



# Practical Data Science

## Assignment A1

Giagkos Stylianos  
f3352410

# Summary

Section A  
“Emotion Annotation on Tweets”

3/42

Practical Data Science Assignment 1

Section B  
“Greek Proverbs Analysis”

27/42

Practical Data Science Assignment 1

Q&A

41/42

Practical Data Science Assignment 1

# Section A

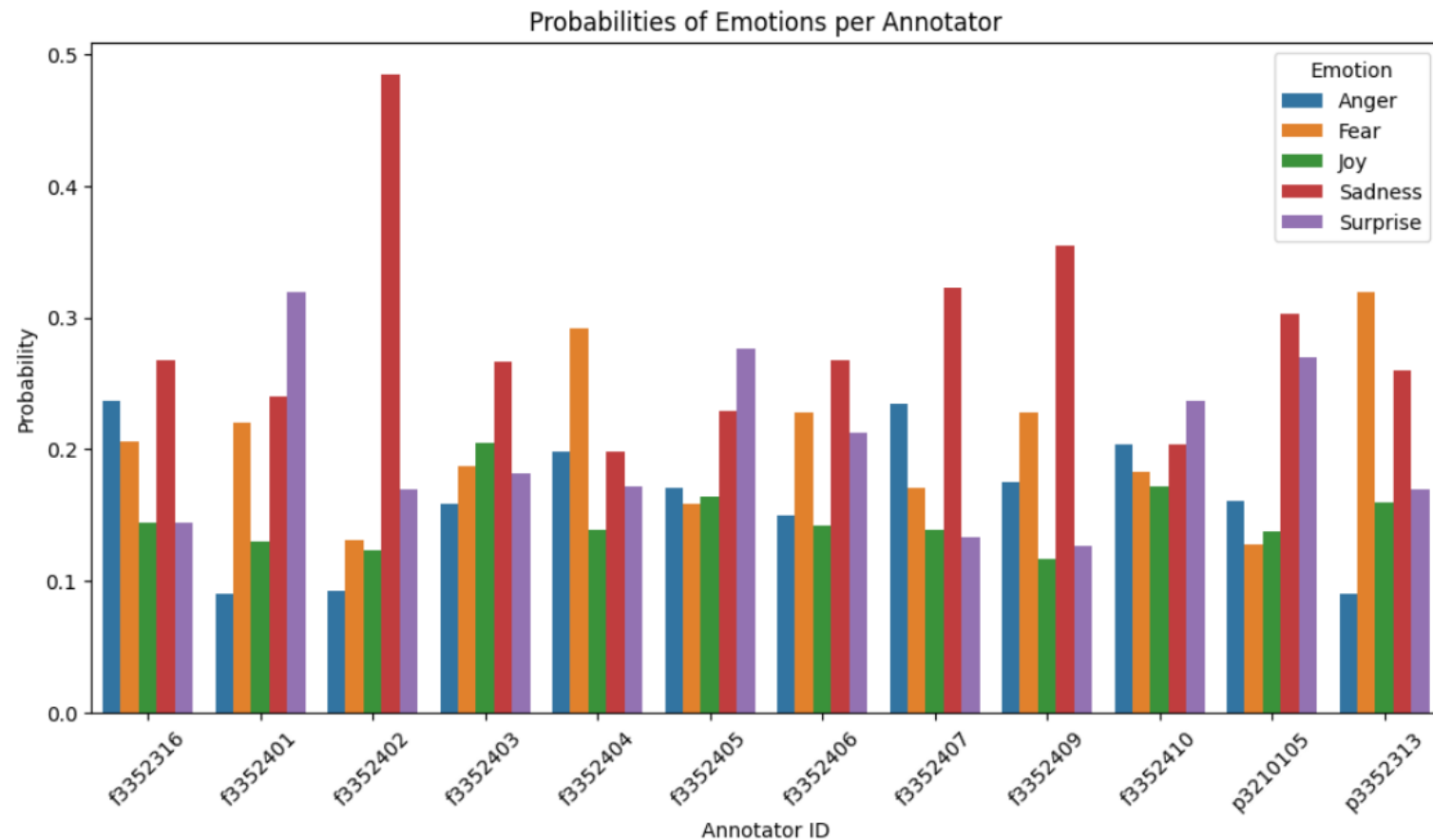
“Emotion Annotation on Tweets”

# 1. Basic

# Clarification

---

- Calculating probabilities: Divide total counts of each emotion by sums.
- Resulting probabilities: Probabilities sum to 1 across all emotions calculated.
- Previous method: Counts calculated from total annotations caused inaccuracies in probabilities.
- Normalization best practice: Normalize probabilities based on total counts across all rows.



