Text Mining
Project Paper

Capturing political and ideological biases with word embeddings

Natalie Widmann
s4499972

January 7, 2017

Newspapers provide us with information about the world and have significant influence on our opinions, as well as our behaviour. Despite the journalistic principle of objectivity, the extent at which certain topics are covered in media, as well as the opinions expressed about political situations depend heavily on the newspaper and the country it is published in.
In this project we will investigate whether political biases in news articles can be captured with word embeddings. Training a neural network on a large corpus of text documents creates a vector space that is able to preserve semantic relations between words [7]. These so called `word2vec` models dependent on the word context that is provided by the text documents they are trained on. Therefore, consistent political and ideological bias in newspapers should affect the relations in the `word2vec` model such that compared with a different model these biases can be identified.
Based on the assumption that media in the US and Russia express diverging opinions regarding certain political issues, two models trained on these databases will be analyzed and compared regarding their view on their own and the opposing state.

Introduction

Free and objective media is often described as the key to a well-informed public and a functioning democracy. Gerber et. al. [5] underline the influential position of media by showing that even short exposure to daily newspapers has significant effect on the political and public opinion, as well as the voting behavior of their readers. In 1988, Edward Herman and Noam Chomsky published the *propaganda model* [6] in which they claim that the profit-oriented structure of mass media corporations and the interests of their stakeholders limit free reporting and the coverage of certain topics.
Hence, to form a critically reflected opinion about political issues it is necessary to have knowledge about who is behind a newspaper and what opinions they might want to convey. Detecting such affiliations is not trivial, as intermediary companies often hid direct relation or financial support of media corporation from political parties or influential persons[1].

[1]See for example the Media Ownership Monitor: http://www.mom-rsf.org/

In this project we will focus on the content of newspaper articles which has the advantage that the influence of political affiliations can be directly assessed by analyzing the vector space model which is based on a large collection of news articles.
The basic assumption in this proof of concept is that US and Russian newspapers express diverging opinions specific political topics, especially referring to the moral standards of the other country. We assume that in newspapers from the US, the context in which Russia appears is closer related to negatively associated terms like *threat, war*,etc. while the US themselves appear in a more positive context. The opposite is expected to be true for Russian newspapers.
As word embeddings depend on the word context provided by the articles used, consistent tendencies to describe a state in a negative or respectively positive manner, should have a significant influence on the corresponding `word2vec` model. We hypothesize that it is possible to identify such consistent biases by comparing the Russian and US model.

Related work

An approach to detect unequal reporting, e.g. a gender bias of the personalities covered in Wikipedia, is the assessment of statistical properties like the frequency or predictability of gender associated words (see for example [10]). More insights into the semantic relations of words give neural networks that are trained on a large collection of documents and map the occurring words into a vector space[7]. These `word2vec` models provide the possibility to explore the vector space via cosine similarity which makes it even possible to create analogies like the famous example $king : man :: woman : queen$. By generating such analogies based on the relation $she : he$ Bolukbasi et. al. [3] revealed that the word embeddings trained on the Google News articles show a significant gender bias such that occupations like *homemaker, nurse* and *receptionist* are highly associated with the word *she*, while for example *maestro,philosopher* and *boss* are very similar to *he*. Similarly, Schmidt [9] analyzed the stereotypical language used on platforms to rate teachers and concludes that 'students have a far more elaborate vocabulary to criticize women for being *unprofessorial* than to criticize men'.
Both projects also make an attempt to remove such a gender bias without disturbing other semantic relations. This is of special interest as word embedding are often used in classification or recommendation tasks, as well as other applications of natural language processing. Zliobaite [11] argues that these negative effects of biased machine learning models play an important role when it comes to the recommendation of products, automated screening of job applications, profiling of perpetrators, etc. and expresses the need for an increase in awareness regarding such issues within the machine learning community.
In this project we will make use of such biases in two `word2vec` models in order to identify diverging opinions and ideologies in the real world.

