, such as: „fashion". They are distinguished by words that only relate to one of them such as solid and gas. The ratio of cooccurrence probabilities is better able to distinguish the relevant words since the strong correlation of ice and steam with water cancels each other out in the ratio, hence the argument for using the ratios of cooccurrence probabilities, instead of the probabilities themselves.

The basic concept is that a given F function based on the word vectors, which are the inputs of F, gives the ratio of cooccurrence probabilities.

$$F(w_i,\ w_j, w_k) = \frac{P(k|i)}{P(k|j)}$$

This basic structure of the ***F*** function is then further elaborated on. Given that word vectors are linear a natural start for ***F*** would be subtraction, and then the dot product is taken so that ***F*** does not mix the dimensionality of the vectors. Giving us the following function.

$$F\left((w_i - w_j)^T\ w_k\right) = \frac{P(k|i)}{(p(k|j)}$$

Here the use of subtraction mirrors the concept of analogy, which has been first showcased by the word2vec models. Since in the case of an FCM, the role of the word and context word is interchangeable as it is shown in Table 3, this would mean that we can exchange both w_k and w_i but also X and X^t, therefore $\boldsymbol{F}$ has to be a homomorphism.

$$F\big((w_i - w_j)^T\, w_k\big) = \frac{F(w_i^T\, wk)}{F(w_j^T\, wk)}$$

This then can be solved the following way:

$$w_i^T\, w_k = \log(P_{ik}) = \log(X_{ik}) - \log\,(X_i)$$

Following this step $\log(X_i)$ is absorbed into a bias term (this is possible since $\log(X_i)$ is independent of k), then adding a bias term b_k to w_k restores the symmetry.

$$w_i^T\, w_k + b_i + \, b_k = \log(X_{ik})$$

Finally, a weighing function is added $f(X_{ij})$ to the cost function:

$$J = \sum_{i,j=1}^{V} f\big(X_{ij}\big)\,[w_i^T\, w_j + b_i + \, b_j - \, \log(X_{ik})]^2$$

Besides the novel use of FCM GloVe also differs from CBOW and Skip-Gram by virtue of using negative sampling instead of softmax. The GloVe model is capable of providing vector representations of words that are on par with or perform better than previous models based on comparison with the use of word similarity datasets(Pennington et al., 2014).

2.2.4. Word embeddings - Fasttext

Fasttext is an improvement on the previously showcased word2vec architectures CBOW and Skip-Gram. As the title suggests „Enriching word vectors with subword information" the novelty of the Fasttext model is the inclusion of subword information which in this case means n-grams that are composed of letters. While previously mentioned embedding techniques represent words as vectors, Fasttext instead represents character-level n-grams as vectors and word vectors are derived as the sum of these n-gram vectors.

The advantage of this new approach is twofold. On the one hand, Fasttext models can handle OOV as in out-of-vocabulary cases, where we want to represent a word in vector space that was not present in our training data. Previously mentioned models would have to ignore that particular feature in the corpora, but Fasttext can represent it based on the n-grams that are present within the word. The second advantage is that this model explicitly incorporates morphological features of the vocabulary into the embedding process. Word2vec and Glove implicitly account for morphological features as the embeddings of inflected words and their original form will unavoidably have similar embeddings. Fasttext's model should allow better results, especially in the case of morphologically rich languages.

The use of n-grams necessitates the inclusion of two further parameters into the CBOW and Skip-Gram architectures (both of which are compatible with Fasttext) and these are the minimum and maximum numbers for n. For example, the word „weather" would have the following n-grams if the minimum value is 2, and the maximum value is 4.

<we>, <ea>, <at>, <th>, <he>, <er>, <wea>, <eat>, <ath>, <the>, <her>, <weat>, <eath>, <athe>, <ther>

2.2.5. Context-dependent Word Embedding Models

The previously detailed three-word embedding models are context-independent, meaning that although the word vectors are learned based on the context of words, but the embedding of individual words is not dependent on the context. The theoretical background section of this thesis focuses on context-independent word embedding models because context-dependent word embeddings are not applicable for the purposes of my analysis for reasons that I detail in the analysis section of my thesis. Despite this, it is important to mention context-dependent models, since these are the current state of the art when it comes to embedding models.

The context of words includes a great deal of useful information. The most obvious example of this is the case of homographs words with the same form but different meanings, and also polysemes words that have different but related senses, examples include lead or fair. These words are represented by the exact same string of characters, but depending on the context could have different meanings.

„Lead the way!”

„These weights are made out of lead.”

„We should go to the fair.”

„This is not fair at all.”

Similarly, when it comes to sarcasm, we may mean the exact opposite of a word we use, which can only be detected through the word's context. Context-independent word embeddings create a matrix with the vector representations of each word present in the training data's vocabulary. Of course, FastText can represent words outside of the vocabulary, but all of the previously mentioned word embedding methods represent the same string of characters with the same vector, while context-dependent word embeddings utilize the information in the word's context to create different continuous vector representations for the same word. Examples of context-dependent word embeddings include ELMo (Peters et al., 2018), BERT (Devlin et al., 2019), and GPT-2(Radford et al., 2019).

2.3. Studying media bias with word embeddings

In the intersection of the scientific literature on word embeddings and media bias is a set of studies that employ word embedding models to study media bias. The use of these models is attractive for such research purposes because they record the semantic relationships of words based on large corpora. These semantic relationships reflect the biases present in texts. Previous research has studied and documented how word embeddings reproduce the biases present in their training data (Bolukbasi et al., 2016; Caliskan et al., 2017; Ethayarajh ET AL., 2019; Llorens, 2018). Measures and tests have been suggested for the detection of bias and the debiasing of word embeddings (Elsafoury et al., 2022; Papakyriakopoulos et al., 2020). This characteristic of word embeddings is obviously relevant due to their use in text classification tasks and generative AI, allowing for these biases to affect the results of these tasks.

While the encoding of bias into word embeddings' vector spaces is a hurdle in many NLP tasks it is also a possible tool for researchers to investigate gendered, cultural, and other biases in texts. One possible approach is to use the information present in the direction of

word embeddings by computing an issue-specific subspace, in effect this means nothing more, than taking two relevant terms for example men and women taking the difference of their vector representations, which we can consider to be the direction of gender bias (Durrheim et al., 2023). We can make our bias more robust by taking other terms, to stick with the example of gender let's say she, her, and he, him. We can then take the centroid of terms related to females and the centroid of terms related to males and consider their differences as the direction of gender bias. This approach lends itself to the investigation of gender biases since the two dimensions and their corresponding definitional terms are given (Friedman et al., 2019; Wevers, 2019) but it has also been used to investigate political bias where the selection of definitional terms is far from obvious. Gordon et al. use word pairs that different political groupings use to describe the same phenomena as definitional terms (2020). This approach also leaves open the possibility of computing multiple axes, as in multiple dimensions of bias. David Rozado and Musa al-Gharbi use a sentiment dictionary to measure the distance of politically relevant words from the negative and positive terms in the sentiment dictionary, their results based on the embedding correlated substantially with human evaluations of bias (2022). This approach is analogous to a directed sentiment analysis on the level of an entire corpus.

2.4. Previous related theses from the faculty

Several previous theses from this faculty have concerned themselves with word embeddings or employed them for specific research questions. Below are a few examples that I have found most useful and/or analogous to the subject of my own thesis. The most common use of word embeddings is as an input for other NLP tasks, and this is reflected in the representation of word embeddings within previous theses. Zsolt Zsabó uses word embeddings for the clustering of universities' web pages (2020). Flóra Bolonyai uses different approaches for a gender-based classification of online texts word embeddings are but one method of numeric representation of texts in her work (2019) similarly, Máté Buda also uses word embeddings to represent texts in a text classification problem (2020)

Enikő Csaba combines word embeddings and Procrustes transformation to bring the vectors of two different word embeddings into the same vector space (2023), her work also relates to mine since the texts she uses are also from Hungarian online media. Finally, Márton

Varsányi uses word embeddings to study the context of terms related to agrarian production in historic Hungarian texts (2020)

3. Data

3.1 Selection of the dataset and its context

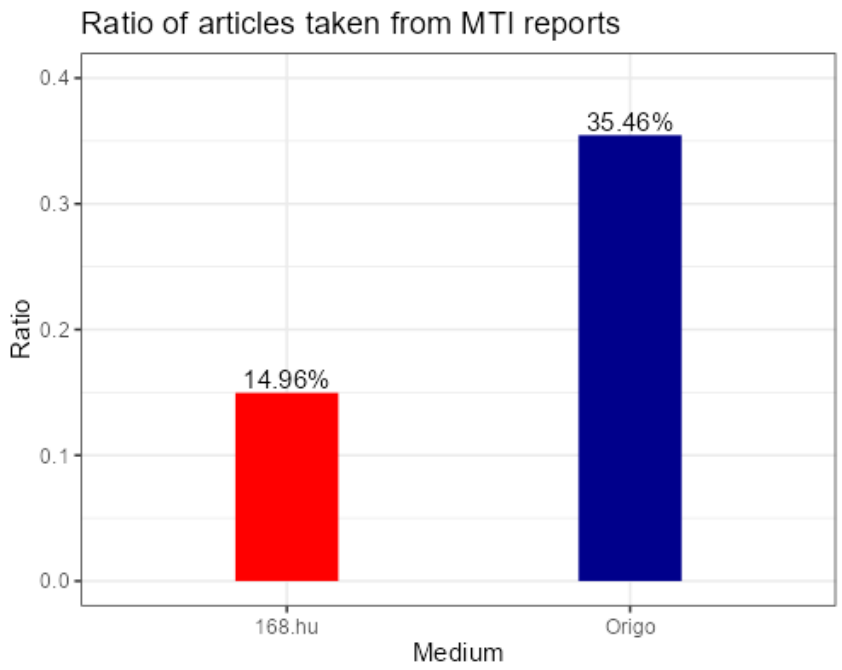
When it came to collecting my data for the analysis, I based my choices on the purpose of the thesis as previously defined. Since I wanted to explore the use of word embeddings in the study of media bias, focusing on the methodological aspects rather than an exact case study, therefore I did not intend to collect corpora that would reflect the entirety of Hungarian online news sources, but rather one that is generally representative of a sub-section of online news in Hungary. My criteria for the inclusion of articles were that they should cover Hungarian political news, and I aimed to collect a set of corpora that diversely covers Hungarian news, I also considered the comparability of their editorial style, the structure of the sections into which the news content is divided, and the readership of the mediums. Finally, I had to also take into account the time period of the collected articles. I chose to select a larger 5-year window for the text collection so that I would have more observations for each medium included in the corpus, the corpus covers the entire time period beginning with 2015 up until the end of 2019. A disadvantage of collecting articles from such a long time period is not just the time and computational cost, but also the fact that mediums that do not have overlapping content, or do not have content from the entire period cannot be included, such as Telex.hu a popular online Hungarian news site founded in 2020.