- **Sadness Dominance:** Higher probabilities, notably for f352402, f352407, and f352409.
- **Significant Fear Levels:** High for annotators f352401, f352403, f352404, and p3352313.
- **Joy & Anger:** Generally low, indicating a dataset skewed towards negative emotions.
- **Surprise:** Moderate levels, often comparable to Fear or Joy.
- **Anger:** Low overall; higher values from few annotators (e.g., f352316, f352407).
- *Variation among annotators reflects diverse interpretations, possibly influenced by individual biases or perceptions.*
- **Examples:**
  - f352405, p3210105: Balanced across emotions.
  - f352402, f352407: Skewed towards Sadness and Fear.

## 2. IAA

- Percentage Agreement

$$\text{Percentage Agreement} = \frac{\text{Number of agreements}}{\text{Total number of items}} \times 100$$

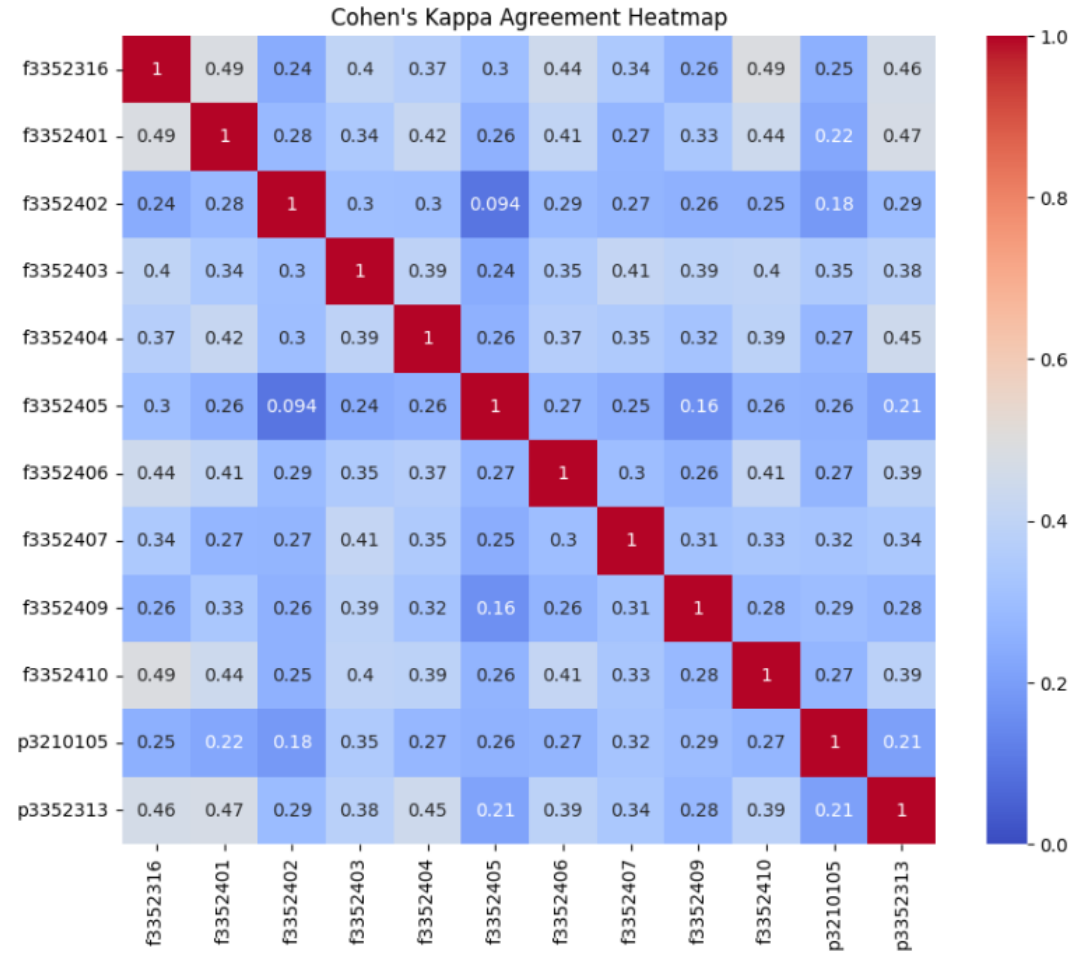
- Cohen's Kappa

$$\kappa = \frac{P_o - P_e}{1 - P_e} \qquad P_e = \sum_{i=1}^k (p_{i1} \times p_{i2})$$



