# Linguistic Bias in Crowdsourced Articles and the Aspect of Time

Evangelos Mathioudis
University of Nicosia
Nicosia, Cyprus
mathioudis.e@live.unic.ac.cy

## ABSTRACT

Wikipedia is considered as a reliable and accurate source of information by many users of the Web. A lot of people will go through Wikipedia to learn about something they are interested in. In addition, scientists possibly use Wikipedia as their starting point for their research by searching on the references for the topic they are interested in. Wikipedia's content is edited and maintained by volunteers. With that said, anyone can add or edit the content and through this express personal or political beliefs. Thus, it is important for the content to be unbiased from personal beliefs. In this work, we are going through political biographies and we are presenting a new approach to detect how linguistic biases have changed over time. Specifically we compare the level of subjectivity of the content in biographies about males and females for 7 European politicians. We compare them by using a custom metric the *Mean Abstract Level*. Based on Linguistic Category Model there are four levels of abstraction when we describe a person. The *Mean Abstract Level* measures the mean level of abstraction in a text.

## CCS CONCEPTS

• **Information systems** → *World Wide Web*; • **Applied computing** → *Law, social and behavioral sciences.*

## KEYWORDS

natural language process, wikipedia, sentiment analysis, subjectivity, linguistic bias

**ACM Reference Format:**
Evangelos Mathioudis. . Linguistic Bias in Crowdsourced Articles and the Aspect of Time. In . ACM, New York, NY, USA, 4 pages.

## 1 INTRODUCTION

Wikipedia is a source of information for many users on the web. Additionally, a large number of researchers of various domains are using Wikipedia as a starting point for their research. But, Wikipedia is an open-collaboration encyclopedia and everyone is free to add content and express their personal opinion on the context that is described. For this reason, Wikipedia's article content needs to be unbiased from personal statements and experiences.

Personal statements and experiences are expressed through linguistic biases. *Linguistic biases* come in two different ways. The *Linguistic Expectancy Bias* or LEB is referred to how we communicate information related to a person or an event based on what our expectations of these are. We tend to use more abstract language when we talk about something consistent with our expectations [4, 5]. The *Linguistic Intergroup Bias* or *LIB* describes how we express ourselves regarding actions and atributes of people who belong to our social group. Our expectations suggest that in-group members act positively, thus we are going to talk about the positive actions and characteristics more descriptively. On the other hand, we tend to describe a negative behavior with less information and more concretely [5, 8].

### 1.1 Motivation

Our work is based on previous work of Otterbacher et al.. We have focused our research on how linguistic biases on biographies about politicians have changed over time. By using the Wikipedia API MediaWiki we downloaded the histories of revisions for several biographies about politicians. Additionally, we have used the tool *Didaxto*[1] to extract domain-specific dictionaries and we have analyzed the biographies concerning the number of verbs, adverbs, and adjectives they contain. Didaxto implements an unsupervised learning approach to detect patterns and extract domain specific dictionaries.

## 2 RELATED WORK

Due to its collaboratively structure, Wikipedia has attracted the interest of scientists and researchers from different domains. Wikipedia's content has been the object of research related to bias from different perspectives.

### 2.1 Gender-Race Bias

Worku et al. investigates if content that is preferred by females has greater chances of deletion due to the imbalanced number of male and female contributors. A different approach is presented in [2]. This work examines the inequality of geographical coverage on biographies of notable individuals. The results reveal that it is clear that Wikipedia contains more articles related to the Global North and the Traditional Western powers [2].

### 2.2 Linguistic Bias

Otterbacher et al. focuses on the linguistic bias on biographies of notable individuals concerning gender and ethnicity on three different locations, the whole content, the first paragraph, and on the knowledge panel on the search engine results page. Besides that, it compares the result across the English and Greek Wikipedias. The conclusion indicates that there are linguistic biases both across Wikipedias and in each separately.

### 2.3 Prejudice

In this work [7], the idea is to correctly assign political labels to news publishers in Wikipedia. The reason is that most of the users when searching on the Web will take a quick look at the knowledge panels of the search engine result pages to get informed.

*Univeristy of Nicosia, MSc in Data Science, Project in Data Science*
Spring, 2021.

[1] Didaxto http://deixto.com/didaxto/

Wikipedia's content is used on third-party information providers, thus it is important to have been verified before.