Based on these aforementioned criteria I have selected the following four mediums. Origo, 24.hu, 168 óra, and Index. Origo is one of the best-known news sites in Hungary, it was founded in 1998, and originally owned by Hungarian Telekom in 2015 it was sold to New Media Inc. even prior to the change in ownership the news site was allegedly under government pressure (Magyari Péter, 2014). Under new management, the site has gone through a marked reorientation in terms of its news content's slant. While previously it was often critical of the Hungarian government, from 2015 it could be considered a government-friendly news site. 24.hu (its original name: Hír24) is an online news site founded and originally owned by Sanoma Budapest Corporation in 2010. In 2011 the FN.hu news site, which mainly focused on business

and economic news was incorporated into 24.hu. The publisher of 24.hu was purchased in 2014 and was subsequently renamed to Central Médiacsoport. The news site received its current name „24.hu" in 2016. 168.hu is the online news site equivalent of 168 óra a weekly Hungarian newspaper. In 2022 the production of the physical newspaper has been suspended, for now only the online version remains. Index.hu was founded in 1999 by the editorial staff of the popular news site iNteRNeTTo. Around 2020 similarly to Origo Index experienced significant changes in management that were later accompanied by the resignation of much of the editorial staff and public demonstrations (hvg.hu, 2020), but this shift in terms of the news content's slant is outside of this analysis' timeframe.

An important part of the Hungarian media environment is MTI as in Magyar Távirati Iroda, Hungarian Telegraphic Office. MTI was founded in 1880, it is one of the oldest news agencies in the world. The importance of MTI goes beyond the scope of historical significance. Its news reporting regularly appears in the news content of online Hungarian media. The percentage of articles taken from MTI differs from medium to medium based on editorial style and decisions. One potential way to find the ratio of news reports that are original MTI content is based on the author information of articles. Two mediums in my corpus (168.hu, Origo) include MTI in their author section, this means that in their cases we can get the ratio of news reports taken from MTI shown in Figure 4. These numbers also give us a loose idea about the potential ratio of MTI news reports in other mediums.

Figure 4.: Ratio of news reports taken from MTI in the reporting of 168 and Origo based on the author variable



Naturally, the effect of taking news from MTI's reporting will be homogenization across Hungarian news sites. An argument could be made that such content should be filtered out of the corpus, this would only be possible through a complicated text classification analysis. After some consideration, I have decided against such filtering for several reasons. News content taken from MTI is presented the same way to readers as other articles, filtering out MTI content would mean measuring bias and polarization on a dataset that significantly differs from the Mediums' actual content as in the content that is available to readers. Mediums frequently reflect on each other's news content not just MTI's.

3.2 Data collection process

I have collected my corpora through web scraping. Web scraping is the process of collecting and organizing unstructured data from the internet into a structured form that can feasibly be used in research (Sirisuriya, 2015). My process for all four mediums was the same and only differed in minor details. This process can be broken down into the following three steps. First, I collected the links to the articles themselves; this first step would involve either the simulation of user input to access later and later articles or looping through a list of links that

I generated based on the structure of the website. This process can be validated based on the number of links per page, or by comparing the number of links for each date, where a period without any articles would indicate a faulty collection process. After I had gathered the links in the second step I collected the HTML information from these links. Finally, I had to find the right segmentation for the HTML structure for all of the mediums. I first wrote a preliminary segmentation based on a single article's HTML structure, and then I tested it through an iterative process that involved testing the segmentation based on random sampling and also by examining cases where my collected variables either contained missing values or had suspiciously short strings. I only stopped the iterative changes in the segmentation after I found no more errors.

3.3 Description of Preprocessing

The process I used to prepare my corpora can be divided into two separate parts, the selection of articles as in filtering and text preprocessing. Filtering meant the removal of all articles outside of the previously outlined time period, and also several article types that could not be used for this current research, such as articles that only included pictures or video content and no text, newsfeed style ongoing reports were also excluded, these type of articles are collections of shorts post that provided constant reporting on a significant ongoing event, such as elections or demonstrations. Another important decision in terms of the article selection was handling the different section structures of the 4 Mediums, as I previously mentioned the similarity of section structure was a consideration during the selection of Mediums, naturally, there are still differences in the section structures shown in Table 4.

Table 4.: Overlaps in the section structure of the four medium as of March 2024

Origo	Index	168.hu	24.hu
National News	National News	National News	National News
International news	International news	International news	International news
Sport	Sport	Sport	Sport
Economy	Economy	Economy	Economy
Culture	Culture	Culture	Culture
-	-	Lifestyle	Lifestyle
-	-	Entertainment	Entertainment
-	Opinion	-	Opinion

-	Science/technology	Science/technology	
Science	-	-	Science
Television	-	-	-
Films	-	-	-
Travel	-	-	-
Cooking	-	-	-
-	FOMO	-	-
-	Blog	-	-
-	Video	-	-
-	Podcast	-	-
-	-	Creative	-
-	-	Gardening	-
-	-	Crime	-
-	-	-	Technology
-	-	-	Podcast
-	-	-	Civic life
-	-	-	Europe
-	-	-	Business
-	-	-	Video
-	-	-	Championship
-	-	-	Countryside
-	-	-	Lists
-	-	-	Religion

There is significant overlap in the different section structures: national-level news, international news, sport, economy, and culture are ubiquitous categories, but there are also a number of smaller section categories that partially overlap or not at all. The structure of the sections is important for a true like-with-like comparison, this problem is further compounded by the fact that the section structure of one Medium is subject to change over time. One approach would be to take the entirety of the news content regardless of sections. Another would be to select articles only from overlapping parts of the section structure. I have chosen to only include sections that correspond to the category of national-level news, usually named „Belföld" or „Itthon" in Hungarian mediums. The category of national news is omnipresent both over time and across publications, most articles generally fall under this category, and finally, national-level news is the section I would consider most relevant for the study of media bias.

The article texts have been preprocessed in two separate steps before tokenization, lemmatization, and removal of stopwords. For lemmatization, I used the freely accessible

huspacy module (Orosz et al., 2017/2023). Lemmatization is the algorithmic method of determining the lemma (the dictionary form) of inflected words(Balakrishnan & Lloyd-Yemoh, 2014). For the removal of stopwords, I used Poltextlab's list, which is available in the Hunminer R package (*HunMineR*, 2021/2023). Below is a random example of a short article's text before and after preprocessing.

Unprocessed/raw text:

„Pénteken kora délután egy időre sem belépni, sem kilépni nem lehetett Magyarországról az informatikai rendszer meghibásodása miatt. Erről az Országos Rendőr-főkapitányság kommunikációs szolgálata tett bejelentést 12 óra 33 perckor a rendőrség honlapján. 13 óra 14 perckor aztán érkezett a hír, hogy újra lehet használni a határátkelőhelyeket, miután megszüntették az informatikai rendszerben keletkezett hibát.\nRendőrségi közlések nyomán az MTI azt írta, a hiba miatt péntek délelőttől átmenetileg szünetelt a ki- és beléptetés a határátkelőhelyeken, a közúti átkelőhelyek közül több helyen kellett ki egy-kétórát várni\nAz informatikai rendszer hibája a Liszt Ferenc-repülőteret is érintette."

Preprocessed text:

„péntek kora délután idő belép kilép magyarország informatikai rendszer meghibásodás országos rendőr főkapitányság kommunikációs szolgálat tesz bejelentés óra perc rendőrség honlap óra perc érkezik hír használ határátkelőhely miután megszüntet informatikai rendszer keletkezett hiba rendőrségi közlés nyomán mti ír hiba péntek délelőt átmeneti szünetel beléptetés határátkelőhely közúti átkelőhely hely kétórát vár informatikai rendszer hiba liszt ferenc repülőtér érint"

As we can see in this example the lemmatizer finds the original forms of the inflected words in most cases. Table 5. describes the changes in the vocabulary from the lemmatization.

Table 5.: Changes of the vocabulary after preprocessing

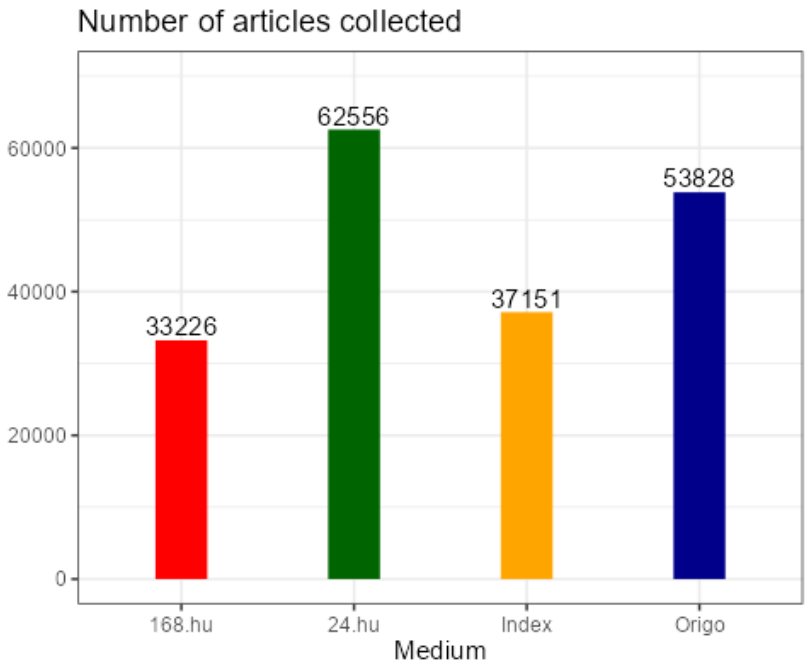
	Number of unique elements in vocabulary	The mean number of appearances in documents	The mean number of appearances in the corpus
Preprocessed text	378193	52,411	71,500
Raw text	1001234	31,072	54,243

I expect that the preprocessing should make my models more robust by simultaneously decreasing the number of unique observations in my corpora and also increasing the number of occurrences for these observations.