Methods

Create Databases

In order to obtain two comparable datasets, one with English news from Russian media and one with articles from US media, the websites of the following newspapers are crawled on a daily basis using the `newspaper`[2] python module: *Russia Today, Pravda Report, Sputnik, New York Times, Washington Post, Washington Times.*
Duplicates and articles shorter than 100 or longer than 10 000 words are removed to avoid that commercials, link descriptions or other texts are included in the databases. In total 1 6097 US

[2] https://github.com/codelucas/newspaper

articles (19 971 465 words) and 5 335 Russian articles (4 816 360 words) are obtained over a period of 7 weeks, starting at the 10th of November 2016.
As the databases are relatively small to obtain a word2vec model able to capture semantic concepts from different contexts, the Brown corpus is used as an additional source of information. Even though, it was collected already in 1961, with its broad variety of text categories ranging from cultural, political, religious and scientific articles to novels and fictive stories[4], the Brown corpus provides contextual information without interfering with current news.

Preprocessing and Word2Vec Model

For both datasets the following preprocessing steps are implemented based on the `NLTK` module[1]: Stopwords, numbers and special characters are removed. The remaining text is changed to lowercase letters, lemmatized and split into sentences.
To train the word2vec models the `gensim` package[8] is used including the functionality to automatically detect multi-word terms. Both word2vec models are trained such that the words in the data are mapped to a 100 dimensional vector space based on a sliding window of size 5. Words that occur less than 3 times in the corpus are ignored.
After obtaining a Russian and a US word2vec model, each of them is updated with the sentences of the Brown corpus that were preprocessed in the same, above described way. The advantage of this procedure is that the Brown corpus adds contextual information to the existing words without changing the original vocabulary of the models.

Model Evaluation

Before comparing the obtained word2vec models, it is necessary to evaluate them based on explicit knowledge. As the database is relatively small, this evaluation gives insights into how well the model captures semantic relations and whether it is reasonable to use it for the analysis of political and ideological biases.
Mikolov et. al. [7] released a test set[3] to evaluate syntactic and semantic regularities based on analogies which cover different aspects like countries and their corresponding capital or currency, e.g. *Athens:Greece::Helsinki:Finland, Mexico:peso::Brazil:real*, family relations like *son:daughter ::brother:sister*, plural forms *eye:eyes::woman:women*, comparatives, e.g. *bad:worse::big::bigger*,etc. As in our model, the words are lemmatized during preprocessing, plurals and tense forms are not included in the vocabulary and therefore, these categories will be removed from the test set. For both, the Russian and the US model, the results with and without the Brown corpus are compared.

Model Comparison

Cosine similarity and Word clouds

A commonly measure used to determine the similarity of two words w_1 and w_2 is the cosine similarity which is defined as:

$$sim(w_1, w_2) = cos(\Theta) = \frac{\mathbf{v}_{w_1} \cdot \mathbf{v}_{w_2}}{||\mathbf{v}_{w_1}|| \; ||\mathbf{v}_{w_2}||}$$

$$\text{with} \qquad ||\mathbf{v}|| = \sqrt{v_1^2 + v_2^2 + ... + v_n^2}$$

$$\text{and} \qquad \mathbf{u} \cdot \mathbf{v} = u_1 v_1 + u_2 v_2 + ... + u_n v_n$$

with $\mathbf{v}_w$ being the vector representation of word w in the word2vec model. The cosine similarity ranges from -1 to 1, where 1 represents the identity of the two vectors, while -1 shows that they

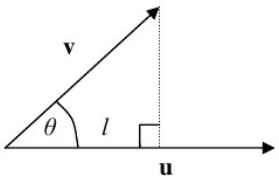
[3]See here for the full list: http://www.fit.vutbr.cz/~imikolov/rnnlm/word-test.v1.txt

have opposite directions. Orthogonal vectors have a cosine similarity of 0 and are often described as unrelated.
For each of the models we will extract the words that have a cosine similarity > 0.75 to the terms *russia* and *unite_state* to get first insights into what the models learned about the two countries. These words are plotted into a word cloud in which the size of the words indicates their similarity.