# 3 METHODOLOGY

## 3.1 Overview

For many people Wikipedia is the main source of reliable information, providing a significantly large number of subjects. One of the most popular categories is the biographies of notable individuals. The idea for this work is to specifically research the linguistic biases that occur in politicians' biographies and their effect on the subjectivity of each biography. Based on the *Linguistic Category Model* [6] there are four predicates categories that our words might fell in. These categories are depicted in Figure 1

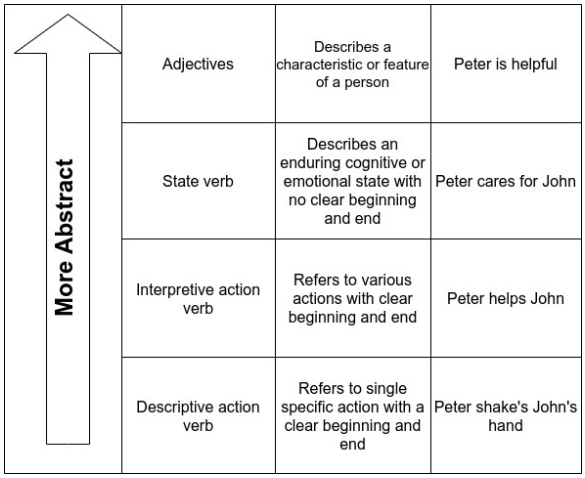


**Figure 1: The Four Predicates based on LCM**

Moving from bottom to top, we describe a person more concretely when we use more descriptive action verbs. On the other hand, we describe a person in a more abstract way if we use more adjectives.

## 3.2 Dataset Description and Analysis

*3.2.1 Dataset.* We have built our dataset by using Wikipedia's API MediaWiki. MediaWiki gives you access to a lot of features and one of them is the history of edits for each biography. The form of the data that Wikipedia returns when someone requests a specific page is on *wikitext*. We parsed the edits using Python's library MediaWiki Parser From Hell[2]. We chose to use history of edits for 7 politicians from Europe. For each edit we have kept track of the **revisionId**, **timestamp**, **userId**, **content** and **tags**.

We will call each edit a **snapshot** of the biography for a specific time $t_i$. A very simplistic way to draw someone the history of edits should look like is the Figure 2 below. It is worth noticing that $snapshot_i$ is not necessarily different from $snapshot_{i+1}$. As have been stated in [3] there are some special kind of edits called *Reverts*. *Revert* is the case when the $user_n$ does some edits at the time $t_n$ and afterwards another user let's say $user_m$ reverts the biography to the snapshot of time $t_{n-1}$ or even prior i.e $t_{n-k}$. We aren't going

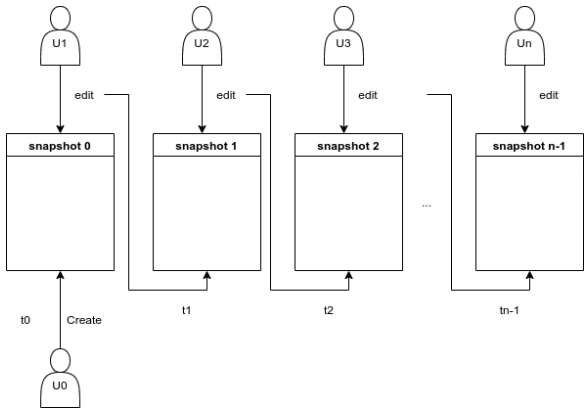


**Figure 2: Snapshots of biography.**

deeper into this, but this is a phenomenon called **Edit Wars** [3] and we have observed a lot of instances of these biographies. So, for each **snapshot** we have got some metadata and the content. There were cases that the content was *hidden* and these cases we chose to drop but it could be a future research question.

*3.2.2 Exploratory Data Analysis.* We fed the contents of the biographies to Didaxto and we have extracted domain-specific dictionaries with positives and negatives words. After that we plugged each of these words in the module *nltk.pos_tag()*[3] and have kept the part of speech that represents. To create a metric of subjectivity we wanted to identify the four predicators that we came up with to the *Linguistic Category Model* which are Descriptive Action Verbs (DAV), Interpretative Action Verbs (IAV), State Verbs (SV), and Adjectives (ADJ). We have inspired our formula from the *mean abstract level* formula from [9]. Unfortunately, the *NLTK POS-tagger* is not able to characterize the words this way, thus we picked everything related to verbs, adverbs, and adjectives.

$$MAL = \frac{2 * numOfVBs + 4 * (numOfAVBs + numOfADJs)}{numOfVBs + numOfAVBs + numOfADJs} \quad (1)$$

As can be seen in Eq. (1) we gave a weight of 2 on the verbs and a weight of 4 on adverbs and adjectives. And then we have divided by the sum of all occurrences. A $MAL = 2$ means that there are no evidence of linguistic biases and a $MAL = 4$ means that the content is biased.

