Romanian leaflet analysis
How complex are they?

1st Semester of 2023-2024

Magureanu Stefan-Ionut and **Sinca Silviu-Gabriel** and **Danciu Maryo-Razvan**
{stefan-ionut.magureanu, silviu-gabriel.sinca, maryo-razvan.danciu}@s.unibuc.ro

Abstract

In this paper we presented an analysis on medical leaflets written in Romanian from 3 categories: leaflets from Romanian drug companies, Polish drug companies and GlaxoSmithKline. Our goal is to show how readability level varies among these categories. Our main target for which we evaluate this level is mainly elderly people coming from an impoverished background.

1 Introduction

A big problem in Romania is people taking different kinds of medication without prescription or without reading the leaflet first. We focused our research on people that got their education during communism, probably not even finishing high school. We assumed that some of those people are still having their roots in that period and didn't try to expand their knowledge.

Initially, the three of us started by gathering the leaflets required for our analysis. Each of us had to extract 50 leaflets from ANM (the Romanian National Drug Agency). We also chose a few books from the communist & pre-communist era that served as our main corpus. As for our personal contributions, here are the most relevant we each had:

1. Stefan
    - came with the idea of selecting a bunch of communist articles from *Gazeta Literara* to improve our existing corpus[1]
    - was responsible for implementing our second analysis method
    - came with the main idea for the corpus of our first analysis method: for it to contain only words that convey emotions (adjectives, verbs & adverbs)
2. Silviu
    - came with the idea of incorporating our second analysis method: adapting Dale-Chall's readability formula
    - plotted the results from our first analysis method
    - implement the T-Test & Permutation Test for the data that resulted from our second analysis method
3. Razvan
    - implemented a way to run the second analysis method
    - plotted the results from our second analysis method

For the analysis, we had 2 main method to estimate the complexity of the leaflets per category

1. Complex word frequency - which turned out not to be so accurate due to the lack of variety of metrics
2. Dale-Chall's readability formula, adapted to our needs - turned out to be the better approach for our purpose. The variety of metrics combined with the data preprocessed into multiple chunks contributed to finding what we were seeking.

The main idea of our research was to present if the medical leaflets are simple enough or not for ordinary people to understand. Our assumption was that the leaflets from Romanian drug companies have a lower level of complexity, due to the lack of language barrier between companies and clients.

We first investigated if there were already any papers written on this subject, but unfortunately we did not find any close to our subject matter. This also fulfills our desire to spark the interest for this particular domain into further research.

[1] https://adt.arcanum.com/ro/collection/RomaniaLiterara/

2 Approach

All the code and resources used for this project is located on our github repository[2]. For the coding part of the project, we developed our program in Google Colab using the Python programming language, version 3.10.12. We also used the following 3rd party libraries: *spacy* for natural language processing, *pandas & numpy* for data analytics, *seaborn, matplotlib & wordcloud* for data visualization and *scipy* for statistics.

Our first step was to gather enough data for our corpus in order to simulate as much as possible the average vocabulary for our target subjects. This was made possible by collecting some well-known literature books that were and still are mandatory to be studied in Romanian high school, some of which being *Ion* - Liviu Rebreanu or *Enigma Otiliei* - George Calinescu. We also added a bunch of old articles from the 50's and 60's from *Gazeta Literara*.

We continued by selecting 150 medical leaflets from ANM [3], 50 for each category we chose.

After we finished with the first step, we started with data preprocessing, which consisted in filtering words by different criteria such as part of speech or number of appearances in the corpus and bringing them to their dictionary form (lemma). We called this data our **vocabulary**.

Subsequent to the initial preprocessing, we started implementing our first analysis method. We defined complex words as being terms that are not present in the vocabulary. This is based on our assumption that people didn't encounter those words upon reading them in the leaflet.

Initial results showed that our assumption was correct, meaning that Romanian Drug Companies produced the least complex leaflets, but to our surprise GlaxoSmithKline ranked considerably the worst out of all three (see Figure 1). You can also visualize a wordcloud of the most frequent complex words for each category in figures 2, 3 and 4. After this discovery, bearing in mind that this method of analysis is very surface level, we investigated further this issue.