3.4 Size and dimensions of the dataset

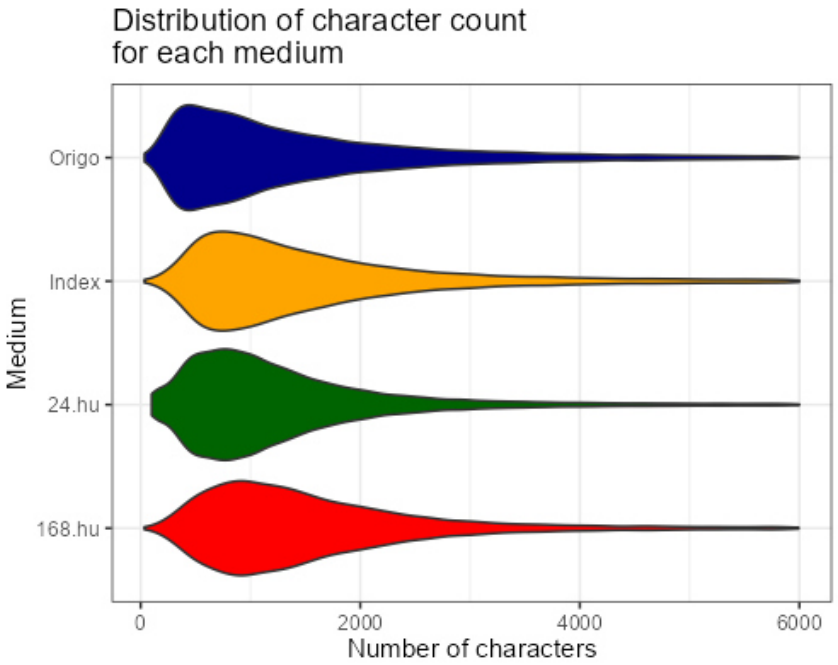
After the corpus has been prepared as described above 186761 observations remained in my dataset. Figure 5 shows the number of observations in each subcorpora belonging to the 4 mediums. The number of observations is roughly comparable.

Figure 6.: Sizes of the subcorpora belonging to each medium



Beyond a simple article count another important point of comparison is the length of articles, the distribution of different length values is shown in Figure 7. The distributions are relatively similar (the largest article length values have been removed for better interpretability).

Figure 7.: The distribution of raw unprocessed text length in character count for each medium



We can see from the figure that the distribution of article length is similar across the 4 different subcorpora.

For each observation of my corpora, I have collected 7 different variables. These are: title, date, author, lead, link, text, source. Later these 7 variables were turned into 8 when I created a raw and a preprocessed version of the text variable. The title contains the title of the given article, date is the date of the article's publication as noted on the website, it accounts for the original publication and not for any later changes. The author is the article's author, this information is in many cases not accessible, or simply the medium's name is given as the author. Lead is the short text before the main text of the article, it is generally stored in a different node of the HTML structure than the main text, not all articles have a lead. The text variable contains the full text of the article, meaning both the lead and the main body of text, this has been used for the analysis after preprocessing the text variable has been divided into

two variables, text for the preprocessed version and raw_text for the text's original version. Finally, the variable source denotes the medium that has published the given article.

4. Analysis

4.1 Steps of the analysis

The Analysis of this thesis is composed of 6 steps including the preprocessing of the collected data as described above. Following this I fit several word embeddings with different models, architectures, and window sizes. In the third step, I evaluate the performance of these models against each other. Both the exact parameters of the fitted models, a discussion on the comparison of word embeddings, and the exact result of the models' comparison are included in the following segment of the thesis.

The best-performing model is then selected. This model is fitted once more on a slightly altered corpus. The corpus is altered by adding tags to a selected set of words, each tag is a string of characters that denotes in which medium the word appeared, for example in a case where the word „Fidesz" is selected it would have 4 different forms after being tagged: „168_Fidesz", „24_Fidesz", „index_fidesz", „origo_Fidesz". The point of tagging is to differentiate the occurrences of relevant words in mediums, the definition of relevance is dependent on the specific research question in every case, this Thesis is methodologically focused, and concerns itself with the application of word embeddings for the research of media bias, therefore I aimed to collect a set of politically relevant words that are as general as possible. The dictionaries used in this Thesis will be detailed in the following sections.

The point of differentiating the appearances of a selected set of words in our corpora is that it allows the comparison of the word vectors of the same word in each medium. Measuring the distance between occurrences of the same word's vector in different mediums relates to both partisan selection bias and partisan presentation bias as defined by Groeling (2013). The vector representation of a word is reflective of the given word's context, which is dependent on both the selection of the news, which news stories (related to that concept) make the cut, and also it is dependent on the presentation of how these news stories are thematized. The distance between tagged word's vectors shows us the relative differences in the given word's context across the collected mediums. It functions like an issue-specific

measure of news polarization. The distances between the same word in different mediums can be visualized with the help of simple dimension reduction methods such as SVD Singular Value Decomposition (Baker, 2005), and UMAP Uniform Manifold Approximation and Projection for Dimension Reduction (McInnes et al., 2020).

I also measure the distances between the tagged words and elements of sentiment dictionaries. This helps to interpret the differences in reporting that we observe. Finally, I repeat the tagging process and the measurement of vector distances with a randomly selected set of words from the vocabulary, so that my results can be checked against a set of control terms for validation.

4.2 Word embeddings, parameters, and the method of comparison

Unlike many other NLP tasks, word embeddings do not have a straightforward measure for their quality. The literature on the evaluation of word embeddings differentiates two ways of testing the performance of word embeddings extrinsic and intrinsic methods (Bakarov, 2018; Jastrzebski et al., 2017; Schnabel et al., 2015). Extrinsic evaluations test word embeddings based on their usefulness as feature vectors in downstream tasks. Word embeddings can be used potentially in any NLP task. The idea is that by performing the same NLP task in the same way with the only difference being the word embedding used for the representation of the corpus, the accuracy or quality of the given NLP task's result will be reflective of the quality of the word embeddings that are being compared.

Amir Bakarov (2018) lists several examples of NLP tasks that have been used to evaluate the performance of word embeddings. These examples include noun phrase chunking (Collobert et al., 2011; Schnabel et al., 2015; Turian et al., 2010), named entity recognition (Chinchor & Marsh, 1998; Collobert et al., 2011; Sang & De Meulder, 2003; Schnabel et al., 2015), sentiment analysis (Maas et al., 2011; Schnabel et al., 2015; Tsvetkov et al., 2015), shallow syntax parsing (Andreas & Klein, 2014; Bansal et al., 2014; Collobert et al., 2011; Köhn, 2016), semantic role labeling (Collobert et al., 2011; Ettinger et al., 2016; Palmer et al., 2005), negation scope (Ettinger et al., 2016), part of speech tagging (Collobert et al., 2011; Toutanova et al., 2003), text classification (Tsvetkov et al., 2015), metaphor detection (Tsvetkov et al., 2014, 2015), paraphrase detection (Bakarov & Gureenkova, 2018; Baumel et al., 2016; Dolan

& Brockett, 2005), textual entailment detection (Baumel et al., 2016; Bowman et al., 2015; Marelli et al., 2014; Mostafazadeh et al., 2016), and input layer for neural networks (Kocmi & Bojar, 2017).

So-called intrinsic methods of evaluation work by comparing the relationship of word's vector representations with human judgment concerning the relationship of those same words. For example, the word analogy tasks used by Mikolov et al. (2015) belong in the category of intrinsic evaluation methods. Both intrinsic and extrinsic evaluation methods have advantages and disadvantages. The two methods do not correlate with each other.

„Absence of correlation between intrinsic and extrinsic methods. Performance scores of word embeddings, when measured with two existing evaluation approaches (intrinsic and extrinsic), do not correlate between themselves. It is unclear what class of methods is more adequate" (Bakarov, 2018)

Furthermore, a comparative study by Tobias Schnabel et al. has found that different extrinsic evaluation methods do not correlate with each other (2015). Different methods of evaluation have different advantages. In cases where we are using word embeddings as inputs for specific NLP tasks, then evaluating based on the performance of that task is obvious, in other cases weighing the advantages and disadvantages of these methods is important. Considering that different methods for the evaluation of word embeddings have been shown to be uncorrelated, and there is no consensus in the literature regarding the best method it is important to reflect on the specific application of word embeddings.

In the case of my thesis, the word embeddings are not an input, I do not use so-called „off the shelf" models, because I am interested in the semantic relations between words that are specific to my own corpus, therefore I consider intrinsic evaluation to be a good fit for the current analysis. Another consideration is the language of my corpus, which is Hungarian, intrinsic evaluation methods require premade tools that have been either created by researchers or through the use of crowd-sourcing (Liza & Grzes, 2016). Hungarian is generally considered to be a low or mid-resource level language, meaning that NLP resources are not as readily available as in the case of high-resource languages such as English. In light of the above considerations, I used a Hungarian language analogy dictionary created by Makrai Márton

(2015). This dictionary follows the logic set out by Mikolov et al. of validating word embeddings by testing the relations with a set of predefined analogy pairs similar to the previously mentioned example of the vectors of the expressions, man, woman, king, and queen.

This analogy dictionary has a total of 21410 sets of words that belong to 14 different categories: capitals of common countries, capitals across the world, centers of counties, currencies, and 10 different types of grammatical relations across words. In my case due to the use of lemmatization grammatical relations could not be included in the training process, also I had to remove every test where one of the 4 terms was not present in my vocabulary leaving me with 4020 sets of test terms. Examples of the analogy word sets are shown in Table 6.

Table 6.: Examples of the analogy dictionary

Type of analogy	1. term	2. term	3. term	4. term
Capitals of common countries	Budapest	Magyarország	Moszkva	Oroszország
Capitals of common countries	London	Nagy-Britannia	Peking	Kína
Capitals of common countries	Bukarest	Románia	Stockholm	Svédország
Capitals across the world	Accra	Gána	Bécs	Ausztira
Capitals across the world	Ankara	Törökország	Kijev	Ukrajna
Capitals across the world	Katmandu	Nepál	Tehrán	Irán
Currencies	Ukrajna	hrivnya	Törökország	líra

Currencies	Csehország	korona	Svájc	frank
Currencies	Magyarország	forint	Oroszország	Rubel
Centers of counties	Bács-Kiskun	Kecskemét	Baranya	Pécs
Centers of counties	Békés	Békéscsaba	Nógrád	Salgótarján
Centers of counties	Borsod-Abaúj-Zemplén	Miskolc	Csongrád	Szeged

The relationship I am testing for is the same as in the men, women, king, and queen example.

term1 – term2 + term3 = term4*

I consider each test to be successful if the vector of term3 added to the difference of term1's and term2's vector produces such a new vector that has the term4 as the counterpart in the vocabulary. Due to the size of my training corpus and other reasons, I will be later elaborating on I do not expect my embedding models to have a state-of-the-art performance, therefore I have also included a more careful test that does not only look for exact matches with the closest vectors but rather also checks if the target term **term4** is present in the words corresponding to the top 10 closest vectors.

I have tested a total of 12 different word embeddings word2vec and FastText models with both Skip-gram and CBOW architecture, and also GloVe models. I only used context-independent embeddings, since in later parts of my research I will be comparing the vector representations of specific tagged terms, meaning that I will compare these word representations without their specific context. I have tried different window sizes and in the case of my GloVe models, I also tried different numbers for epochs, since GloVe models took considerably less time to fit. In the case of the FastText embeddings, I used n-grams with lengths between 3 and 6. Even from the different parameters listed above it is obvious that the same models could have been reasonably fitted in a nigh-infinite number of different ways,

and there are also other parameters such as the length of the vectors that I did not try to optimize for. In an ideal scenario, I would have used a simple grid-search algorithm to optimize along the multiple dimensions of the available parameters, in this case, I resorted to using the 12 simple models with the parameters described below due to the computational costs of fitting large amounts of word embeddings. I fitted my word2vec and FastText models with the gensim package(Rehurek, 2009) and my GloVe models with the text2vec package (Selivanov et al., 2023). Table 7 shows the specific parameters of each of my 12 models and their performance.