Keyword projection

To compare the two word2vec models the projection of keywords onto a basis axis is used. Bolukbasi et. al. [2] apply this method to show that gender stereotypes are consistent across word embeddings.
Manually a list of keywords is created such that the words describe actions or conditions of a state and at the same time have a positive or negative association, e.g. *equality, justice, freedom, censorship, war*, etc[4].
To get insights on whether these keywords are seen as positive or negative and how they are associated with the words *unite_state* and *russia* in the each of the word2vec models we will make use of keyword projection. Mathematically speaking, a keyword with its corresponding vector $\mathbf{v}$ is orthogonally projected onto a basis axis $\mathbf{u}$ such that a scalar value l indicates the length of vector $\mathbf{v}$ on the axis $\mathbf{u}$. See the formula and figure below for further illustration:

$$l = ||v|| \cos(\phi) = ||\mathbf{v}|| \frac{\mathbf{u} \cdot \mathbf{v}}{||\mathbf{u}||\ ||\mathbf{v}||}$$



In this project we will use two different basis vectors, one representing the subjective value of a word, so to say showing how positive or negative a word is interpreted, hereinafter called positive-negative axis. The other axis displays a words closeness to the US and Russia, the so called East-West axis.
For each model, the positive-negative axis is obtained by subtracting the vector of the word *bad* from the vector of the word *good*. Similarly, for the East-West axis the *unite_state* vector is subtracted from the vector for the word *russia*. Afterwards, each word in the list of keywords is projected to the model specific positive-negative and East-West axis.
The big advantage of this method is that it allows us to directly compare the results of the two models and put them together in a scatter plot.
With the projection onto the positive-negative axis we can analyze whether the two states agree on the moral value of the keywords (indicated by a positive correlation). If so we would expect a negative correlation representing disagreement when these words are projected onto the East-West axis, as then positively associated words are closer related to the own state while negative words have a higher similarity with the opposing country.

Difference in Cosine Similarity

Another method to compare the two word2vec models is by having a closer look at the differences of the cosine similarities regarding the terms *unite_state* and *russia*. Therefore, for each keyword

[4]The full keyword list contains: *good, wealth, freedom, justice, equality, peace, social, free_speech, democracy, economy, human_right, authority, army, power, crime, dictatorship, repression, propaganda, censorship, right_wing, conspiracy, terrorism, threat, danger, war, bad*

we compute its cosine similarity to *unite_state* and *russia* and then calculate the difference.
If the difference is > 0 the keyword is has a higher similarity with *unite_state*, while if the difference is < 0 the association with *russia* is higher. Close to 0 the keyword can be seen as neutral.
Based on the initial assumption that states tend to speak better about themselves than about opposing governments, we expect that for negatively associated keywords the opposing state has higher similarity scores while positive words have a higher association with the own state.

Results

Model evaluation

The table below shows the model performance regarding syntactic and semantic relations based on the test set provided by Mikoholov et. al. [7]. Both models are evaluated with and without the additional Brown corpus. In both conditions the US model shows generally better results than the Russian model which can be explained by the larger number of news articles used for training. The addition of the Brown corpus increases the accuracy in most of the categories excepting currency.

Accuracy in %	US	US + Brown	Russia	Russia + Brown
capital-common-countries	9.5	11.7	1.5	4.8
capital-world	4.0	5.0	1.8	2.5
currency	0.9	0.0	0.6	0.0
city-in-state	3.3	2.5	0.4	1.3
family	19.9	25.4	4.3	10.5

Even though there is a lot of room for improvement, we consider the models US + Brown and Russia + Brown as sufficient for a further analysis, as they are able to capture the basic relations from different categories. Furthermore, the test set contains many relations between low frequent countries which are not relevant for purpose of capturing ideological biases in newspapers.
For further analysis we will use the both models in combination with the corpus and for convenience refer to them respectively to the Russian model and the US model.

Model Comparison

Cosine Similarity and Word clouds

Figure 1 shows the word clouds of terms that have a cosine similarity greater than 0.75 with *unite_state* (first row) and *russia* (second row). Before looking in more detail at the meaning of the words, it is striking that the number of words falling in this similarity range differs a lot. While in the Russian model (first column) 63 words are similar to *russia* and even 791 words have a cosine similarity higher than 0.75 with *unite_state*, the US model (second column) shows with 39 words for *russia* and only 6 words for *unite_state*, less words in the same cosine similarity range. Also note that in both models the number of closely related words is higher when the keyword is the opposing state than when it is the own country, e.g. the word *russia* in the Russian model has a lower number of similar words than the word *unite_state*. A Fisher's Exact Test indicates that the data does not provide sufficient evidence to conclude that this difference is significant (two-tailed p-value of 0.15).
A closer look at the meaning of the words indicates that in Russian newspapers the United States are mainly associated with negotiations, the European Union, trade treaties, but also words like *sanction* and *undermine* (see Figure 1 upper left). In contrast US newspapers associate the Ukraine, Iran and Syra, as well as *Putin*, the *Kremlin* and*Assad* with Russia. In both model