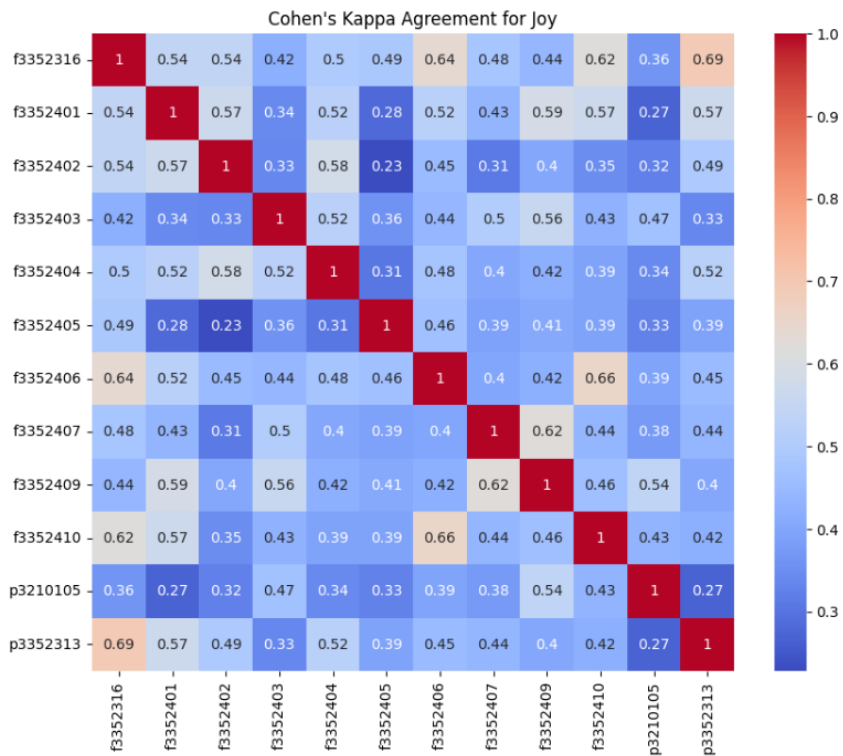
# Overall Agreement For Human Annotators

Average Kappa is: 0.38 Average percentage is: 0.32

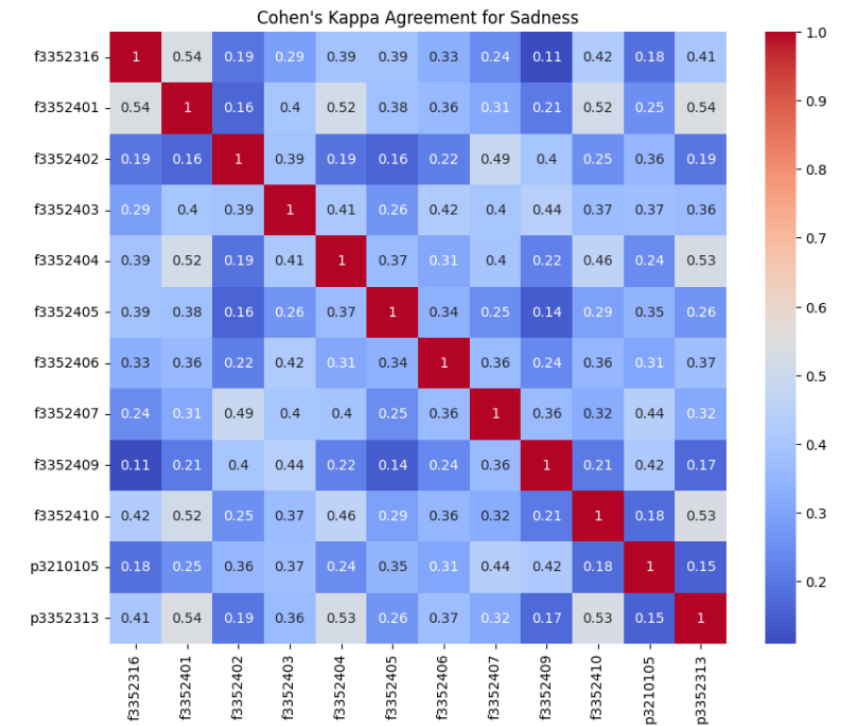


- **Purpose of overall agreement:** Measures annotator agreement across all emotions efficiently.
- **Benefits of the approach:** Quick view, easy to understand, saves time.
- **Resulting matrices:** Aggregates results into pairwise\_kappa and pairwise\_percentage matrices.

# Per-Emotion Agreement For Human Annotators



- **Per-Emotion Metrics:**  
Analyzing agreement per emotion highlights disagreements.
- **Joy's High Agreement:**  
Joy has high agreement; moderate reliability.
- **Sadness's Moderate Agreement:**  
Sadness shows reasonable agreement, moderate reliability.
- **Agreement vs. Reliability:**  
High agreement, low Kappa suggests inflated reliability.