Table 7.: Performances of the different word embeddings

Modell	Architecture	window size	Number of epochs	Vector sizes	Exact prediction rate	Right prediction is within the top 10 vectors
Word2vec	CBOW	5	10	100	1.667%	14.65%
Word2vec	CBOW	8	10	100	2.512%	17.26%
Word2vec	Skip-gram	5	10	100	3.955%	30.697%
Word2vec	Skip-gram	8	10	100	4.50%	32.01%
FastText	CBOW	5	10	100	0.622%	6.69%
FastText	CBOW	8	10	100	1.04%	7.437%
FastText	Skip-gram	5	10	100	4.65%	28.58%
FastText	Skip-gram	8	10	100	5.40%	34.4%

GloVe	-	5	10	100	2.29%	5.75%
GloVe	-	8	10	100	2.29%	5.80%
GloVe	-	5	15	100	1.84%	4.88%
GloVe	-	8	15	100	2.44%	5.42%

Despite my expectations, the FastText models have performed best, and out of those the Skip-gram model with a window of size 8 had the best results. I expected FastText to relatively underperform since the lemmatization has removed much of the morphological information that makes the utilization of subword-level information attractive and useful. It seems that even after lemmatization the vocabulary still contained relevant morphological information. Based on the above results the final model for my analysis was a FastText model with Skip-Gram architecture, a window size of 8, 10 epochs, and the vectors of my model had 100 dimensions.

Another important observation regarding my results is that they are in terms of their performance at these word analogy tasks far behind those published in the original articles. Several reasons could explain this. Firstly, the difference in languages, Hungarian is more morphologically rich than English, of course, has been partially negated to an unknown degree by the use of lemmatization in the preprocessing. The size of the corpus is also relevant, mine contains a little over 186.000 observations, which would be considered large in the context of many other NLP tasks, but all-purpose, „off the shelf" word embedding models are fitted with the help of at least a few millions of documents. Finally, the most important thing in my opinion is the specificity of my corpus.

I have previously narrowed down the collected texts by only selecting articles from the sections relating to national-level news. This has surely caused my embeddings to perform worse, not just because I have less training data this way, but also because important subjects such as international news and economic news have been excluded from the corpus, these

documents would have improved the performance of the models on the analogy questions relating to currencies and world capitals. This highlights the troubles of selecting the right evaluation method for word embeddings. In my specific case, narrowing down the data involved in the analysis simultaneously undermines the models' performance in my evaluation method and also helps the models to better fit the particular research question.

4.3 Dictionaries used in the analysis

The analysis as outlined above makes use of three different dictionaries. The first one is compiled by me and is composed of a set of politically relevant terms. For this set of words, I wanted to compile a list that is both general and complete meaning that they undeniably have political relevance, and the bounds of the set are as clearly defined as possible. Based on these considerations I have chosen to use the names of relevant political parties. I considered a party to be relevant if they throughout the relevant period (2015-2019) had enough support to gain parliamentary representation in an actual or hypothetical parliamentary election. This meant that I excluded three parties despite them having parliamentary representation Együtt (together), Párbeszéd a Zöldek Pártja (Dialogue The Green's Party), and MLP (Hungarian Liberal Party) these parties gained representation through an electoral coalition and did not have the sufficient support on their own. Following the above criteria, I have also included Momentum which did not have parliamentary representation but in a hypothetical election could have gained enough votes to be represented in Parliament.

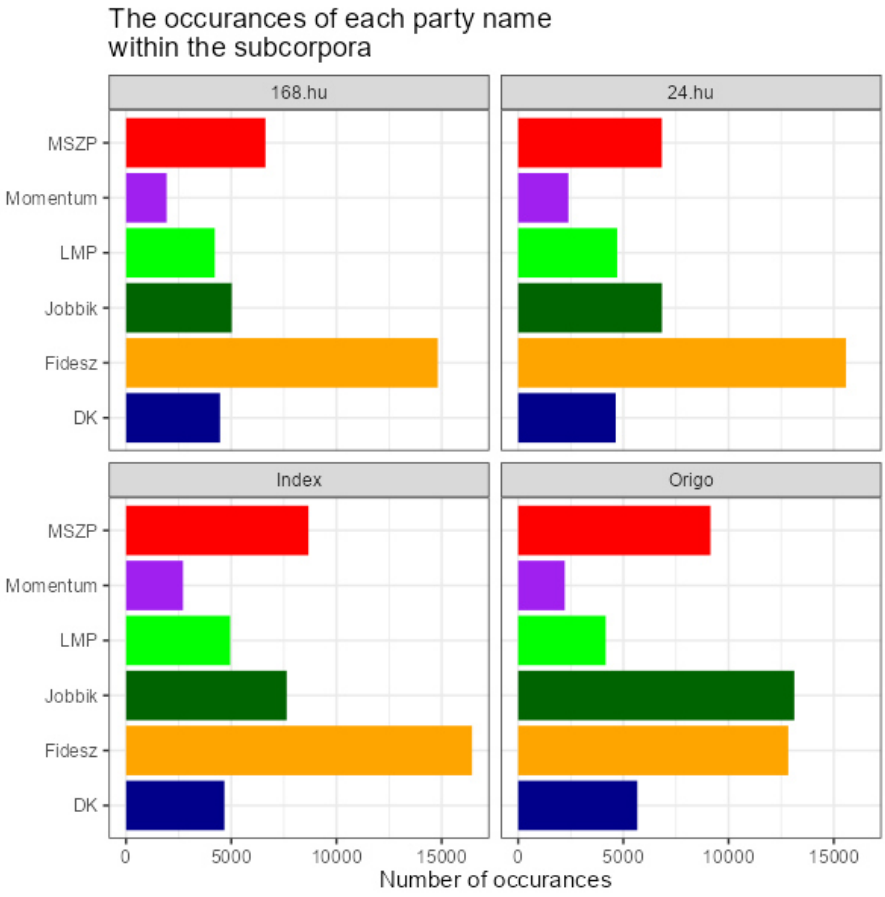
When I tagged the parties' names I tried to account for each common phrasing, Table 8 shows all of the vector representations of parties in the corpus that I accounted for during the tagging process.

Table 8.: The parties and their string representations

Parties	String representations
MSZP	MSZP, Magyar Szocialista Párt
Momentum	Momentum, Momentum Mozgalom, MOMO
LMP	LMP, Lehet Más a Politika
Jobbik	Jobbik, Jobbik Magyarországért Mozgalom
Fidesz	Fidesz, Fidesz-KDNP, Fiatal Demokraták Szövetsége
DK	DK, Demokratikus Koalíció

After tagging, I counted the number of occurrences for each party. The result of the count is shown in Figure 8. The number of occurrences is more or less in line with the given parties' support, with Momentum being a notable underperformer which is explained by the fact that it was only founded in 2017 meaning that it was not present throughout the entirety of the period under investigation. Finally, one more interesting relation that should be noted is that Fidesz has the most occurrences in every subcorpora except for Origo the only government-friendly medium, where Jobbik is the most commonly mentioned party.

Figure 8.: The number of party name occurrences in each subcorpora



The second dictionary that I use in my analysis is a sentiment dictionary, I use the dictionary provided in the HunMiner package. The sentiment dictionary is used to measure the distance between the tagged words and the words in the sentiment dictionary. The third and final dictionary that I use is a set of control words, these words have been randomly selected 20 from all elements of the vocabulary that appear in the corpus at least 6000 times. I decided to only select from the words that have a certain minimum number of occurrences for two reasons, firstly this way I can know that the randomly selected words will appear in each subcorpora, and secondly the party names that I have tagged are also rather common, therefore it would be more accurate to compare them, with similarly common terms.

5. Results

5.1 Relative distance of tagged words

I have calculated the similarity of each possible combination of the tagged terms based on their cosine similarity, a measure that is very commonly used in NLP research. After that, I also calculated the average cosine similarity for each potential combination of mediums. Table 9 shows the cosine similarity for each party and medium combination while Table 10 shows the average cosine similarities for each medium combination.

Table 9.: The cosine similarities for each combination of tagged words and mediums

Party	Medium1	Medium2	Cosine similarity
Jobbik	Index	Origo	0,945769
Jobbik	Origo	Index	0,945769
Jobbik	Index	24.hu	0,932317
Fidesz	Origo	24.hu	0,927918
Fidesz	24.hu	Origo	0,927918
Momentum	Index	24.hu	0,909157
Momentum	168.hu	24.hu	0,909105
LMP	Origo	Index	0,898862
LMP	Index	Origo	0,898862
Jobbik	168.hu	Origo	0,898762
MSZP	Index	Origo	0,896069
Fidesz	Index	24.hu	0,885795
Fidesz	24.hu	168.hu	0,884513
Momentum	Origo	Index	0,884153
Jobbik	168.hu	Index	0,881306
Jobbik	Index	168.hu	0,881306
Momentum	Index	168.hu	0,875658
DK	24.hu	168.hu	0,874245
DK	Index	Origo	0,87357

DK	Origo	Index	0,87357
Momentum	24.hu	Origo	0,86795
MSZP	168.hu	24.hu	0,86621
LMP	168.hu	24.hu	0,865974
LMP	24.hu	168.hu	0,865974
MSZP	Origo	24.hu	0,864923
Momentum	Origo	168.hu	0,853802
Fidesz	168.hu	Index	0,85145
MSZP	24.hu	Index	0,846304
DK	Index	24.hu	0,838002
LMP	24.hu	Index	0,827583
MSZP	168.hu	Origo	0,812437
MSZP	Index	168.hu	0,778787
LMP	168.hu	Origo	0,737693
DK	168.hu	Origo	0,736123
DK	Origo	168.hu	0,736123
DK	168.hu	Index	0,73155

Table 10.: The average cosine similarity of the tagged terms for each combination of mediums

Medium1	Medium2	Average cosine similarity
Index	Origo	0,902
24.hu	Origo	0,897
168.hu	24.hu	0,878
24.hu	Index	0,873
168.hu	Index	0,833
168.hu	Origo	0,796

Following this I have also compared the distribution and descriptive statistics of the cosine similarity values in the case of my tagged party names, and the randomly selected tagged

terms. The distributions are displayed in Figure 9 and the descriptive statistics are shown in Table 11.

