Natural Language Processing

Supervisor: Dr. Seyed Saleh Etemadi

News Classification

Author:

Elnaz Rezaee elnazrezaee80@gmail.com

2023/2024

Contents

1 Introduction

The purpose of this project is to analyze and process fact-checking data obtained from Politifact. The data consists of statements from different categories, including ”true,” ”mostly true,” ”half true,” ”barely true,” ”false,” and ”pants on fire.” The data was collected by scraping the Politifact website and subsequently cleaned and preprocessed for further analysis.

2 Accurate data source

The data source of this project is the PolitiFact website, which categorizes the data into 6 classes with the titles stated in the introduction section. Also, for the data to be balanced, a specific number (2500) of data has been collected from each category.

3 Data collection methodology

The data collection for this report involved a systematic approach utilizing various steps and tools. The following outlines the methodology employed:

- **Selection of data sources**: The data for this report was collected from Politifact, a reputable fact-checking organization known for its unbiased analysis of statements made by politicians and public figures. Politifact provides a comprehensive repository of fact-checks, ratings, and statements, making it an ideal source for this research.
- **Web scraping**: The data was collected from the Politifact website using web scraping techniques. The `BeautifulSoup` library in Python was utilized to extract information from the HTML pages. The `requests` library facilitated the HTTP requests to retrieve the web pages.
- **Data Cleaning**: The collected data underwent a cleaning process to remove any irrelevant or noisy information. The `pandas` library was used to read the raw data, and cleaning operations were performed to ensure consistency and accuracy.
- **Text Processing**: To analyze the textual content of the statements, various text processing techniques were applied. The `nltk` library was employed to tokenize the statements into sentences, while the `spacy` library facilitated the tokenization of words for further analysis.

These steps were carried out using the Python programming language and various libraries such as `csv`, `json`, and `urllib3`. The collected data was stored in CSV format for easy accessibility and further analysis.

By following this robust data collection methodology and utilizing appropriate tools, the integrity and accuracy of the data used in this report were ensured.

4 Data format

The collected data is organized in a specific file and folder structure to ensure efficient storage and retrieval. The following outlines the structure of each file and provides an explanation of the different tags:

4.1 File Structure

The data is stored in CSV (Comma-Separated Values) format, which is a widely used format for tabular data. The data folder, contains 4 categories including raw, clean, sentencebroken and wordbroken. Each of these folders, have 6 CSV files as follows:

```
- true.csv
- mostly-true.csv
- half-true.csv
- barely-true.csv
- false.csv
- pants-fire.csv
```

Each CSV file corresponds to a specific truthfulness category and contains two columns: `Statement` and `Label`. The `Statement` column stores the fact-check statements, while the `Label` column represents the truthfulness rating associated with each statement.

4.2 Differentiation of Tags

The data collected from Politifact utilizes different tags to represent the truthfulness rating of each statement. The tags used and their meanings are as follows:

- **True**: This tag represents statements that have been fact-checked and determined to be completely accurate and truthful.
- **Mostly True**: Statements labeled as ”Mostly True” indicate that they contain some degree of truth but may also include some inaccuracies or misleading information.
- **Half True**: This tag is assigned to statements that are partially true but also contain some false or misleading elements.
- **Barely True**: Statements labeled as ”Barely True” are mostly false but may contain a small element of truth.
- **False**: This tag represents statements that have been fact-checked and determined to be false or misleading.
- **Pants on Fire**: The ”Pants on Fire” tag is used for statements that are blatantly false or entirely fabricated.

By categorizing the statements into different tags based on their truthfulness, the data allows for further analysis and evaluation of the reliability and accuracy of the statements made by politicians and public figures.

5 Pre-processing

The following pre-processing steps were performed:

5.1 Sentence separation

Sentences were separated using the nltk library. So, the method of sentence breaking is performed using the *nltk.sent_tokenize*() function.

5.2 Tokens/Words separation

Tokens/words were separated using the spaCy library. The statement is passed through the spaCy pipeline to obtain a processed document object (doc). The tokens in the document object are extracted and stored as a list of strings. So, the text is splitted by words.