## | 3. Ground-truth

```
print(ground_truth)
```

	Anger	Fear	Joy	Sadness	Surprise
0	0.0	1.0	0.0	0.0	1.0
1	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	1.0
3	0.0	0.0	0.0	1.0	0.0
4	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...
111	1.0	0.0	0.0	0.0	0.0
112	0.0	0.0	0.0	0.0	1.0
113	0.0	0.0	0.0	1.0	0.0
114	0.0	0.0	0.0	1.0	0.0
115	0.0	0.0	0.0	0.0	0.0

```
[116 rows x 5 columns]
```

- **Multiple 1 values:** Indicates majority agreement on emotions for each instance.
- **Meaning of agreement:** Several annotators recognizing emotions reflect in ground truth.
- **Example scenario:** Three out of five annotators identifying Fear and Surprise.
- **Recording agreement:** Both emotions recorded as 1 in the ground truth.

## | 4. LLMs

# Creating a Blank DataFrame for LLM Annotation with Prompting



Platform: Ollama  
Model: Llama3.2

	text	Anger	Fear	Joy	Sadness	Surprise
0	My mouth fell open `` No, no, no... I..					
1	You can barely make out your daughter's pale f...					
2	But after blinking my eyes for a few times lep...					
3	Slowly rising to my feet I came to the conclus...					
4	I noticed this months after moving in and doin...					
...	...	...	...	...	...	...
111	"ARCh stop your progression.					
112	This 'star', starts to move across the sky.					
113	and my feet hurt.					
114	so i cried my eyes out and did the drawing.					
115	They were coal black.					

```
import ollama
import re

def annotate_emotions(text):
    # Generate text using ollama with the desired prompt
    prompt = (f"Analyze the following text for emotions (Anger, Fear, Joy, Sadness, Surprise):\n"
              f"\"{text}\"")
    f"Return a binary output (0 or 1) for each emotion without explanations.")

    # Call ollama.generate to get the response
    response = ollama.generate(model='llama3.2', prompt=prompt)

    # Access the text content within the response dictionary
    response_text = response["response"]

    # dictionary to hold emotion values
    emotion_dict = {'Anger': 0, 'Fear': 0, 'Joy': 0, 'Sadness': 0, 'Surprise': 0}
```

```
Text: ; ) In the evening we did go down and put our feet in the water-I got to about my waist actually.

Emotions: {'Anger': 0, 'Fear': 0, 'Joy': 0, 'Sadness': 0, 'Surprise': 1}
```

# | 5. Agents

# Code Highlights for Annotation Process

```
prompt = (f"Analyze the following text\n"
  f" << \"{text}\" >> for emotions (Anger, Fear, Joy, Sadness, Surprise) \n"
  f"considering you think like a: {age} years old {gender} from {region}.\n"
  f"Return a binary output (0 or 1) for each emotion without explanations.")
```