Figure 9.: The distribution of cosine similarities for all combinations of party names and all combinations of randomly selected terms

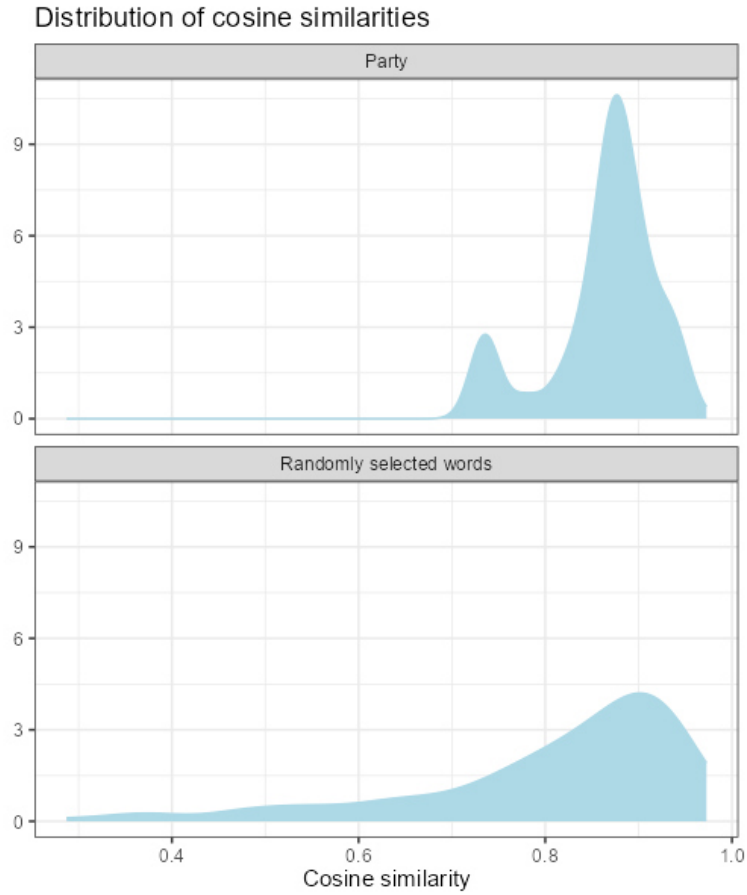
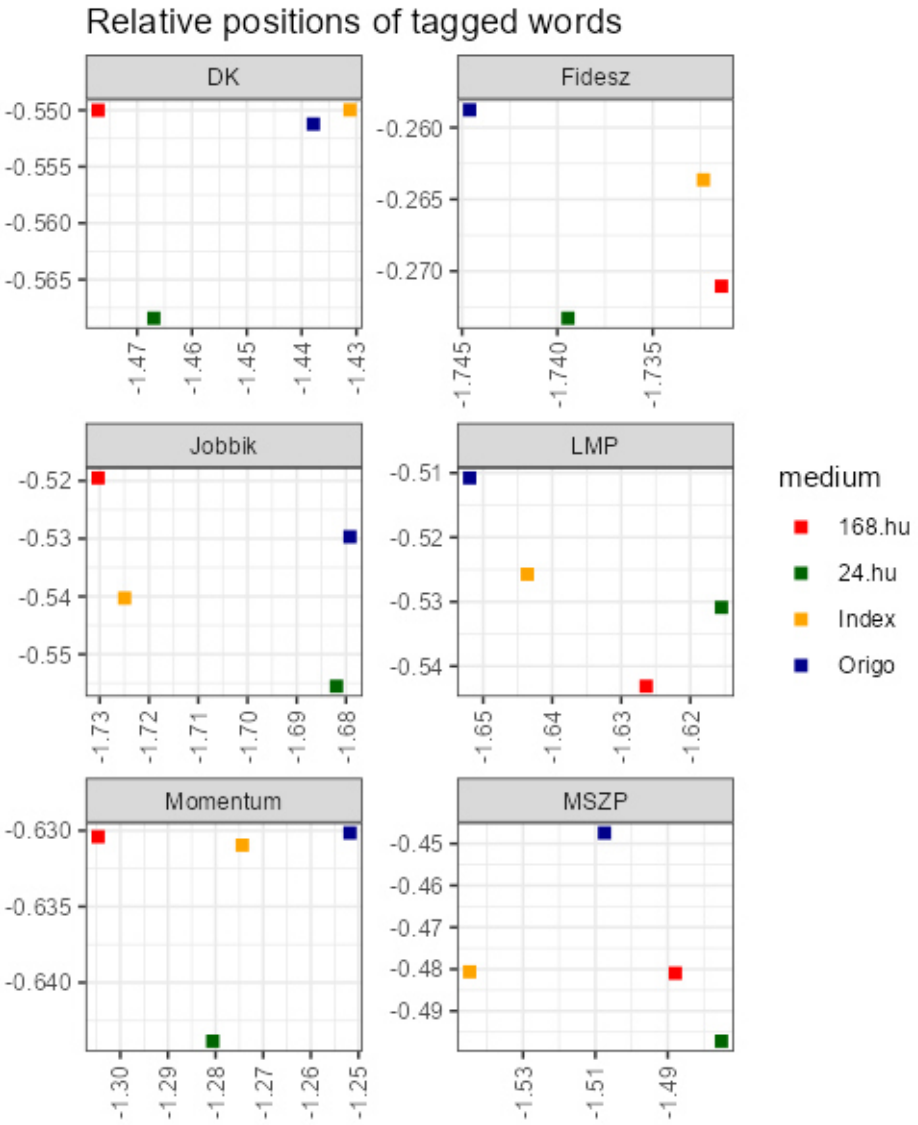


Table 11.: Descriptive statistics of the cosine similarity values in the case of both sets of tagged words

Tagged words	Variance	Mean	Median
Party representations	0,003314	0,863486	0,873907
Randomly selected words	0,021866	0,805635	0,855587

I can also look at the relation of specific tagged terms in different mediums, and not just the overall similarity of the whole set of tagged terms. To visualize these relations, I have used UMAP to reduce the 100 dimensions assigned to my vocabulary to 2. Figure 10. shows the relative positions of each tagged term after I have reduced the dimensions.

Figure 10.: Position of the different tagged mediums in 2-dimensional space after dimension reduction

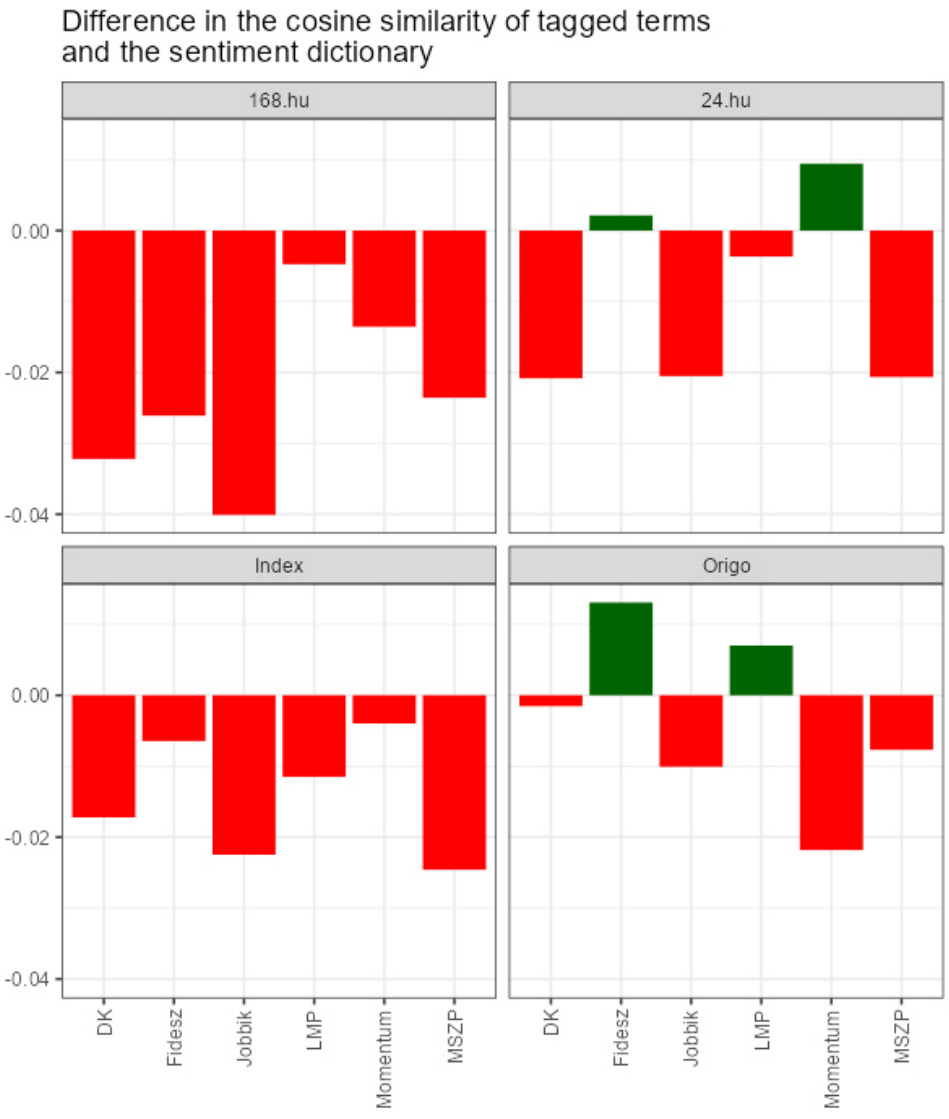


5.2 Distance of tagged words from sentiment dictionary

These tools for visualization can be helpful in some cases, to depict the relative relations of specific words in the vector space of word embedding models, but they do not offer an easily

interpretable output. To gain more insight into the relevant differences between the vector representation of words across the mediums I use a sentiment dictionary similarly to previous research (Rozado & Al-Gharbi, 2022). As I have mentioned I use the sentiment dictionary included in the HunMiner R package (*poltextlab/HunMineR*, 2021/2023) this includes 2588 negative and 2299 positive terms obviously, I could only use the terms that are present in my vocabulary, which leaves me with 2313 negative and 2084 positive terms. I have calculated the cosine similarity of the tagged words and each of these. For each of the tagged terms, the average of their similarity to each negative term and each positive term is separately calculated, their difference is then taken, and the result is shown in Figure 11.

Figure 11.: Differences between parties average cosine similarity to negative and positive terms in the sentiment dictionary.



The Figure shows that in most cases the vector representations of the parties have been closer to the negative terms in our sentiment dictionary, than the positive ones. These results are of course dependent on the vocabulary of the corpus, and the specific dictionary that is used, therefore it is advisable to compare the relative distances across mediums rather than take these results as an absolute measurement. The governing party Fidesz-KDNP had the most positive sentiment associated with their vector representation within the Origo subcorpus, which is in line with my expectations. Simultaneously Momentum is the closest to the negative terms within the Origo subcorpora. Two things that have surprised me about the results of this part of the analysis is that in the case of 24.hu Fidesz is slightly closer to the positive terms, similarly in the case of Origo the vector representation of LMP is on average slightly closer to the positive terms of the dictionary, than the negative ones.