5.3 Data cleaning

The data was cleaned using the implemented data cleaning method. The `clean_data` function takes a `text` parameter, which represents the input text that needs to be cleaned.

Within the function, the `re.sub()` method is used to perform regular expression-based substitution in the text. The regular expression pattern `r'[^a-zA-Z0-9. $]'` matches any character that is not an alphabet letter (lowercase or uppercase), a digit, a period, a space, or a dollar sign.

The `re.sub()` method replaces all occurrences of the matched pattern with an empty string, effectively removing them from the text. This helps in eliminating unwanted characters or symbols from the text. The cleaned text is then returned from the function.

5.4 Data size Before/After Cleaning

The data size before and after cleaning is as follows:

- Data Size Before Cleaning: 1878218 bytes
- Data Size After Cleaning: 1770027 bytes

6 Tagging unit and tagging method

The tagging unit in this project claims from the Politifact dataset.The labels represent different categories such as ”true”, ”mostly-true”, ”half-true”, ”barely-true”, ”false”, and ”pants-fire”.

The tagging method is based on the labels provided by Politifact, which are assigned to the statements based on their assessment of truthfulness. Each statement is associated with a specific label that indicates its categorization or classification.

The tagging process involves extracting the statements from the dataset and assigning the corresponding label to each statement. This is done by parsing the HTML content of the Politifact website, retrieving the statement and label information, and storing them in separate CSV files.

7 Data statistics

7.1 Number of units of data

Label	Count
TRUE	2580
mostly-true	2580
half-true	2580
barely-true	2580
FALSE	2580
pants-fire	2580

7.2 Number of sentences

Label	Sentences Count
TRUE	2970
mostly-true	3002
half-true	3075
barely-true	3080
FALSE	2807
pants-fire	2948

7.3 Number of words

Label	Words Count
TRUE	49988
mostly-true	51841
half-true	53107
barely-true	51355
FALSE	42005
pants-fire	43586

7.4 Unique word count

Label	Unique Words Count
TRUE	7142
mostly-true	7138
half-true	7578
barely-true	8012
FALSE	8232
pants-fire	8393

7.5 Number of unique words common and uncommon between labels

Common	Uncommon
1929	19207

7.6 10 most frequent non-shared words from each label

TRUE	mostly-true	half-true	barely-true	FALSE	pants-fire
11yearold	1.1	sea	safetynet	azidothymidine	porn
boy	rebates	Gore	Gormans	AZT	Whitmers
Infertility	cuz	polar	poem	Ma	thwart
differently	covers	workingage	giant	presenting	blackout
excluded	lastminute	choosing	cloudbased	software	Utopia
MiamiDade	antidiscrimination	preventing	central	Quantum	series
Venezuela	operates	promotions	grid	AI	intentional
particular	theme	objects	convince	Publix	sterilize
remains	benefiting	rescinded	Quad	traditional	Boiling
unchanged	Walt	Individuals	stability	began	tap

7.7 Top 10 common words of each label compared to other labels based on relative frequency

TRUE	mostly-true	half-true	barely-true	FALSE	pants-fire
the	the	the	the	the	the
of	in	in	to	to	to
in	of	of	of	a	a
to	to	to	in	in	in
a	a	a	a	of	of
and	and	and	and	and	Says
for	Says	Says	Says	is	and
is	for	for	for	for	is
Says	is	that	is	that	for
that	that	is	that	Says	that

7.8 Top 10 words of each label based on TF-IDF(w_i)

TRUE	mostly-true	half-true	barely-true	FALSE	pants-fire
11yearold	differently	nicaragua	unchanged	player	troop
boy	excluded	venezuela	homeownership	wisconsin	nycs
saved	infertility	14000	particular	brewers	scout
lives	treated	miamidade	relatively	plays	cookies
same	often	haiti	fair	or	shelters
day	issues	cuba	1968	generate	please
two	coverage	10000	remains	baseball	consider
an	insurance	came	housing	basketball	entirely
people	other	countries	african	games	buying
the	from	schools	signed	football	6000

