УДК 004.912

DEVELOPMENT OF A CLASSIFICATION MODEL FOR DETERMINING THE BIAS OF POLITICAL TEXTS IN ENGLISH

Ларионова М.В. (Университет ИТМО)
Научный руководитель – к.пед.н., доцент ФИКТ Валитова Ю.О.
(Университет ИТМО**)**

The aim of the study was to develop a classification algorithm to determine whether a sentence from a political text is biased or neutral. During the study, machine learning methods were used.

Введение. Politics plays an important role in the life of society. In this area, the appeal to emotions is widespread, for example, in pre-election speeches or speeches of party members, since this allows politicians to focus on the most pressing issues. At the same time, political texts are often biased. This can lead to a distorted formation of the position of the listener, which makes the assessment of the bias of political texts relevant.

Основная часть. To conduct the study, a dataset containing labeled text data in English was used. The dataset was based on messages by politicians posted on social networks. Texts preprocessing was carried out, including tokenization, lemmatization and stop words removal, after which the texts were vectorized. For this purpose, Python libraries such as NLTK, spaCy, gensim, re, applied for natural language processing, were used. After that, the messages were classified by means of machine learning methods. The models developed in the study were applied to a new dataset containing Democratic and Republican politicians' speeches in the 2020 US campaign. Moreover, the words with the highest frequency in biased sentences and their subject matter were identified. Thus, text preprocessing and classification models were developed for subsequent prediction of the text class: biased or neutral.

Выводы. During the study, the appropriateness of applying the developed models to new political text data was determined. It was found that biased sentences are less common than neutral ones, and that the most common component of biased sentences is words related to the current political, economic and social situation in the country, to members of the party giving a speech or, on the contrary, to members of the opposing party. This solution can be relevant in an interdisciplinary approach that considers speeches of politicians from the phenomenological and sociocultural point of view.

Ларионова М.В. (автор) Подпись
ms.maria.larionova@gmail.com

Валитова Ю.О. (научный руководитель) Подпись

УДК 004.912

РАЗРАБОТКА КЛАССИФИКАЦИОННОЙ МОДЕЛИ ДЛЯ ОПРЕДЕЛЕНИЯ ПРЕДВЗЯТОСТИ ТЕКСТОВ ПОЛИТИЧЕСКОГО СОДЕРЖАНИЯ НА АНГЛИЙСКОМ ЯЗЫКЕ

Ларионова М.В. (Университет ИТМО)
Научный руководитель – к.пед.н., доцент ФИКТ Валитова Ю.О.
(Университет ИТМО**)**

Целью исследования является разработка алгоритма классификации, позволяющего определить, является ли предложение из текста политического содержания предвзятым или нейтральным. В ходе работы применялись методы на основе машинного обучения.

Введение. Политика занимает важное место в жизни общества. В этой сфере актуальна апелляция к эмоциям, например в предвыборных речах или выступлениях партийных членов, так как это позволяет политическим деятелям акцентировать внимание на наиболее актуальных проблемах. При этом зачастую политические тексты являются предвзятыми. Это может привести к искаженному формированию позиции слушателя, что делает актуальной оценку предвзятости политических текстов.

Основная часть. Для проведения исследования был использован датасет, содержащий размеченные текстовые данные на английском языке, которые были сформированы из сообщений политиков, содержащихся в социальных сетях. Была осуществлена предобработка текстов, включающая токенизацию, лемматизацию и удаление стоп-слов, после чего была проведена их векторизация. Для этого использовались такие библиотеки языка Python, как NLTK, spaCy, gensim, re, применяемые для обработки естественного языка. После этого при помощи методов машинного обучения была проведена классификация сообщений. Модели, разработанные в ходе исследования, были применены к новому набору данных, содержащему выступления политиков от партий демократов и республиканцев в предвыборной кампании США 2020 года. Также были выявлены слова, наиболее часто встречающиеся в предвзятых предложениях, и их тематика. Таким образом, были разработаны модели предобработки текста и классификации для последующего предсказания класса текста: предвзятый или нейтральный.

Выводы. В ходе исследования была определена целесообразность применения разработанных моделей к новым текстовым данным политического характера. Было выявлено, что предвзятые предложения встречаются реже, чем нейтральные и что наиболее частой составляющей предвзятых предложений являются слова, связанные с актуальной политической, экономической и социальной ситуацией в стране, с членами выступающей партии или, напротив, с членами противоборствующей партии. Данное решение может быть актуально в междисциплинарном подходе, рассматривающем выступления политиков с феноменологической и социокультурной точек зрения.

Ларионова М.В. (автор) Подпись
ms.maria.larionova@gmail.com

Валитова Ю.О. (научный руководитель) Подпись