We continued with exploring different kinds of text analysis: sentence length, syllables count, connector words count, word length. However, we felt that we needed a reference point for our own assumption about the way texts, in this con-

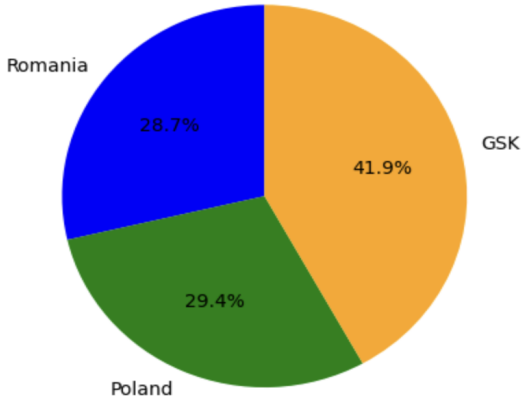

Figure 1: Complex words distribution

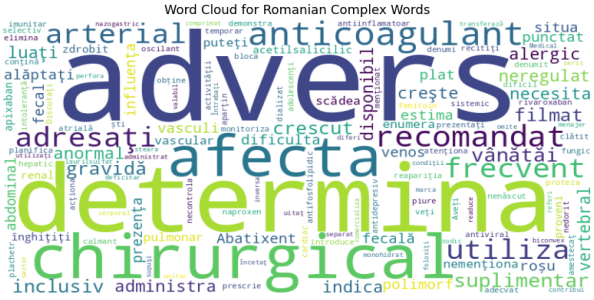

Figure 2: Romania

Figure 3: Poland

Figure 4: GSK

[2]https://github.com/silviusinca/romanian-leaflet-analysis
[3]https://nomenclator.anm.ro/medicamente

Score	Notes
4.9 or lower	easily understood by an average 4th-grade student or lower
5.0–5.9	easily understood by an average 5th- or 6th-grade student
6.0–6.9	easily understood by an average 7th- or 8th-grade student
7.0–7.9	easily understood by an average 9th- or 10th-grade student
8.0–8.9	easily understood by an average 11th- or 12th-grade student
9.0–9.9	easily understood by an average college student

Figure 5: Dale-Chall readability score https://en.wikipedia.org/wiki/Dale-Chall_readability_formula

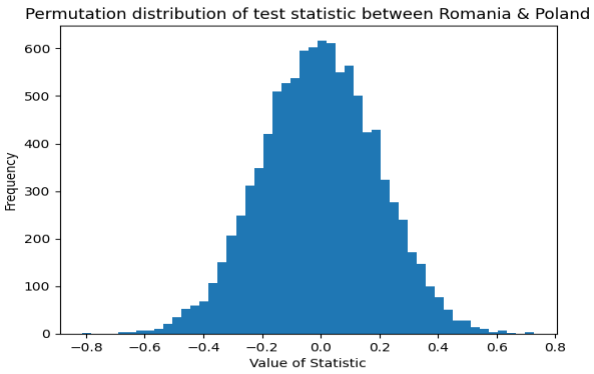

Figure 6: Permutation Test with leaflets from Romania & Poland

text leaflets, should be analysed. Therefore, we stumbled upon Dale-Chall's readability score from 1948[4].

```
if complex_word_percentage < 5:
    raw_score = 0.1579 *
     complex_word_percentage +
     0.0496 * average_length_sents
else:
    raw_score = 0.1579 *
    complex_word_percentage +
     0.0496 * average_length_sents
     + 3.6365
```

Although this scoring method (see Figure 5) was created in the 40's United States, we observed that the scores are relevant even today to literature books written in Romanian, by testing the formula on a couple of relevant books, way different from the ones used in the corpus, but keeping the same period in mind. Our output was as follows:

```
[SAMPLE] Ciresarii Score:
6.568284606645282
[SAMPLE] Little Prince Score:
6.223922440220505
```

This means that the program approximated said books as being easily understood by a 7th grade student, which makes the formula still relevant.

```
Romania Score:  10.2172380546041
Poland Score:  10.474540732983243
GSK Score:  9.97651142265883
```

Surprisingly, GlaxoSmithKline leaflets readability score was much better compared to the first analysis, but also Romania and Poland have similar scores, which is in agreement with our initial findings.