To get a point of comparison I have also calculated the cosine similarities between my control words and the elements of the sentiment dictionary. Figure 12 shows the distribution of these cosine similarities, while Table 12 shows the descriptive statistics of the same cosine similarity values.

Figure 12.: Distribution of cosine similarities between the tagged terms and the elements of the sentiment dictionary

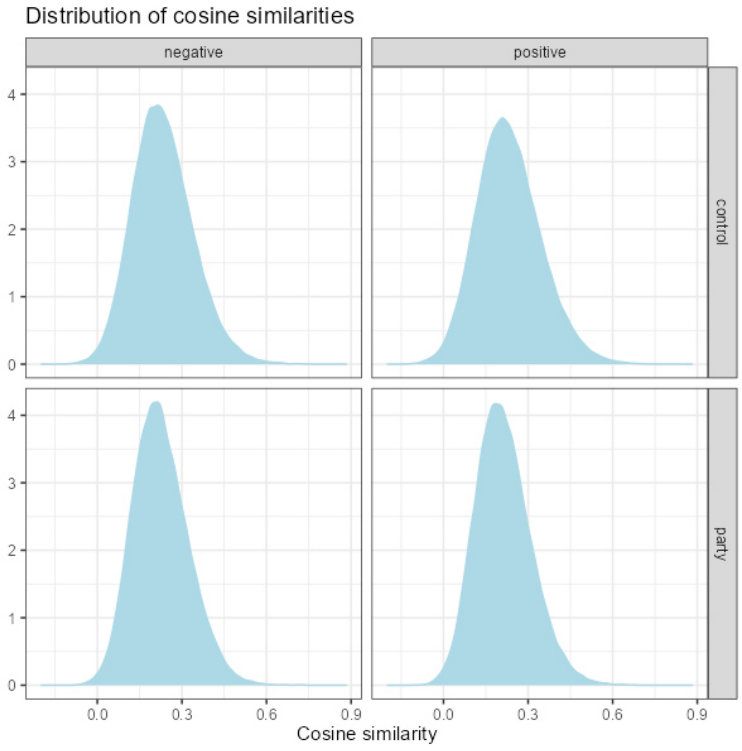


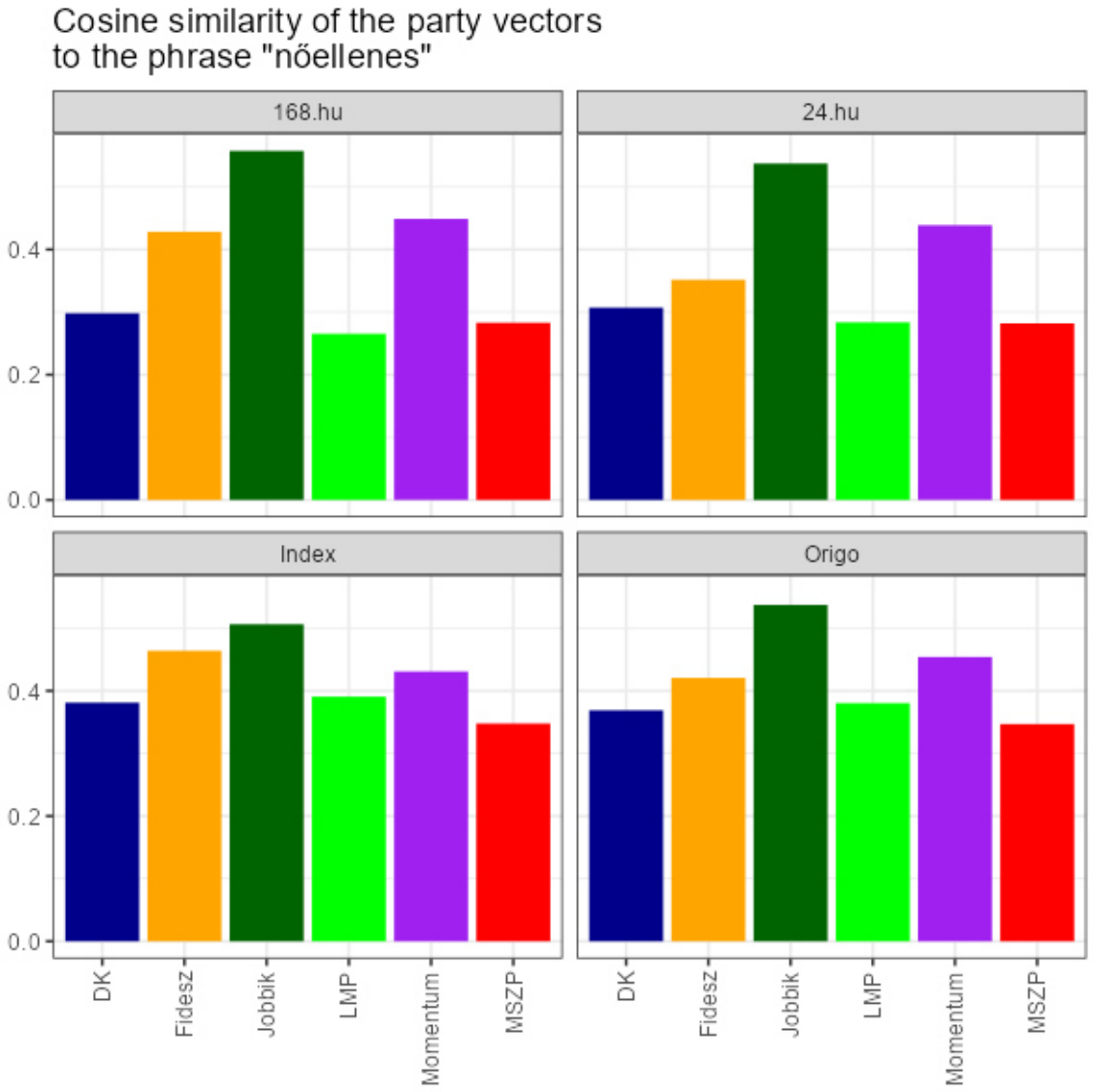
Table 12.: Descriptive statistics of the cosine similarity values between the two types of tagged words and elements of the sentiment dictionary

Tagged words	Sentiment words	Variance	Mean	Median
Randomly selected words	negative	0,012	0,233	0,226
Randomly selected words	positive	0,013	0,233	0,226
Party representations	negative	0,010	0,225	0,218
Party representations	positive	0,010	0,212	0,206

These results show that overall the text representations of parties are at a slightly larger distance from both the negative and positive elements of the sentiment dictionary and also have a bit less variation amongst the cosine similarity values. The smaller amount of variation in the case of the party names is likely because party names more often appear in the same context and therefore have more similar vector representations than randomly selected common words.

Finally, I also wanted to look into a more distinct case of vector relations, where we not only measure the similarity of the party's vector representation and a set of words relating to positive and negative sentiments, but something specific where we could expect a greater deal of variation between the values of cosine similarity. To this end, the phrase "nőellenes" as in anti-women/ misogynist is used.

Figure 13.: Cosine similarity of the party's vector representation to the vector representation of the term "nőellenes"



The results are more or less in line with expectations Jobbik a hard right party is consistently closest to the vector representation of "nőellenes" and Fidesz also performs highly. What is interesting is that Momentum a liberal party also has a high cosine similarity with the term. To interpret this particular result more detailed work would be needed.

6. Conclusion, and limitations of the thesis

This thesis set out to overview the applications of word embedding models to the study of media bias and contribute to this cross-section of the literature through an analysis of my own

on a newly collected corpus of online Hungarian news mediums. My approach to the use of word embeddings differs from previous examples in the literature, in so far that I did not compute an issue-specific subspace, but rather I used tags added to specific strings to compute a subspace alongside the representations of the same concept in the specific subcorpora of my corpus. To my knowledge, this is the first time that this method has been used for the study of media bias, of course, it is not completely novel, but rather a tweak on existing approaches.

My analysis was composed of two separate parts. The comparison of my tagged word's relative distances, and the comparison of the tagged word's distances from the elements of a sentiment dictionary. In the case of the tagged word's relative distances, there were no discernable patterns between the aggregated distances. It may be the case, that with some less frequent terms, where their appearance in the different mediums is more differentiated the relative distances of vector representations would better show the relations of the mediums. Looking at the distances from the sentiment dictionaries terms the results were in line with my prior expectations, although a larger corpus with more mediums would be necessary to better evaluate these relations.

The advantages of word embeddings in the study of media bias have been shown through my analysis. I believe that continuous representations of semantic relations are a natural fit for this research area. Word embeddings are scalable for large corpora since they do not necessitate qualitative work from the researcher. They are flexible because many different semantic relations can be tested for which allows for issue-specific analysis, and also word embeddings are a well-researched NLP method. One important drawback of word embeddings that I have discussed in the analysis section is the lack of an evident and clear-cut method of evaluation.

My thesis is limited in two ways that should be reflected on. Those two limitations are the corpus I use, and the extent of cross-validation. When I was selecting my corpus I aimed to include a diverse set of mediums, that could highlight the advantages and drawbacks of word embeddings in the study of media bias. A study that aims to make notable observations about the extent of bias across mediums and over time in Hungarian media would need to include many more mediums, the corpus of this study is not representative of online Hungarian media and even less so of Hungarian media in general.

The second limitation is the extent of cross-validation within my analysis by this, I not only mean the number of different models and parameters, that I use but also tests for the effect of other elements of the analysis. Such as the sentiment dictionary, testing with multiple dictionaries, and comparing the results was not feasible in my case due to the lack of time. Finally, the whole process could be cross-validated with expert opinion or other methods of measuring media bias, preferably both. I believe that it would be fruitful to build on my results and carry out a similar analysis, with an extended corpus of Hungarian (or other) media, and more detailed cross-validation, such research would be an important contribution to the study of media bias.