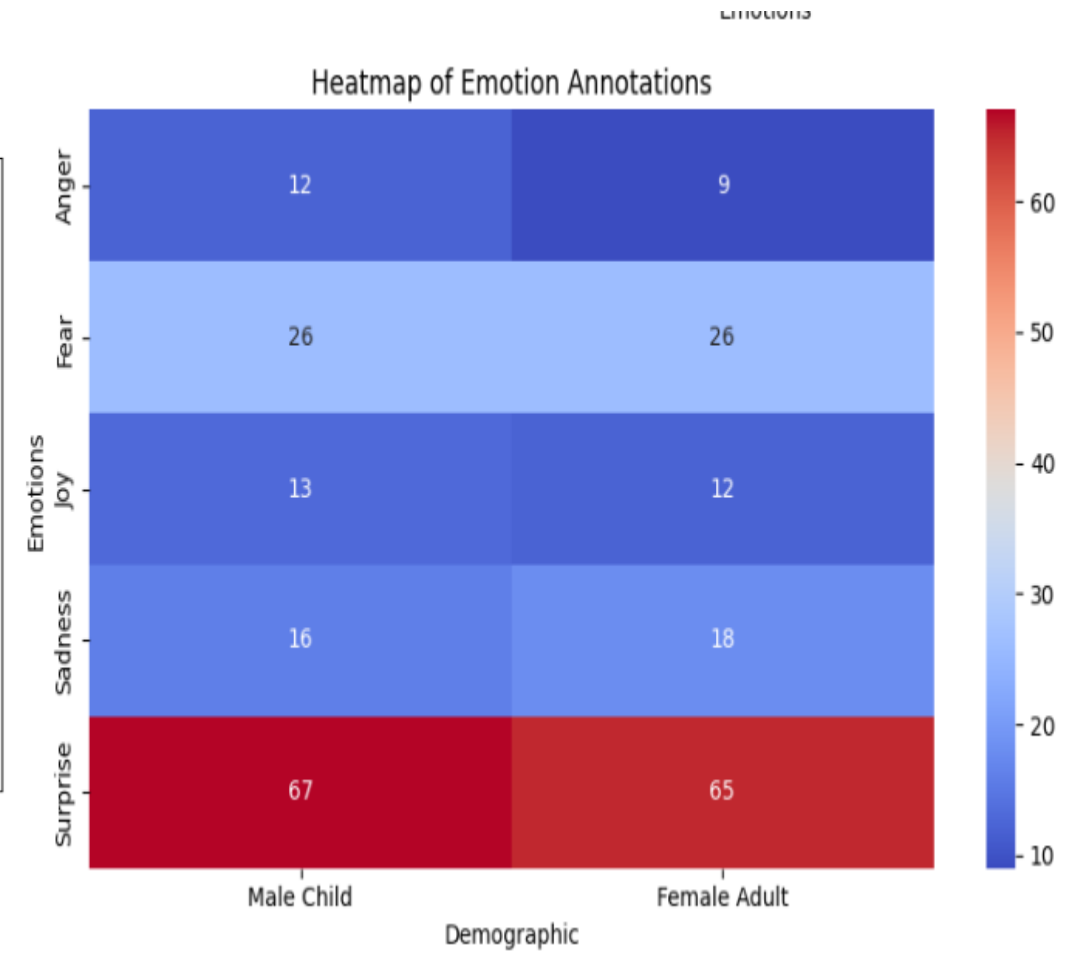
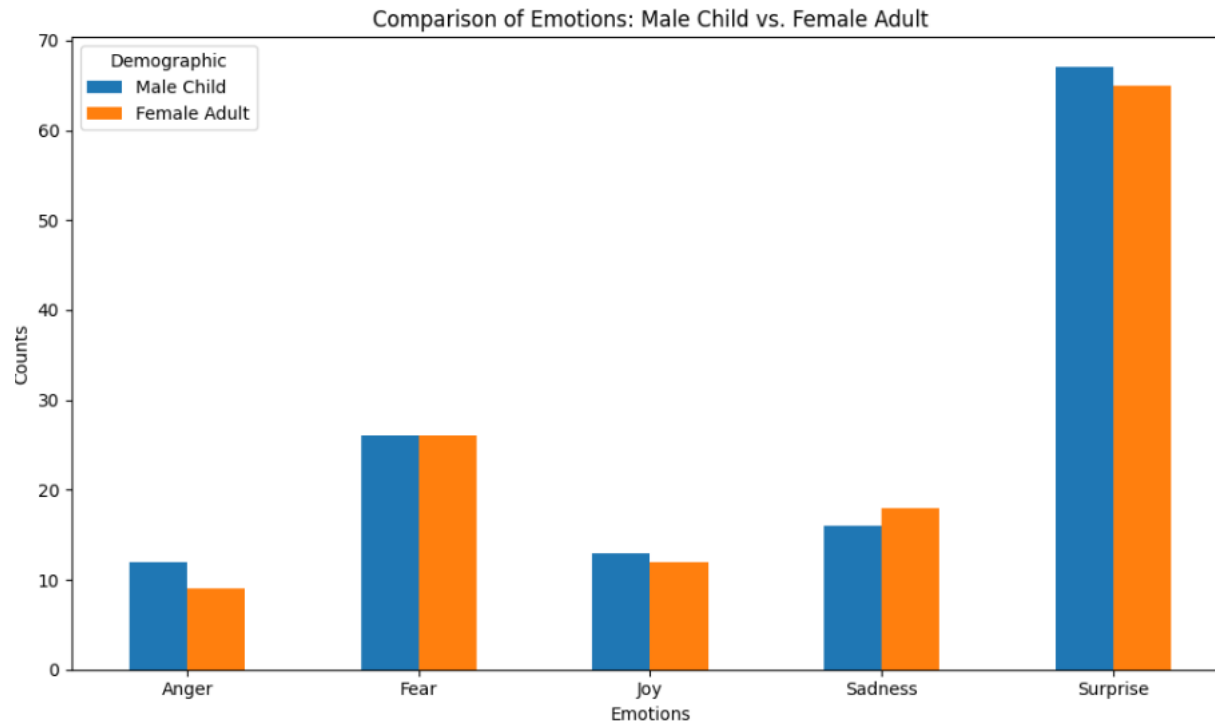
```
#[-*]?: Matches an optional '-' or '*' at the start.
#\s*: Matches any whitespace (space, tab, etc.) zero or more times.
#(\w+): Matches one or more word characters (letters, digits, or underscores) and captures them in a group.
#:: Matches a literal colon.
#\s*: Matches any whitespace zero or more times.
#(\d+(?:\.\d+)?)?: Matches a number (one or more digits).
```

```
emotions_male = annotate_emotions_instructed(row['text'], '17', 'male', 'New York')

# Annotate for Female adult
emotions_female = annotate_emotions_instructed(row['text'], '56', 'female', 'Russia')
```