In addition, for each **snapshot** we have calculated the relative frequencies of positive and negative words. Also, we have extracted the relative frequencies for adjectives, verbs and adverbs. Furthermore, we have created a continuous variable named *mean_abstract_level* and a feature called *gender* to be able to do aggregations based on gender and research whether there are linguistic biases related to the gender of politicians.

We have noticed that some days were very "busy" and there was a lot of "action" happening in very short intervals. For that reason, we resample our dataset and have kept the last measure for each of the above measurements. We have concluded to a dataset that consists of 17721 rows, each one represents a snapshot of a specific

[2]mwparserfromhell https://pypi.org/project/mwparserfromhell/

[3]Nltk POS-tagger https://www.nltk.org/book/ch05.html

biography[4]. Table 1 compares the average values of the *length of biography* and the *mal score* for males and females politicians. We observe slightly different *mean_abstract_level scores* which we

**Table 1**

| gender | length | mean_abstract_level |
|---|---|---|
| f | 4656.6 | 2.09278 |
| m | 8904.95 | 2.08482 |

would like to investigate if it could be by chance or not. Thus, we applied a hypothesis test. Our hypothesis are:

> *Null Hypothesis $H_0$: The average of female MAL score is equal to average of male MAL*
> *Alternative Hypothesis $H_1$: The average of female MAL score is higher than the average of male MAL*

Because our entries represent an event that happened to a specific time $t_n$, we assume that each one of these entries is a single article in Wikipedia. This assumption is helping us to be able to apply *T-test* on data that represent time series data.

**Table 2: Results**

| | T-statistic | p-Value |
|---|---|---|
| 0 | 4.06077 | 2.45858e-05 |

Table 2 shows the results of our hypothesis test. For a confidence level of 95% we get a $p \asymp 0.00002$. We conclude that there are evidences to reject the null hypothesis.

*3.2.3 Connections to Real Events.* It is interesting to see how do politicians' action affect the world's opinion and if this can be identified on social media or in Wikipedia's articles. As have been observed in [1] there is a lot of attention on political biographies during election times related to signs of engagement on other social media.

In Figure 3 we plot the normalized values of the *mean abstract level*, and the relative frequencies of positive and negative words of *Recep Tayyip Erdoğan*'s[5] biography. We are going to give some

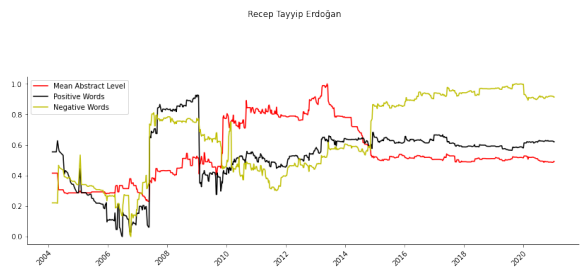


**Figure 3: Erdogan's Weekly MAL Scores, positive and negative words concetrations**

highlights of this plot:

(1) There are picks for positive and negative concentrations after the elections on 2007[6]
(2) In the summer of 2008, the Turkish prime minister visited Baghdad after some period of tense between two countries[7].
(3) A big drop of positives words at the beginning of 2009 is followed by an increase in the MAL score. These changes could be related to Erdogan's meeting with the Israeli president on Davos[8].
(4) After 2009, the mean abstract level is increasing for about 4 years. The positive and negative words frequencies at the same period follow this increase but at a smaller rate. There is a pick for *mal score* which could correlate with an award that the Turkish prime minister received[9].
(5) In the summer of 2013 CNN posted an article[10] with the title "Turkey's Erdogan: Successful leader or 'dictator'?".


**Figure 4: Merkel's Weekly MAL Scores, positive and negative words concetrations**

On the other hand, in Figure 4 we have plotted the weekly changes in *Mean Abstract Level* and the changes in concentration of positive and negative words for Merkel's biography. We highlight the following:

(1) It is clear that Merkel as a woman is being described more abstractly than Erdogan.
(2) The subjectivity remains stable from the beginning.
(3) Positive and negative words have almost the same increase for the first one and a half years.
(4) In May 2005, Merkel won the national elections.
(5) It is almost after the national elections of 2005 until now that positive words dominate negative words in the biography.