7.9 Histogram of the number of repetitions of each unique word

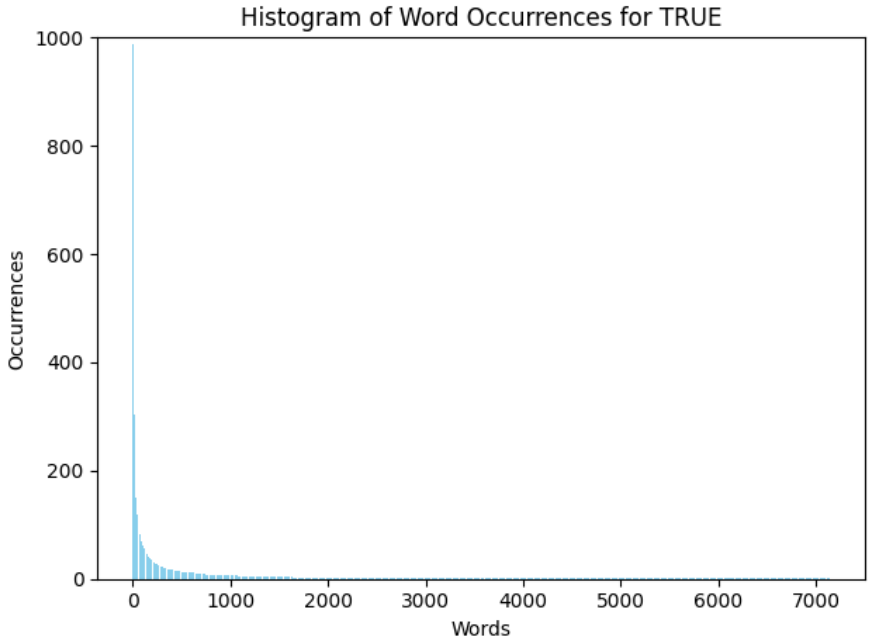

Figure 1: Histogram of label "True"

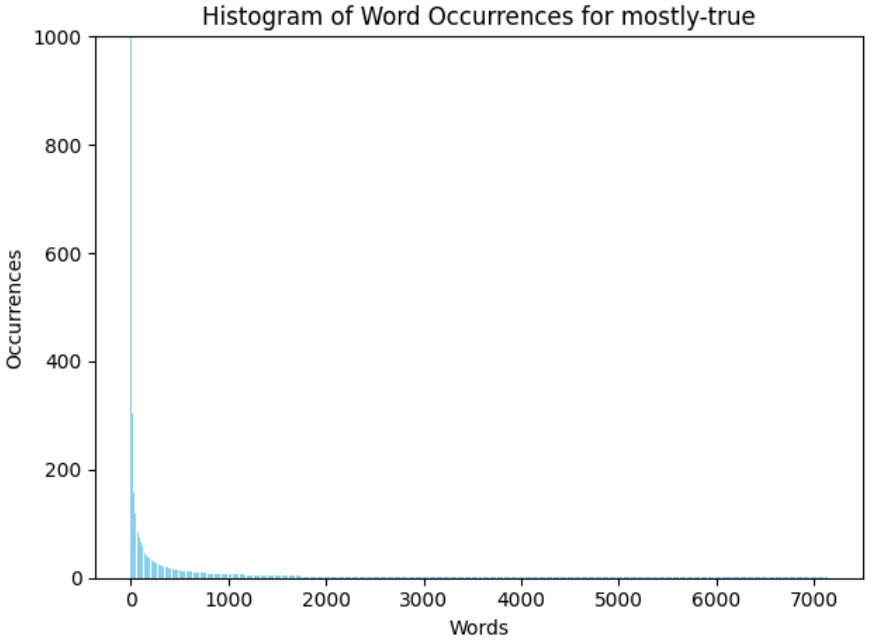

Figure 2: Histogram of label "Mostly True"