# Relevant plots



# Relevant plots

	Emotion	Percentage Difference
0	Anger	-25.000000
1	Fear	0.000000
2	Joy	-7.692308
3	Sadness	12.500000
4	Surprise	-2.985075]

Chi-Square Statistic: 0.5560431070246703

P-Value: 0.9678208001393466



**Observed differences:** Notable disparities in Anger, Joy, Sadness annotations between groups.

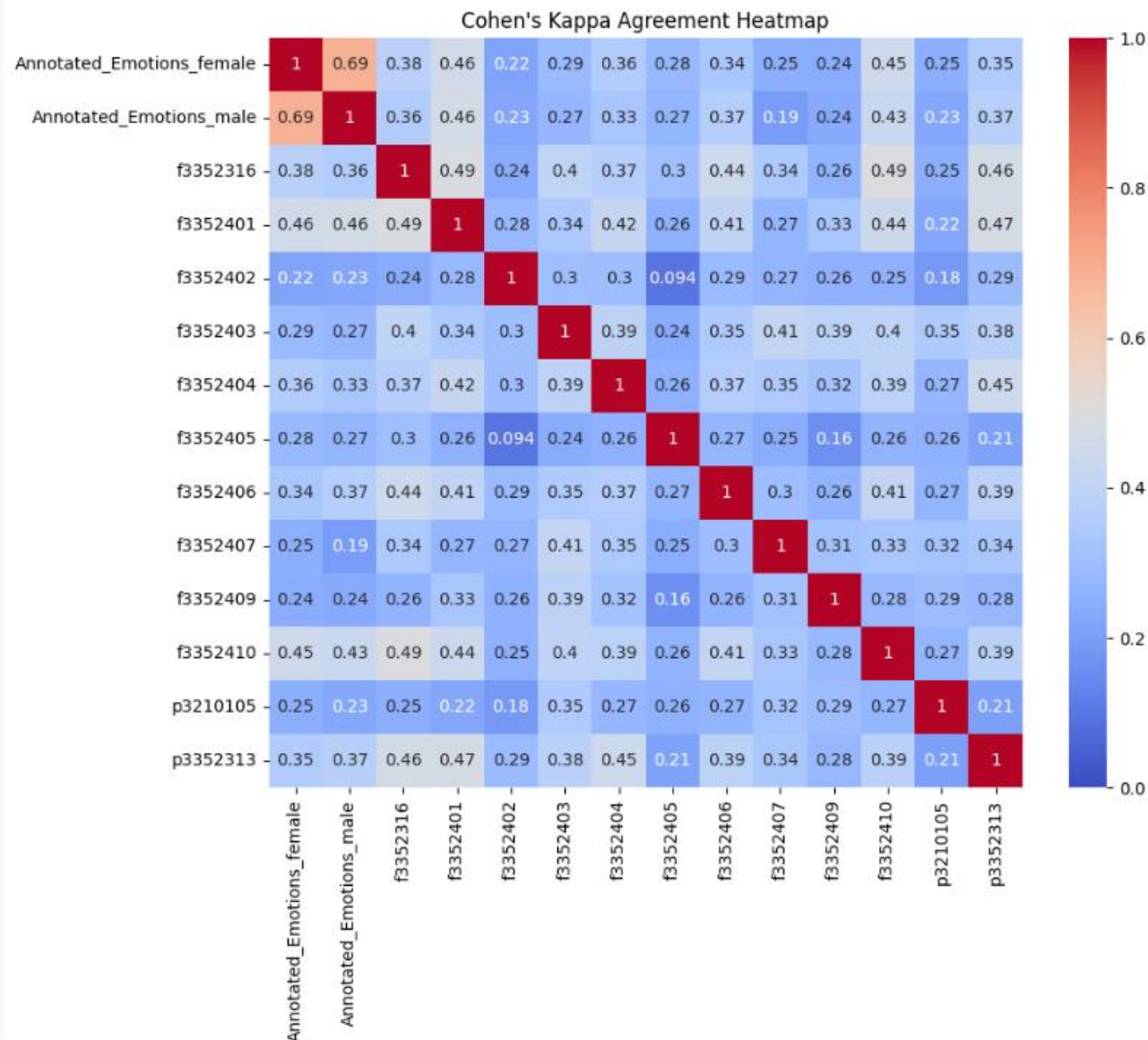
**Statistical test results:** Chi-square test reveals no significant differences in emotion counts.