[4]See references

In order to reject the null hypothesis (scores from two different chunks are equal) we ran a few T-Tests and a few Permutation Tests (see figures 6, 7 and 8). We accomplished this by splitting leaflets into 2000 word chunks of text. According to the *p-values* resulted, we are confident to state that the null hypothesis is false and our ranking is accurate.

T-Test:

```
score_romania =
    np.array(complexity_chunks_romania)
score_poland =
    np.array(complexity_chunks_polonia)
score_gsk =
    np.array(complexity_chunks_gsk)

res_romania_poland =
    stats.ttest_ind(score_romania,
        score_poland)
res_romania_gsk =
    stats.ttest_ind(score_romania,
        score_gsk)
res_poland_gsk =
    stats.ttest_ind(score_poland,
        score_gsk)
```

T-Test results:

```
Romania v Poland p-value, statistic:
0.0136855431754127 -2.502596789450564
Romania v GSK p-value, statistic:
0.01461816201622772 2.4820409597836184
Poland v GSK p-value, statistic:
1.6641950810274703e-06 5.03592011603592
```

3 Limitations

The process of gathering data was rather tedious, especially with the leaflets. For example, initially

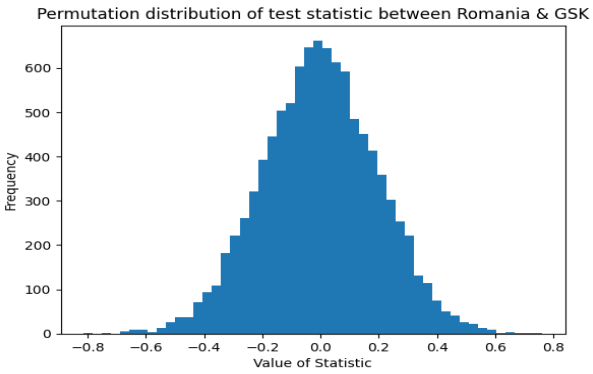

Figure 7: Permutation Test with leaflets from Romania & GSK

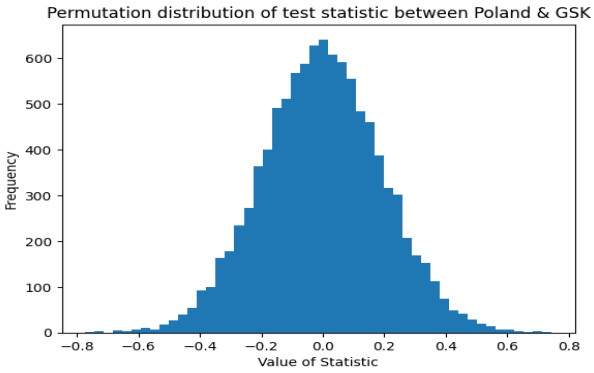

Figure 8: Permutation Test with leaflets from GSK & Poland

we wanted to inspect Pfizer leaflets instead of GSK, but we only manage to find around 20 samples, and that was not enough. Another big impediment was trying to find the words from the basic vocabulary of the Romanian language in order to enhance the relevance of our initial corpus.

4 Conclusions and Future Work

We are satisfied with our results, but there is still room for improvement. A way of doing this is by expanding the study to other social categories and also improving our corpus by adding relevant texts.

As for the course of Archaeology of Intelligent Machines, we really loved it because of the unique teaching methods and the freedom of choice for the projects.

We enjoyed the process of doing this research and came to the conclusion that there are still lots of areas that wait to be investigated and that the fun never stops.

References

Maqsood S, Shahid A, Tanvir Afzal M, Roman M, Khan Z, Nawaz Z, Aziz MH. Assessing English language sentences readability using machine learning models. PeerJ Comput Sci. 2022 Jan 4;8:e818. doi: 10.7717/peerj-cs.818. PMID: 35111913; PMCID: PMC8771811.

Dale, Edgar, and Jeanne S. Chall. "The Concept of Readability." Elementary English, vol. 26, no. 1, 1949, pp. 19–26. JSTOR, http://www.jstor.org/stable/41383594. Accessed 5 Feb. 2024.