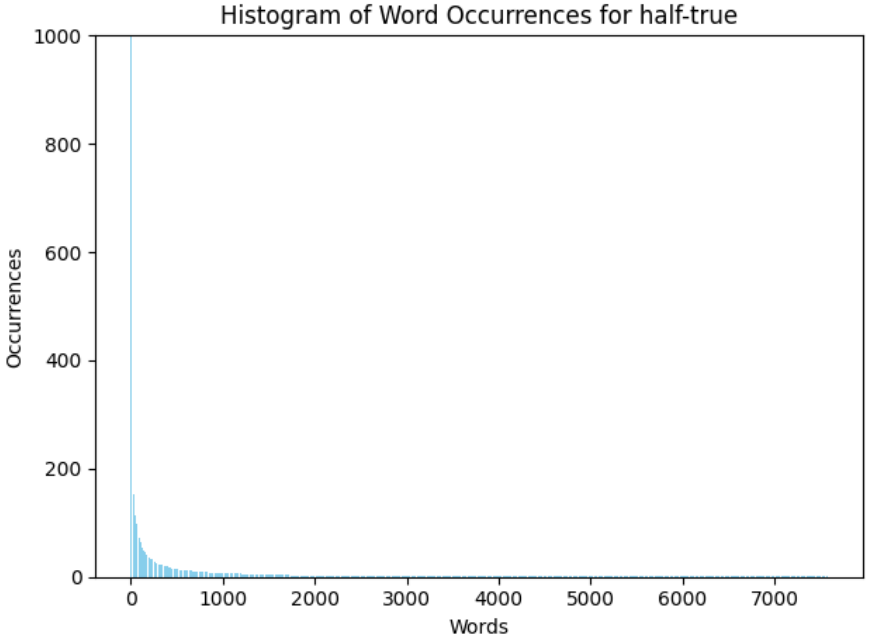

Figure 3: Histogram of label "Half True"

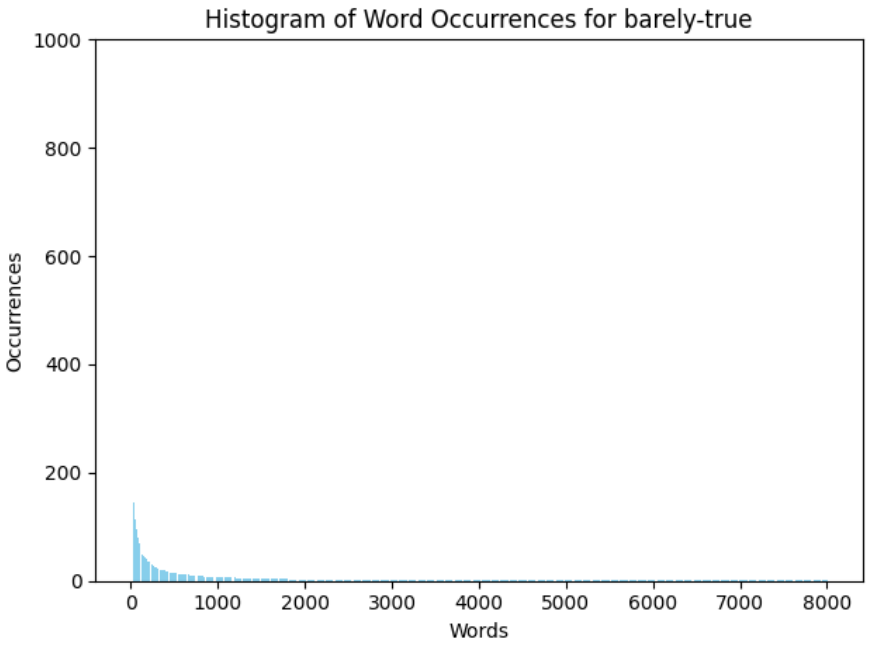

Figure 4: Histogram of label "Barely True"