**Interpretation:** Observable differences likely not meaningful due to statistical insignificance.

**Summary insight:** Demographic factors may not significantly influence emotional expression here.

- Overall Inter Annotation Agreement between Agents and Human Annotators

Average Kappa is: 0.37 Average percentage is: 0.31



### ■ Highest Agreement:

Female Annotator & Male Annotator:

Highest Kappa score of 0.69, suggesting similar standards in emotional interpretation (both generated by the same LLM).

### ■ Other Strong Agreements:

Female Annotator & f3352401: High Kappa score, indicating similar annotation style.

### ■ Areas of Low Agreement:

Annotators like f3352402: Frequently display low Kappa scores (0.2-0.3), reflecting differences in emotional interpretation.

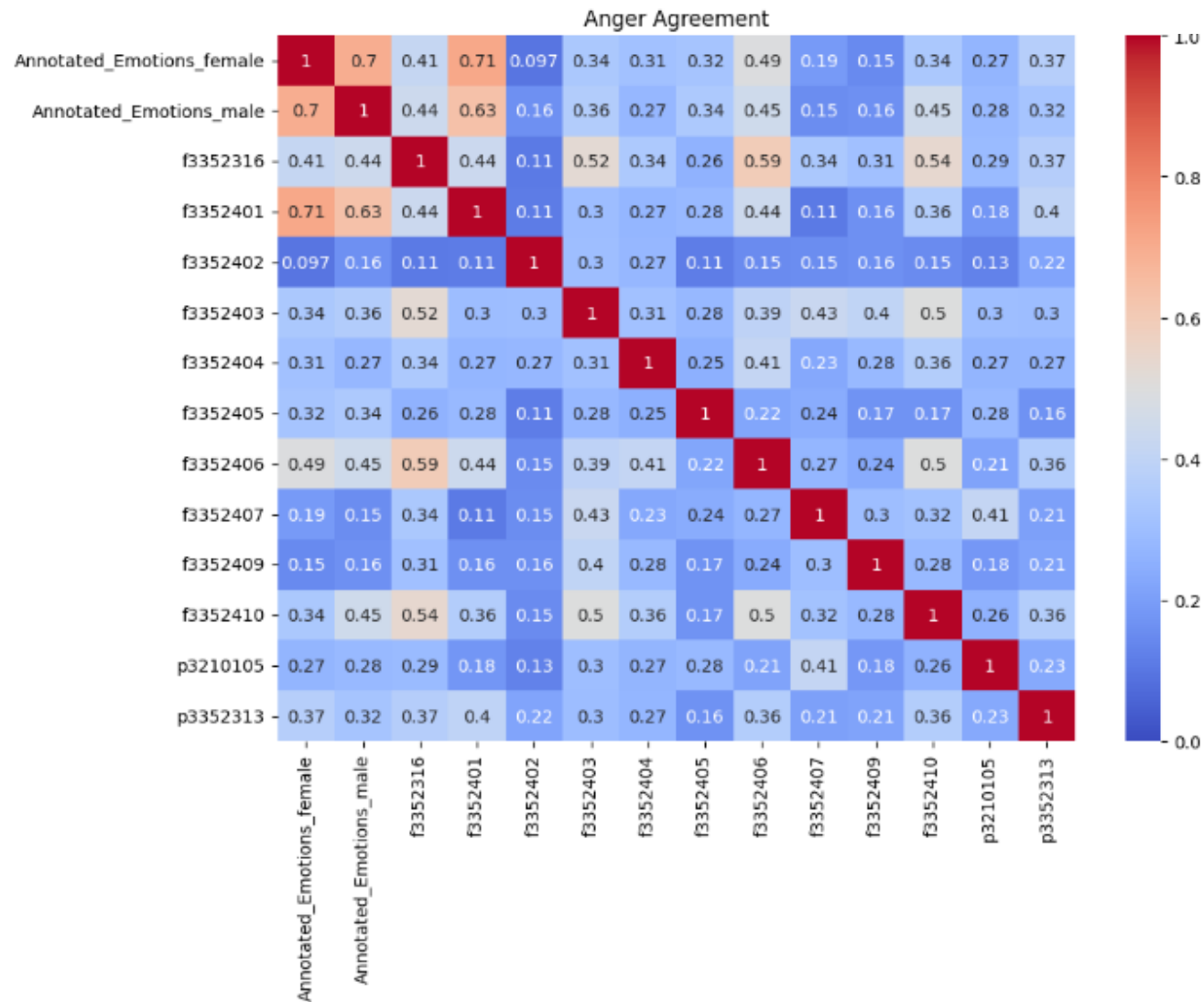
*Variability may suggest a need for clearer guidelines or training for annotators with low scores.*