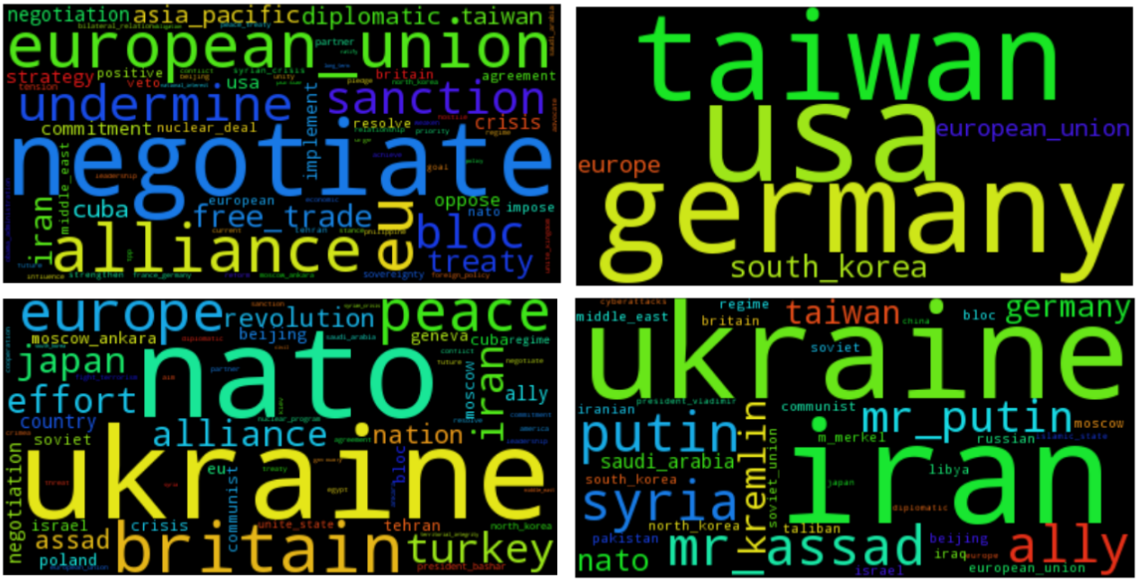


Figure 1: Word clouds showing the terms with a cosine similarity > 0.75 to the words *unite_state* (first row) and *russia* (second row) for the Russian (first column) and US (second column) word2vec model. The mores space a words take up, the more similar it is to the corresponding keyword.

other states play an important role, especially when talking about their own country (see Figure 1 lower left for Russia and upper right for US). Besides, the Russian model also contains terms like *nato, peace, effort, nation, revolution, crisis,...* that are similar to *russia*.
In total, it seems that the US model has a more objective view containing the president, government and partner relations of the opposing country, whereas in the Russian model also terms with a positive or negative association appear, e.g. *peace,crisis,sanction*. But the divergence in the number of words and their meaning make a statistical comparison difficult.

Keyword projection

Figure 2 displays the keyword projections on the positive-negative (left) and the East-West axis (right). Obviously, the words *good* and *bad*, which were also used for the basis axis, span the keyword word range. Besides *war*, which is as expected closely related to the negative end of the basis axis, most words can be found in the center. Also the *authority* stands out as it seems that it is associated more positively in Russian newspapers than it is in US media. In total, the words show a positive correlation of 0.524 (excluding the words *good* and *bad* as they only served as reference points).
On the right side of Figure 2 the keyword projection onto the East-West axis is displayed such that positive values are closer to Russia and negative values closer to the US end of the East-West basis axis. It seems that in the Russian model most of the term are located in the interval $[-1.5, 0]$ while in the US model the keywords are wider spread from -1.5 to 2. Both models seem to agree that *army* and *war* are closer related to *russia* than to the *unite_state*. All in all, the keywords (again excluding *good* and *bad*) show a correlation of 0.258.
As most of the words are centered and show only small variations we will have a closer look at the relation of each model to the own and the opposing country by looking at the difference in cosine similarity for the same keywords.