References

1. Aday, S. (2010). Chasing the Bad News: An Analysis of 2005 Iraq and Afghanistan War Coverage on NBC and Fox News Channel. *Journal of Communication*, *60*(1), 144–164. https://doi.org/10.1111/j.1460-2466.2009.01472.x
2. Andreas, J., & Klein, D. (2014). How much do word embeddings encode about syntax? *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 822–827. https://aclanthology.org/P14-2133.pdf
3. Ansolabehere, S., Lessem, R., & Snyder Jr, J. M. (2006). The orientation of newspaper endorsements in US elections, 1940-2002. *Quarterly Journal of political science*, *1*(4), 393–404.
4. Bakarov, A. (2018). *A Survey of Word Embeddings Evaluation Methods* (arXiv:1801.09536). arXiv. http://arxiv.org/abs/1801.09536
5. Bakarov, A., & Gureenkova, O. (2018). Automated Detection of Non-Relevant Posts on the Russian Imageboard "2ch": Importance of the Choice of Word Representations. In W. M. P. Van Der Aalst, D. I. Ignatov, M. Khachay, S. O. Kuznetsov, V. Lempitsky, I. A. Lomazova, N. Loukachevitch, A. Napoli, A. Panchenko, P. M. Pardalos, A. V. Savchenko, & S. Wasserman (Szerk.), *Analysis of Images, Social Networks and Texts* (Köt. 10716, o. 16–21). Springer International Publishing. https://doi.org/10.1007/978-3-319-73013-4_2
6. Baker, K. (2005). Singular value decomposition tutorial. *The Ohio State University*, *24*, 22.
7. Balakrishnan, V., & Lloyd-Yemoh, E. (2014). *Stemming and lemmatization: A comparison of retrieval performances*. http://eprints.um.edu.my/13423/
8. Banning, S., & Coleman, R. (2009). Louder than Words: A Content Analysis of Presidential Candidates' Televised Nonverbal Communication. *Visual Communication Quarterly*, *16*(1), 4–17. https://doi.org/10.1080/15551390802620464

9. Bansal, M., Gimpel, K., & Livescu, K. (2014). Tailoring continuous word representations for dependency parsing. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 809–815. https://aclanthology.org/P14-2131.pdf

10. Barrett, A. W., & Barrington, L. W. (2005). Bias in Newspaper Photograph Selection. *Political Research Quarterly*, *58*(4), 609–618. https://doi.org/10.1177/106591290505800408

11. Barrett, A. W., & Peake, J. S. (2007). When the President Comes to Town: Examining Local Newspaper Coverage of Domestic Presidential Travel. *American Politics Research*, *35*(1), 3–31. https://doi.org/10.1177/1532673X06292816

12. Baum, M. A., & Groeling, T. (2009). Shot by the Messenger: Partisan Cues and Public Opinion Regarding National Security and War. *Political Behavior*, *31*(2), 157–186. https://doi.org/10.1007/s11109-008-9074-9

13. Baumel, T., Cohen, R., & Elhadad, M. (2016). Sentence embedding evaluation using pyramid annotation. *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, 145–149. https://aclanthology.org/W16-2526.pdf

14. Bengio, Y., Ducharme, R., & Vincent, P. (2000). A neural probabilistic language model. *Advances in neural information processing systems*, *13*. https://proceedings.neurips.cc/paper_files/paper/2000/hash/728f206c2a01bf572b5940d7d9a8fa4c-Abstract.html

15. Bengio, Y., & LeCun, Y. (2007). *Scaling learning algorithms toward AI*. https://direct.mit.edu/books/edited-volume/chapter-pdf/2286500/9780262255790_can.pdf

16. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993–1022.

17. Bolonyai, F. (2019). *Szöveganalitikai modellek szerzők nemi profilozására*. Eötvös Loránd Tudományegyetem.

18. Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in neural information processing systems*, *29*. https://proceedings.neurips.cc/paper_files/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html

19. Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). *A large annotated corpus for learning natural language inference* (arXiv:1508.05326). arXiv. http://arxiv.org/abs/1508.05326

20. Buda Jakab, M. (2020). *Szövegklasszifikáció rekurrens neurális háló alapú nyelvi modell segítségével*. Eötvös Loránd Tudományegyetem.

21. Butler, D. M., & Schofield, E. (2010). Were Newspapers More Interested in Pro-Obama Letters to the Editor in 2008? Evidence From a Field Experiment. *American Politics Research*, *38*(2), 356–371. https://doi.org/10.1177/1532673X09349912

22. Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

23. Chen, W.-F., Al-Khatib, K., Stein, B., & Wachsmuth, H. (2020). *Detecting Media Bias in News Articles using Gaussian Bias Distributions* (arXiv:2010.10649). arXiv. http://arxiv.org/abs/2010.10649

24. Chinchor, N., & Marsh, E. (1998). Muc-7 information extraction task definition. *Proceeding of the seventh message understanding conference (MUC-7), Appendices*, 359–367. https://catalog.ldc.upenn.edu/docs/LDC2001T02/guidelines.IEtask42.ps

25. 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*, 2493–2537.

26. Covert, T. J. A., & Wasburn, P. C. (2007). Measuring Media Bias: A Content Analysis of Time and Newsweek Coverage of Domestic Social Issues, 1975–2000*. *Social Science Quarterly*, *88*(3), 690–706. https://doi.org/10.1111/j.1540-6237.2007.00478.x

27. Csaba, E. (2023). *Szóbeágyazási vektorterek illesztési problémájának megoldása Prokrusztész transzformációval*. Eötvös Loránd Tudományegyetem.

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

29. Dolan, B., & Brockett, C. (2005). Automatically constructing a corpus of sentential paraphrases. *Third international workshop on paraphrasing (IWP2005)*. https://www.microsoft.com/en-us/research/publication/automatically-constructing-a-corpus-of-sentential-paraphrases/

30. Dumais, S. T. (2004). Latent semantic analysis. *Annual Review of Information Science and Technology (ARIST)*, *38*, 189–230.

31. Durante, R., & Knight, B. (2012). Partisan Control, Media Bias, and Viewer Responses: Evidence from Berlusconi's Italy. *Journal of the European Economic Association*, *10*(3), 451–481. https://doi.org/10.1111/j.1542-4774.2011.01060.x

32. Durrheim, K., Schuld, M., Mafunda, M., & Mazibuko, S. (2023). Using word embeddings to investigate cultural biases. *British Journal of Social Psychology*, *62*(1), 617–629. https://doi.org/10.1111/bjso.12560

33. Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, *7*(2–3), 195–225. https://doi.org/10.1007/BF00114844

34. Elsafoury, F., Wilson, S. R., Katsigiannis, S., & Ramzan, N. (2022). *SOS: Systematic offensive stereotyping bias in word embeddings*. https://durham-repository.worktribe.com/output/1136394

35. Enikolopov, R., & Petrova, M. (2015). Chapter 17 - Media Capture: Empirical Evidence. In S. P. Anderson, J. Waldfogel, & D. Strömberg (Szerk.), *Handbook of Media Economics* (Köt. 1, o. 687–700). North-Holland. https://doi.org/10.1016/B978-0-444-63685-0.00017-6

36. Eshbaugh-Soha, M. (2010). The Tone of Local Presidential News Coverage. *Political Communication*, *27*(2), 121–140. https://doi.org/10.1080/10584600903502623

37. Ethayarajh, K., Duvenaud, D., & Hirst, G. (2019). *Understanding Undesirable Word Embedding Associations* (arXiv:1908.06361). arXiv. http://arxiv.org/abs/1908.06361

38. Ettinger, A., Elgohary, A., & Resnik, P. (2016). Probing for semantic evidence of composition by means of simple classification tasks. *Proceedings of the 1st workshop on evaluating vector-space representations for nlp*, 134–139. https://aclanthology.org/W16-2524.pdf

39. Friedman, S., Schmer-Galunder, S., Chen, A., & Rye, J. (2019). Relating word embedding gender biases to gender gaps: A cross-cultural analysis. *Proceedings of the first workshop on gender bias in natural language processing*, 18–24. https://aclanthology.org/W19-3803/

40. Gans, J. S., & Leigh, A. (2012). How partisan is the press? Multiple measures of media slant. *Economic Record*, *88*(280), 127–147.

41. Gentzkow, M., Petek, N., Shapiro, J. M., & Sinkinson, M. (2015). Do Newspapers Serve the State? Incumbent Party Influence on the US Press, 1869–1928. *Journal of the European Economic Association*, *13*(1), 29–61. https://doi.org/10.1111/jeea.12119

42. Gentzkow, M., & Shapiro, J. M. (2010). What Drives Media Slant? Evidence From U.S. Daily Newspapers. *Econometrica*, *78*(1), 35–71. https://doi.org/10.3982/ECTA7195

43. Gentzkow, M., Shapiro, J. M., & Stone, D. F. (2015). Chapter 14 - Media Bias in the Marketplace: Theory. In S. P. Anderson, J. Waldfogel, & D. Strömberg (Szerk.), *Handbook of Media Economics* (Köt. 1, o. 623–645). North-Holland. https://doi.org/10.1016/B978-0-444-63685-0.00014-0

44. Gordon, J., Babaeianjelodar, M., & Matthews, J. (2020). Studying Political Bias via Word Embeddings. *Companion Proceedings of the Web Conference 2020*, 760–764. https://doi.org/10.1145/3366424.3383560

45. Grabe, M. E., & Bucy, E. P. (2009). *Image Bite Politics: News and the Visual Framing of Elections*. Oxford University Press, USA.

46. Groeling, T. (2008). Who's the Fairest of them All? An Empirical Test for Partisan Bias on ABC, CBS, NBC, and Fox News. *Presidential Studies Quarterly*, *38*(4), 631–657. https://doi.org/10.1111/j.1741-5705.2008.02668.x

47. Groeling, T. (2013). Media bias by the numbers: Challenges and opportunities in the empirical study of partisan news. *Annual Review of Political Science*, *16*, 129–151.

48. Groseclose, T., & Milyo, J. (2005). A Measure of Media Bias*. *The Quarterly Journal of Economics*, *120*(4), 1191–1237. https://doi.org/10.1162/003355305775097542

49. Hamborg, F. (2020). Media bias, the social sciences, and nlp: Automating frame analyses to identify bias by word choice and labeling. *Proceedings of the 58th annual meeting of the association for computational linguistics: student research workshop*, 79–87. https://aclanthology.org/2020.acl-srw.12/

50. Hamborg, F., Donnay, K., & Gipp, B. (2019). Automated identification of media bias in news articles: An interdisciplinary literature review. *International Journal on Digital Libraries*, *20*(4), 391–415. https://doi.org/10.1007/s00799-018-0261-y

51. Hehman, E., Graber, E. C., Hoffman, L. H., & Gaertner, S. L. (2012). Warmth and competence: A content analysis of photographs depicting American presidents. *Psychology of Popular Media Culture*, *1*(1), 46–52. https://doi.org/10.1037/a0026513

52. Ho, D. E., & Quinn, K. M. (2007). *Assessing political positions of media*.

53. Ho, D., & Quinn, K. (2008). Measuring Explicit Political Positions of Media. *Quarterly Journal of Political Science*, *3*. https://doi.org/10.1561/100.00008048

54. Jastrzebski, S., Leśniak, D., & Czarnecki, W. M. (2017). *How to evaluate word embeddings? On importance of data efficiency and simple supervised tasks* (arXiv:1702.02170). arXiv. http://arxiv.org/abs/1702.02170

55. KEPPLINGER, H. M. (1982). VISUAL BIASES IN TELEVISION CAMPAIGN COVERAGE. *Communication Research*, *9*(3), 432–446. https://doi.org/10.1177/009365082009003005

56. Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: State of the art, current trends and challenges. *Multimedia Tools and Applications*, *82*(3), 3713–3744. https://doi.org/10.1007/s11042-022-13428-4

57. Kim, M. Y., & Johnson, K. M. (2022). CLoSE: Contrastive Learning of Subframe Embeddings for Political Bias Classification of News Media. In N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, & S.-H. Na (Szerk.), *Proceedings of the 29th International Conference on Computational Linguistics* (o. 2780–2793). International Committee on Computational Linguistics. https://aclanthology.org/2022.coling-1.245

58. Kocmi, T., & Bojar, O. (2017). *An Exploration of Word Embedding Initialization in Deep-Learning Tasks* (arXiv:1711.09160). arXiv. http://arxiv.org/abs/1711.09160

59. Köhn, A. (2016). Evaluating embeddings using syntax-based classification tasks as a proxy for parser performance. *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, 67–71. https://aclanthology.org/W16-2512.pdf

60. Liza, F. F., & Grzes, M. (2016). *An improved crowdsourcing based evaluation technique for word embedding methods*. http://kar.kent.ac.uk/57326/

61. Llorens, M. (2018). Text analytics techniques in the digital world: Word embeddings and bias. *Irish Communication Review*, *16*(1), 6.

