Brief Raport, Mateusz Guściora, 02.04.2024

## Steps

a) Plan, design and Set hypothesis
b) Implement:
    I. Loading libraries and data files
    II. Quick Overview-general info
    III. Cleaning data and preliminary analysis
    IV. Text preprocessing
    V. Sentiment analysis-adding new information to dataset
    VI. Testing Hypothesis
    VII. Data presentation and evaluation
c) Raport

## Data

Set used: "*job_postings.csv*"

Attributes with possible interesting informations: *description(job), company_description, views, applies, formatted_work_type, sponsored, skills_desc, formatted_experience_level*

Analysis script: "*zadanko-linkledin-job-posting.ipynb*"

## Hypothesis

**Hypothesis 1** (H1): *Job postings with positive emotional tone in their descriptions are associated with a higher number of applicant views or/and applications, compared to those with neutral or negative tones.*

H0: *There is no association between the emotional tone of job postings (positive, neutral, negative) and the number of applicant views or applications.*

- if p_value < 0.05 it supports H1
- if p_value >=0.05 it supports H0 and therefore no significant findings
- Furthermore results can be either**: consistent findings** (if all methods show similar results) or **inconsistent findings** (if more than one of methods show different results)

## Questions and Answers during analysis

Q: Does sentiment have impact on number of views, applies?

A: Sentiment does not appear to significantly impact no. of views or applications

Q: Does methods for hypothesis testing are right for this data?

A: There is no normal distribution in data so ANOVA should not probably be used. The rest methods are appropriate but reveal no strong patterns in the data.

Q: How does the method of sentiment analysis (TextBlob vs. NLTK) impact evaluation?

A: Lead to different evaluation outcomes (sentiment scores) and different results in tests.

Q: Is there consistency in results? A: There is not. While most indicate no relation. Three results does.

Q: Can the applies-to-views ratio be a more insightful measure than raw applies or views counts?

A: It can, it shows relation between applies and views but not for this problem as there is not concluding pattern or association between sentiment and former.

Q: What additional factors, beyond sentiment, might influence the number of views and applications a job posting receives? (not tested))

## Evaluation of hypothesis and results

Analysis suggests a lack of consistent association between job posting sentiment and number of views or applications, with no strong evidence to support a significant impact of sentiment on applicant behavior, leaning towards the null hypothesis (H0). While there are instances where certain sentiment categories show potential significance, such as with the Chi-Square test for applications using TextBlob, these are not uniformly observed across all tests [*Table1*].

| | Method | Test | P-value | Potential Association |
|---|---|---|---|---|
| 0 | TextBlob | ANOVA_views | 4.066679e-01 | No |
| 1 | TextBlob | ANOVA_applies | 4.459067e-01 | No |
| 2 | TextBlob | ANOVA_ratio | 3.241595e-04 | Yes |
| 3 | TextBlob | Kruskal_Wallis_views | 6.603355e-01 | No |
| 4 | TextBlob | Kruskal_Wallis_applies | 4.035592e-01 | No |
| 5 | TextBlob | Kruskal_Wallis_ratio | 4.958077e-02 | Yes |
| 6 | TextBlob | Chi_Square_views | 1.033024e-01 | No |
| 7 | TextBlob | Chi_Square_applies | 1.300143e-40 | Yes |
| 8 | TextBlob | Chi_Square_ratio | 1.334098e-01 | No |
| 9 | NLTK | ANOVA_views | 9.937567e-01 | No |
| 10 | NLTK | ANOVA_applies | 2.794443e-01 | No |
| 11 | NLTK | ANOVA_ratio | 3.732230e-01 | No |
| 12 | NLTK | Kruskal_Wallis_views | 6.116439e-01 | No |
| 13 | NLTK | Kruskal_Wallis_applies | 2.797095e-01 | No |
| 14 | NLTK | Kruskal_Wallis_ratio | 2.501903e-01 | No |
| 15 | NLTK | Chi_Square_views | 9.250615e-01 | No |
| 16 | NLTK | Chi_Square_applies | 6.390631e-01 | No |
| 17 | NLTK | Chi_Square_ratio | 4.476123e-01 | No |

Table1: P-value Results of testing hypothesis with evaluation (yes, no, dk)

The visualized data also does not indicate a distinct pattern of increase or decrease in average views or applications in relation to sentiment categories [*Figure1*].

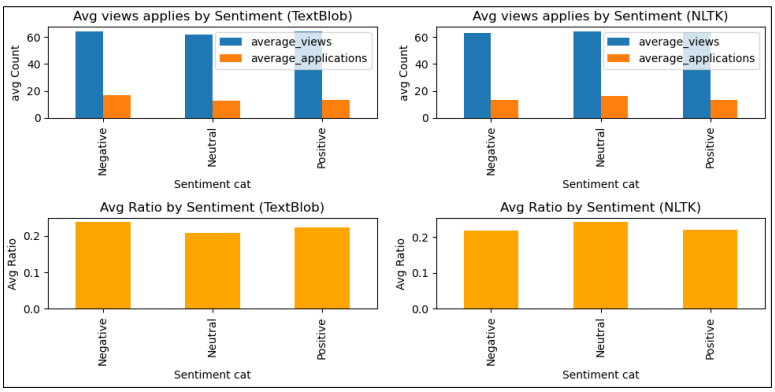


Figure1: Distribution of avg. views, applies by sentiment category on two different methods

## Setup info

Setup: VC, jupyter, python, libraries (*provided in the notebook), CPU

Methods for sentiment score calculation: textblop, nltk vader

Methods for hypothesis testing: ANOVA, Kruskal-Wallis , Chi-square-test

Run time of notebook: 818.93 seconds (13.60 minutes)

Sentiment method textblob: 167.05 seconds (2.70 minutes)

Sentiment method nltk: 640.49 seconds (10.60 minutes)