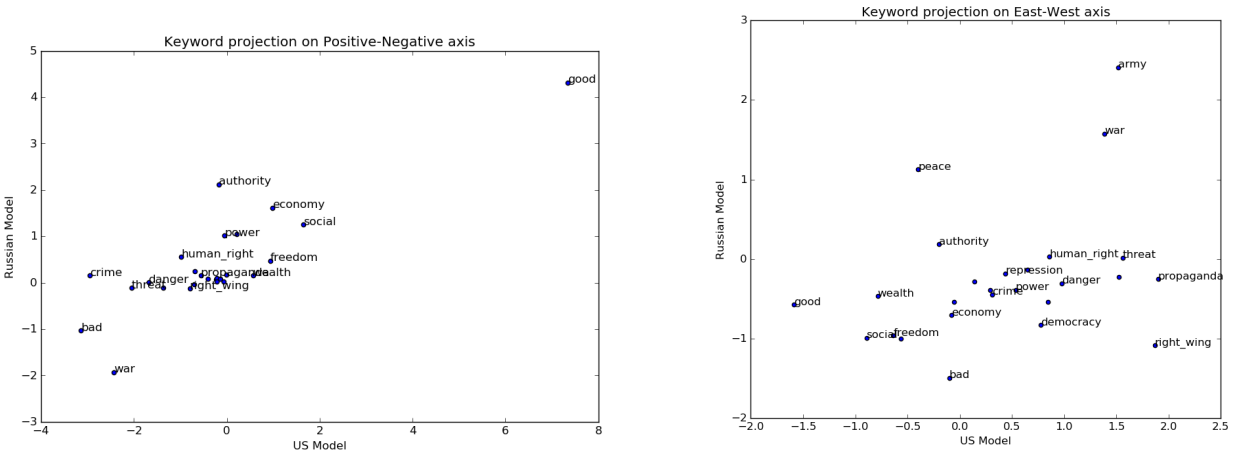


Figure 2: The scatter plots show the keyword projection on the positive-negative axis (left) and on the East-West axis (right). On the left, positive x- or y-values are closer related to the word *good* while negative x- or y-values are more similar to the word *bad*. On the right, positive x- or y-values are closer related to *russia* while negative x- or y-values are more similar to *unite_state*. In both plots, the x-axis corresponds to the projections length of the keywords in the US model, while the y-axis describes the same for the Russian model.

Difference in cosine similarity

The difference in cosine similarity of each keyword regarding the terms *unite_state* and *russia* are displayed in Figure 3. A positive score shows that in the corresponding model this word is closer related to *unite_state* than to *russia*. Respectively, negative scores indicate a higher similarity of the keyword towards the term *russia*.
Similar to the previous result, it can be observed that for the Russian model (blue) most of the keywords have a higher similarity to the US. Exceptions are the words *war* with a value of about -0.03 and *army* which with a value of -0.17 shows the highest deviation. The words with the highest relation to the US are *equality (0.27)* and *right_wing (0.268)*.
In contrast, the US model (red) shows more variations in the resulting scores. It is striking that mainly the first half of the keywords which consists of subjectively negative terms are closer related to Russia, e.g. *right_wing (-0.233), propaganda (-0.230), repression (-0.17)*, while the second half which consist mainly of positive terms are closer related to the US, e.g. *wealth (0.132), good (0.101), freedom (0.095)*). Also note that in the US model 7 keywords show a very neutral score, while in the Russian model only 2 keywords lay in the interval $[-0.04, 0.04]$. Even though, the Russian model does not fit the predicted results, a t-test with a p-value of $1.66 * 10^{-7}$ shows a significant difference between the differences in cosine similarity obtained from the two models.