62. Lowry, D. T. (2008). Network Tv News Framing of Good Vs. Bad Economic News under Democrat and Republican Presidents: A Lexical Analysis of Political Bias. *Journalism & Mass Communication Quarterly*, *85*(3), 483–498. https://doi.org/10.1177/107769900808500301

63. Maas, A., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning word vectors for sentiment analysis. *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, 142–150. https://aclanthology.org/P11-1015.pdf

64. Makrai, M. (2015). Comparison of distributed language models on medium-resourced languages. *XI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2015)*, 22–33.

65. Marelli, M., Bentivogli, L., Baroni, M., Bernardi, R., Menini, S., & Zamparelli, R. (2014). Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, 1–8. https://aclanthology.org/S14-2001.pdf

66. McCombs, M., & Valenzuela, S. (2020). *Setting the agenda: Mass media and public opinion*. John Wiley & Sons. https://books.google.com/books?hl=en&lr=&id=E-UJEAAAQBAJ&oi=fnd&pg=PT3&dq=media+agenda+setting&ots=Sr1foMlLZ8&sig=5zsmmnJZW4NWO0Tbic7rVhZmXvg

67. McInnes, L., Healy, J., & Melville, J. (2020). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction* (arXiv:1802.03426). arXiv. http://arxiv.org/abs/1802.03426

68. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space* (arXiv:1301.3781). arXiv. http://arxiv.org/abs/1301.3781

69. Mikolov, T., Deoras, A., Povey, D., Burget, L., & Černockỳ, J. (2011). Strategies for training large scale neural network language models. *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*, 196–201. https://ieeexplore.ieee.org/abstract/document/6163930/

70. Mikolov, T., Yih, W., & Zweig, G. (2013). Linguistic Regularities in Continuous Space Word Representations. In L. Vanderwende, H. Daumé III, & K. Kirchhoff (Szerk.), *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (o. 746–751). Association for Computational Linguistics. https://aclanthology.org/N13-1090

71. Mnih, A., & Hinton, G. E. (2008). A scalable hierarchical distributed language model. *Advances in neural information processing systems*, *21*. https://proceedings.neurips.cc/paper_files/paper/2008/file/1e056d2b0ebd5c878c550da6ac5d3724-Paper.pdf

72. Mohammed, A. A., & Umaashankar, V. (2018). Effectiveness of hierarchical softmax in large scale classification tasks. *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 1090–1094. https://ieeexplore.ieee.org/abstract/document/8554637/

73. Moriarty, S. E., & Garramone, G. M. (1986). A Study of Newsmagazine Photographs of the 1984 Presidential Campaign. *Journalism Quarterly*, *63*(4), 728–734. https://doi.org/10.1177/107769908606300408

74. Morin, F., & Bengio, Y. (2005). Hierarchical probabilistic neural network language model. *International workshop on artificial intelligence and statistics*, 246–252. http://proceedings.mlr.press/r5/morin05a/morin05a.pdf

75. Mostafazadeh, N., Vanderwende, L., Yih, W., Kohli, P., & Allen, J. (2016). Story cloze evaluator: Vector space representation evaluation by predicting what happens next. *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, 24–29. https://aclanthology.org/W16-2505.pdf

76. Niven, D. (2003). Objective Evidence on Media Bias: Newspaper Coverage of Congressional Party Switchers. *Journalism & Mass Communication Quarterly*, *80*(2), 311–326. https://doi.org/10.1177/107769900308000206

77. Orosz, G., Szabó, G., Berkecz, P., Szántó, Z., & Farkas, R. (2023). Advancing Hungarian Text Processing with HuSpaCy: Efficient and Accurate NLP Pipelines. In *Text, Speech, and Dialogue: TSD 2023. Lecture Notes in Computer Science* (Köt. 14102, o. 58–69) [Python]. https://doi.org/10.1007/978-3-031-40498-6_6 (Original work published 2017)

78. Palmer, M., Gildea, D., & Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, *31*(1), 71–106.

79. Papakyriakopoulos, O., Hegelich, S., Serrano, J. C. M., & Marco, F. (2020). Bias in word embeddings. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 446–457. https://doi.org/10.1145/3351095.3372843

80. Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543. https://aclanthology.org/D14-1162.pdf

81. Péter M. (2014, augusztus 22). Politikai okokból kellett távoznia az Origo főszerkesztőjének, állítja egykori helyettese. *444*. https://444.hu/2014/08/22/politikai-okokbol-kellett-tavoznia-az-origo-foszerkesztojenek-allitja-egykori-helyettese

82. Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep Contextualized Word Representations. In M. Walker, H. Ji, & A. Stent (Szerk.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (o. 2227–2237). Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1202

83. *Poltextlab/HunMineR*. (2023). [R]. poltextLAB. https://github.com/poltextlab/HunMineR (Original work published 2021)

84. Ponmalai, R., & Kamath, C. (2019). *Self-organizing maps and their applications to data analysis*. Lawrence Livermore National Lab.(LLNL), Livermore, CA (United States). https://www.osti.gov/biblio/1566795

85. Prat, A., & Strömberg, D. (2013). The political economy of mass media. *Advances in economics and econometrics*, *2*, 135.

86. Puglisi, R., & Snyder, J. M., Jr. (2015). The Balanced US Press. *Journal of the European Economic Association*, *13*(2), 240–264. https://doi.org/10.1111/jeea.12101

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

88. Rehurek, R. (é. n.). *gensim: Python framework for fast Vector Space Modelling* (4.3.2) [Python; OS Independent]. Elérés 2024. április 15., forrás https://radimrehurek.com/gensim/

89. Rong, X. (2016). *Word2vec Parameter Learning Explained* (arXiv:1411.2738). arXiv. https://doi.org/10.48550/arXiv.1411.2738

90. Rozado, D., & Al-Gharbi, M. (2022). Using word embeddings to probe sentiment associations of politically loaded terms in news and opinion articles from news media outlets. *Journal of Computational Social Science*, *5*(1), 427–448.

91. Sang, E. F. T. K., & De Meulder, F. (2003). *Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition* (arXiv:cs/0306050). arXiv. http://arxiv.org/abs/cs/0306050

92. Schnabel, T., Labutov, I., Mimno, D., & Joachims, T. (2015). Evaluation methods for unsupervised word embeddings. *Proceedings of the 2015 conference on empirical methods in natural language processing*, 298–307. https://aclanthology.org/D15-1036.pdf

93. Selivanov, D., models), M. B. (Coherence measures for topic, & code), Q. W. (Author of the W. C. (2023). *text2vec: Modern Text Mining Framework for R* (0.6.4) [Software]. https://cran.r-project.org/web/packages/text2vec/index.html

94. Shultziner, D., & Stukalin, Y. (2021). Distorting the News? The Mechanisms of Partisan Media Bias and Its Effects on News Production. *Political Behavior*, *43*(1), 201–222. https://doi.org/10.1007/s11109-019-09551-y

95. Sirisuriya, D. S. (2015). *A comparative study on web scraping*. http://ir.kdu.ac.lk/handle/345/1051

96. Spinde, T., Hamborg, F., & Gipp, B. (2020). Media Bias in German News Articles: A Combined Approach. In I. Koprinska, M. Kamp, A. Appice, C. Loglisci, L. Antonie, A. Zimmermann, R. Guidotti, Ö. Özgöbek, R. P. Ribeiro, R. Gavaldà, J. Gama, L. Adilova, Y. Krishnamurthy, P. M. Ferreira, D. Malerba, I. Medeiros, M. Ceci, G. Manco, E. Masciari, … J. A. Gulla (Szerk.), *ECML PKDD 2020 Workshops* (Köt. 1323, o. 581–590). Springer International Publishing. https://doi.org/10.1007/978-3-030-65965-3_41

97. Szabó, Z. (2020). *Szövegek klaszterezése szóbeágyazás alapján*. Eötvös Loránd Tudományegyetem.

98. Szeidl, A., & Szucs, F. (2021). Media Capture Through Favor Exchange. *Econometrica*, *89*(1), 281–310. https://doi.org/10.3982/ECTA15641

99. Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. *Proceedings of the 2003 human language technology conference of the North American chapter of the association for computational linguistics*, 252–259. https://aclanthology.org/N03-1033.pdf

100. Tsvetkov, Y., Boytsov, L., Gershman, A., Nyberg, E., & Dyer, C. (2014). Metaphor detection with cross-lingual model transfer. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 248–258. https://aclanthology.org/P14-1024.pdf

101. Tsvetkov, Y., Faruqui, M., Ling, W., Lample, G., & Dyer, C. (2015). Evaluation of word vector representations by subspace alignment. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2049–2054. https://aclanthology.org/D15-1243.pdf

102. Turian, J., Ratinov, L., & Bengio, Y. (2010). Word representations: A simple and general method for semi-supervised learning. *Proceedings of the 48th annual meeting of the association for computational linguistics*, 384–394. https://aclanthology.org/P10-1040.pdf

103. Varsányi, M. (2020). *Háztáji a szocialista nagyüzemi mezőgazdaságban—Egy digitális társadalomtudományi perspektíva lehetőségei*. Eötvös Loránd Tudományegyetem.

104. Waldman, P., & Devitt, J. (1998). Newspaper Photographs and the 1996 Presidential Election: The Question of Bias. *Journalism & Mass Communication Quarterly*, *75*(2), 302–311. https://doi.org/10.1177/107769909807500206

105. Wevers, M. (2019). *Using Word Embeddings to Examine Gender Bias in Dutch Newspapers, 1950-1990* (arXiv:1907.08922). arXiv. http://arxiv.org/abs/1907.08922

106. Zrt H. K. (2020, július 24). Több ezren tüntettek Orbán hivatalánál a szabad Index mellett. *hvg.hu*. https://hvg.hu/itthon/20200724_Elkezdodott_az_Index_melletti_tuntetes