/It is important here to clarify that all the above haven't verified quantitatively but they are the results of the author's research. The author picked these two politicians because they had the largest numbers of revisions./

## 4 DISCUSSION

We would like to focus a bit more to figure 3. The *mal scores* for all biographies presented several up and downs. But in each

[4]All the code will be uploaded on https://git.io/JOc7Z
[5]Recep Tayyip Erdoğan https://en.wikipedia.org/wiki/Recep_Tayyip_Erdo%C4%9Fan
[6]https://www.forbes.com/2007/07/23/turkey-erdogan-elections-biz-cx_0724oxford.html
[7]https://www.nytimes.com/2008/07/11/world/middleeast/11iraq.html
[8]https://www.theguardian.com/world/2009/jan/30/turkish-prime-minister-gaza-davos
[9]https://news.un.org/en/story/2010/03/330952-turkish-prime-minister-wins-first-ever-un-award-memory-slain-lebanese-leader
[10]https://edition.cnn.com/2013/06/04/world/europe/turkey-erdogan/index.html

case, there was always a plateau. Wikipedia is an open online encyclopedia and everyone can contribute to the content. It is obvious that the maintainers of Wikipedia isn't able to manage this volume of content. As a result, polarized volunteers have the chance to contribute positive or negative [1, 7] depending on the content and in line with the *LEB* and *LIB* models we have already talked about.

The same holds and in the case of *Erdogan's* biography. We observe that at its birth the biography is less subjective. Over years, more polarized contributors add their personal opinion to the biography. This goes on for several years when the more neutral authors as per the subjectivity take action. One thing that must not be confused about is that subjectivity is possible to be expressed with both positive and negative words. That means when the subjectivity increases doesn't mean that the content is in favor of the person of interest.

## 5 CONCLUSION

Ultimately, we conclude on this, our *t-test* has given us evidence that there are linguistic biases on political biographies through time. We have also plotted two biographies of different genders. The comparison of these plots confirms our results in t-test. Of course, we have focused only on a small number of European politicians and this would be the object of future work. Also, we find possible correlations of changes on the subjectivity and real events that relate to each politician. This of course could be an object for more research.

## REFERENCES


[1] Pushkal Agarwal, Miriam Redi, Nishanth Sastry, Edward Wood, and Andrew Blick. 2020. Wikipedia and Westminster. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*. nil. https://doi.org/10.1145/3372923.3404817

[2] Pablo Beytía. 2020. The Positioning Matters. In *Companion Proceedings of the Web Conference 2020*. nil. https://doi.org/10.1145/3366424.3383569

[3] Anamika Chhabra, Rishemjit Kaur, and S. R.S. Iyengar. 2020. Dynamics of Edit War Sequences in Wikipedia. In *Proceedings of the 16th International Symposium on Open Collaboration*. nil. https://doi.org/10.1145/3412569.3412585

[4] Alexandrea Melissa Hunt. 2011. The Linguistic Expectancy Bias and the American Mass Media. http://hdl.handle.net/20.500.12613/1481

[5] Jahna Otterbacher, Ioannis Katakis, and Pantelis Agathangelou. 2019. *Linguistic Bias in Crowdsourced Biographies: A Cross-lingual Examination*. WORLD SCIENTIFIC, 411–440. https://doi.org/10.1142/9789813274884_0012

[6] Gün R. Semin. nil. *The Linguistic Category Model*. SAGE Publications Ltd, 309–326. https://doi.org/10.4135/9781446249215.n16

[7] Khonzodakhon Umarova and Eni Mustafaraj. 2019. How Partisanship and Perceived Political Bias Affect Wikipedia Entries of News Sources. In *Companion Proceedings of The 2019 World Wide Web Conference*. nil. https://doi.org/10.1145/3308560.3316760

[8] William von Hippel, Denise Sekaquaptewa, and Patrick Vargas. 1997. The Linguistic Intergroup Bias As an Implicit Indicator of Prejudice. *Journal of Experimental Social Psychology* 33, 5 (1997), 490–509. https://doi.org/10.1006/jesp.1997.1332

[9] Daniël H. J. Wigboldus, Russell Spears, and Gün R. Semin. 2005. When Do We Communicate Stereotypes? Influence of the Social Context on the Linguistic Expectancy Bias. *Group Processes & Intergroup Relations* 8, 3 (2005), 215–230. https://doi.org/10.1177/1368430205053939

[10] Zena Worku, Taryn Bipat, David W. McDonald, and Mark Zachry. 2020. Exploring Systematic Bias through Article Deletions on Wikipedia from a Behavioral Perspective. In *Proceedings of the 16th International Symposium on Open Collaboration*. nil. https://doi.org/10.1145/3412569.3412573