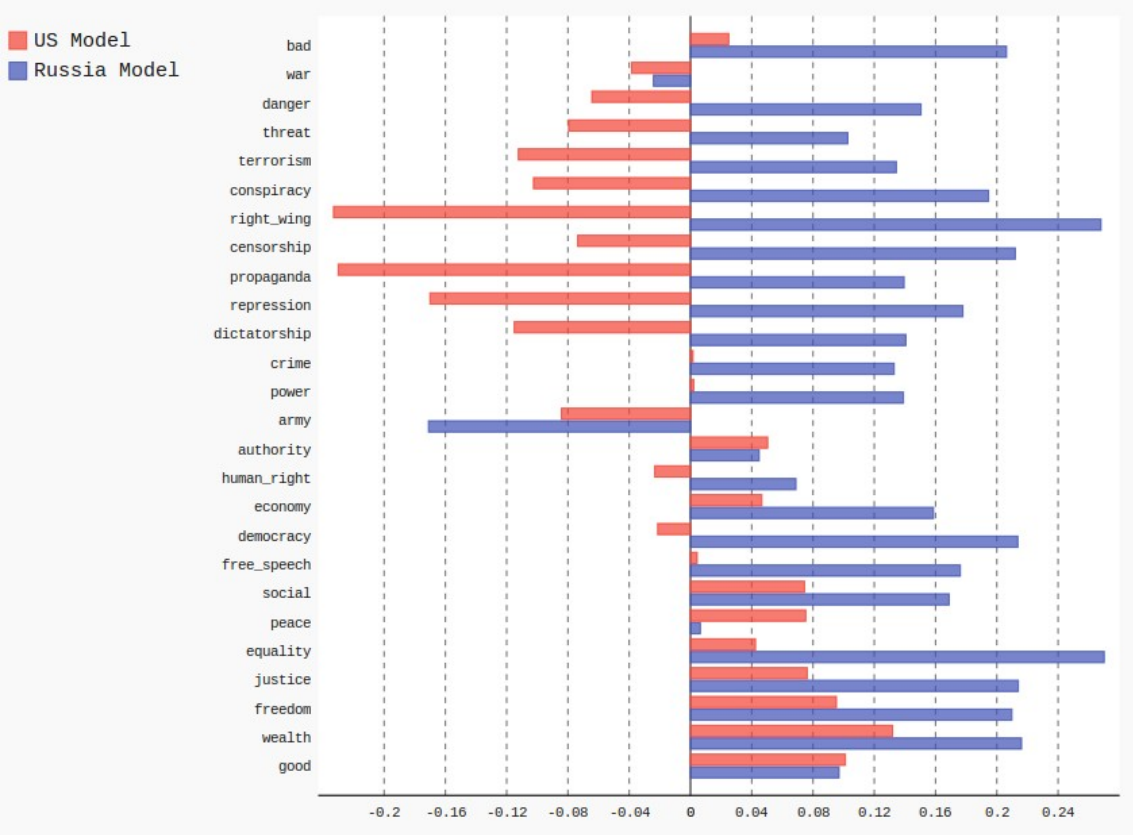


Figure 3: Difference of the cosine similarity of each keyword regarding *unite_state* and *russia* for both, US (red) and Russian (blue) model. A positive score for a word shows that in the corresponding model this word is closer related to *unite_state* than to *russia*. Respectively, negative scores indicate a higher similarity of the keyword towards *russia*.

Discussion

Regarding the evaluation of the word2vec models, we note that the addition of the Brown corpus increases the performance in building correct semantic relations in most of the categories excepting currency, which can be explained with the fact that since 1961, when the Brown dataset was collected, many countries have adopted a new currency, e.g. the euro zone. Nevertheless, the improved accuracy show the positive influence of adding a basis word2vec model. Please note that as the Brown corpus was consistently added to both of the models it can not be accounted for their differences.
In general, the Mikolov test set is only used to get insights into how well the model is able to capture a real world relations. A very high accuracy on the test set is not required, as it includes very detailed information about low frequent countries, which is not relevant for the task of capturing ideological biases in newspapers.

When comparing the most similar words each model to itself or the opposing country we noted a difference in the amount of how many closely related terms were found. A possible explanation is the number and variety of the articles used for training. The more frequent a word appears

in different contexts, the more this word will be associated with different concepts and therefore decrease the cosine similarity to a certain word which only represents one of these concepts. This, would not only explain the differences between the two models, but also the observation that the number of words decreases when the model is asked for the most similar words regarding its own country. Especially, when comparing newspapers from different countries it is obvious that the coverage of the own country is much more diverse, domestic politics, infrastructural and social projects, as well as smaller incidents. Therefore, also the context in which this specific country appears is much more diverse than the context of a foreign country which is often limited to decision that directly influence the own relations and events of world interest. A bigger context provides much more information which makes it more difficult to closely related such a word to another one.
Possible ways to improve the observed results are to either increase the size of at least the Russian database or to filter the articles such that only foreign politics is considered.