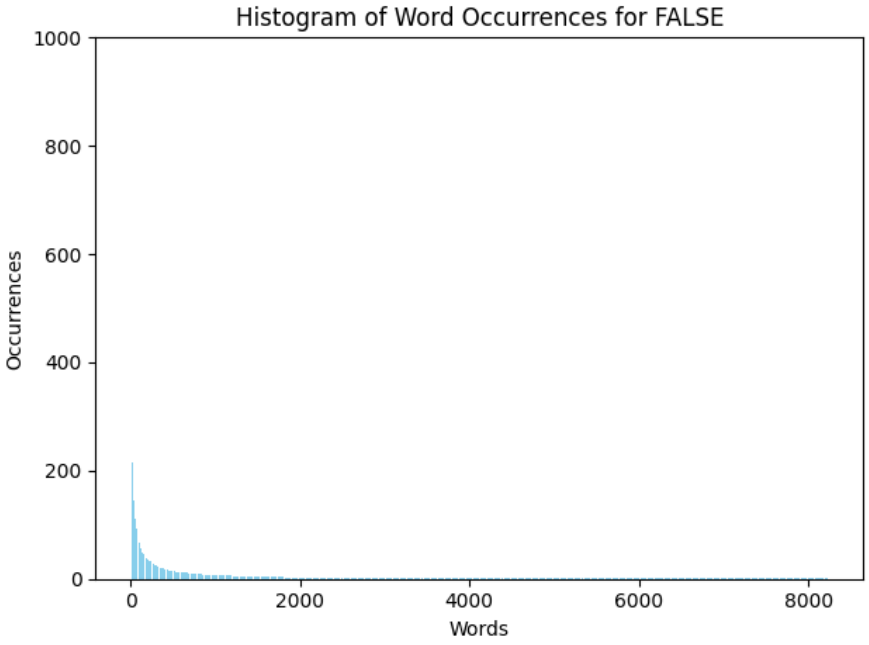

Figure 5: Histogram of label "False"

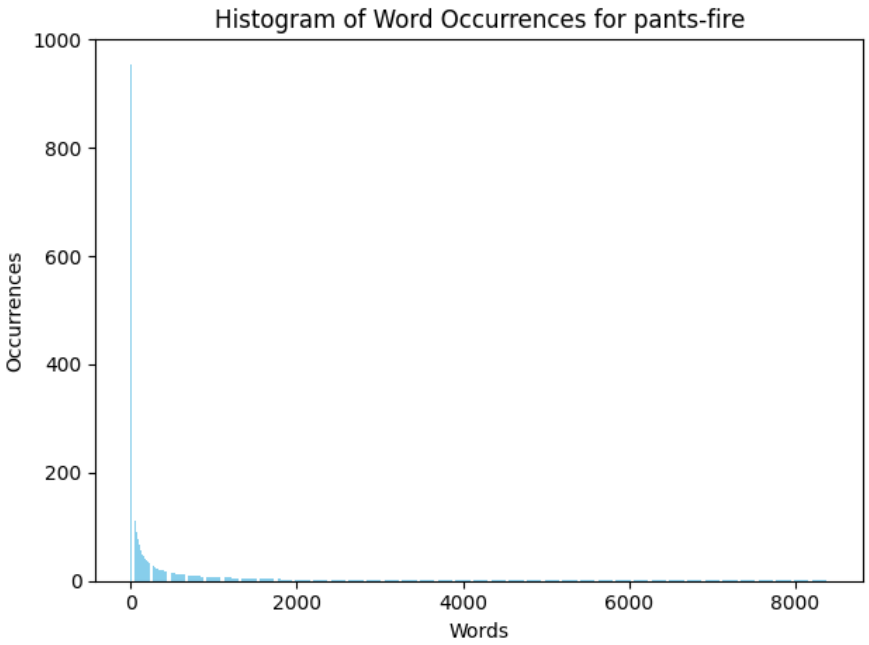

Figure 6: Histogram of label ”Pants on Fire”

Click here to see my github.
Click here to see my dataset.