The keyword projection onto the positive-negative axis showed the expected positive correlation, which could indicate an agreement of the moral assessment of the keywords. However, with a correlation coefficient of 0.524 this is relation is not very strong and as most of the words are very centered, adding or removing certain keywords could lead to a significant change in the correlation. Therefore, a more in-depth analysis is needed with a larger amount of keywords is required.
Even though, the positive correlation in the keyword projection onto the East-West axis might be influenced by outliers, there seems to be a consistent difference in the US and Russian model which becomes clearer when looking at the differences in cosine similarity: While the US models seems to fulfill the prior assumptions that positive terms will be closer related to a models own country while negative terms will show a higher cosine similarity with the opposing state, the Russian models seems to be almost unaffected from the subjective meaning of a word. Besides the words *army* and *war* all keywords are showing a higher similarity to *unite_state* than to *russia*.
Similar to the above stated explanation, the smaller size of the Russian database could be the reason why this model does not behave as expected.

In general, the interpretation of results from word2vec models is limited as word vectors are computed based on context similarity but do not reveal the context itself. For example a negatively associated word could appear often in a very positive context, e.g. *There is no repression in the US, Russia is fighting terrorism*,etc. Such consistent pairing of negative words in positive context, or vice versa, could lead to the interpretation that a negative or respectively positive word is closely related to a certain country suggesting an ideological biases while the news articles themselves would suggest the opposite. However, such wrong interpretations can be prevented by using multi-word expression such that more context is included in the vocabulary, but also by verifying the word2vec results with an in-depth analysis of the original articles.

Another crucial point that can lead to very different results in keyword projection or the difference in cosine similarity is the words that are chosen as a basis axis. In this project we use the words *good* and *bad*, as well as *unite_state* and *russia*. It is important to note that moral judgment is in most written articles transmitted in an implicit way, e.g. no newspaper will write 'country XY is bad', also negatively associated words like *propaganda* or *censorship* might not appear in a negative context, like *propaganda has a bad influence*, because the terms themselves already have this negative association. To capture such moral relations it is again crucial to have a large enough database to make it possible to build up concept cluster to identify related terms.
Also cultural differences come into play as they affect the usage of vocabulary. Therefore, even though two countries use the same word, it might have different meanings or is used in distinct contexts what makes it difficult to directly compare them.

Conclusion

In this report we have seen how the two word2vec models differ when comparing their similarities of keywords in regard to their own or the opposing state. Despite of results that differ in parts from the prior assumptions and their limitations in interpretation, we were able to show that models show a significant difference in the cosine similarities of the keywords. This is a very important starting point for a further analysis which has to include a larger database for the Russian news articles, as well as a balance or control of the number of articles covering issues of domestic and foreign politics. Furthermore, this research shows that there is plenty of room to investigate how political and ideological biases can be captured and identified with word embeddings.

References

[1] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly Media, 2009.

[2] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Quantifying and reducing stereotypes in word embeddings. *arXiv preprint arXiv:1606.06121*, 2016.

[3] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pages 4349–4357, 2016.

[4] W.N. Francis and H. Kucera. Brown corpus manual - manual of information to accompany a standard corpus of present-day edited american english, for use with digital computers. Departement of Linguistics, Brown University, 1971. http://clu.uni.no/icame/brown/bcm.html.

[5] Alan S Gerber, Dean Karlan, and Daniel Bergan. Does the media matter? a field experiment measuring the effect of newspapers on voting behavior and political opinions. *American Economic Journal: Applied Economics*, 1(2):35–52, 2009.

[6] Edward S Herman and Noam Chomsky. Manufacturing consent: A propaganda model. *Manufacturing Consent*, 1988.

[7] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[8] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en.

[9] B. Schmidt. Rejecting the gender binary: a vector-space operation. Ben's Bookworm Blog, 2015. http://bookworm.benschmidt.org/posts/2015-10-30-rejecting-the-gender-binary.html.

[10] Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. It's a man's wikipedia? assessing gender inequality in an online encyclopedia. *CoRR*, abs/1501.06307, 2015.

[11] Indre Zliobaite. A survey on measuring indirect discrimination in machine learning. *CoRR*, abs/1511.00148, 2015. URL http://arxiv.org/abs/1511.00148.