### ■ Overall Metrics:

Average Kappa Score: 0.37 (moderate agreement)

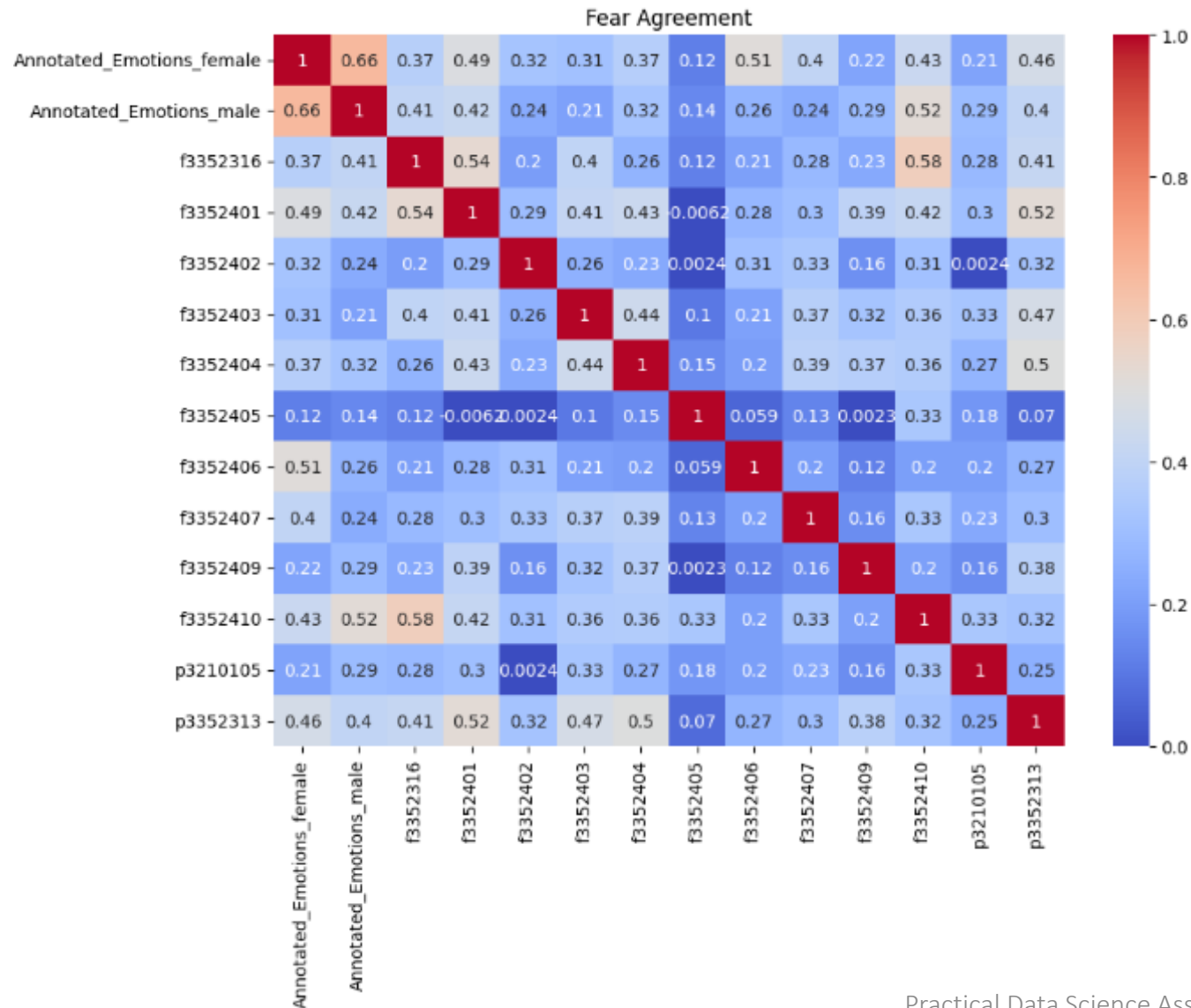
- Per Emotion Inter Annotation Agreement between Agents and Human Annotators

Emotion	Average Kappa	Average Percentage Agreement
Anger	0.35	0.80
Fear	0.34	0.76
Joy	0.49	0.85
Sadness	0.37	0.71
Surprise	0.29	0.70



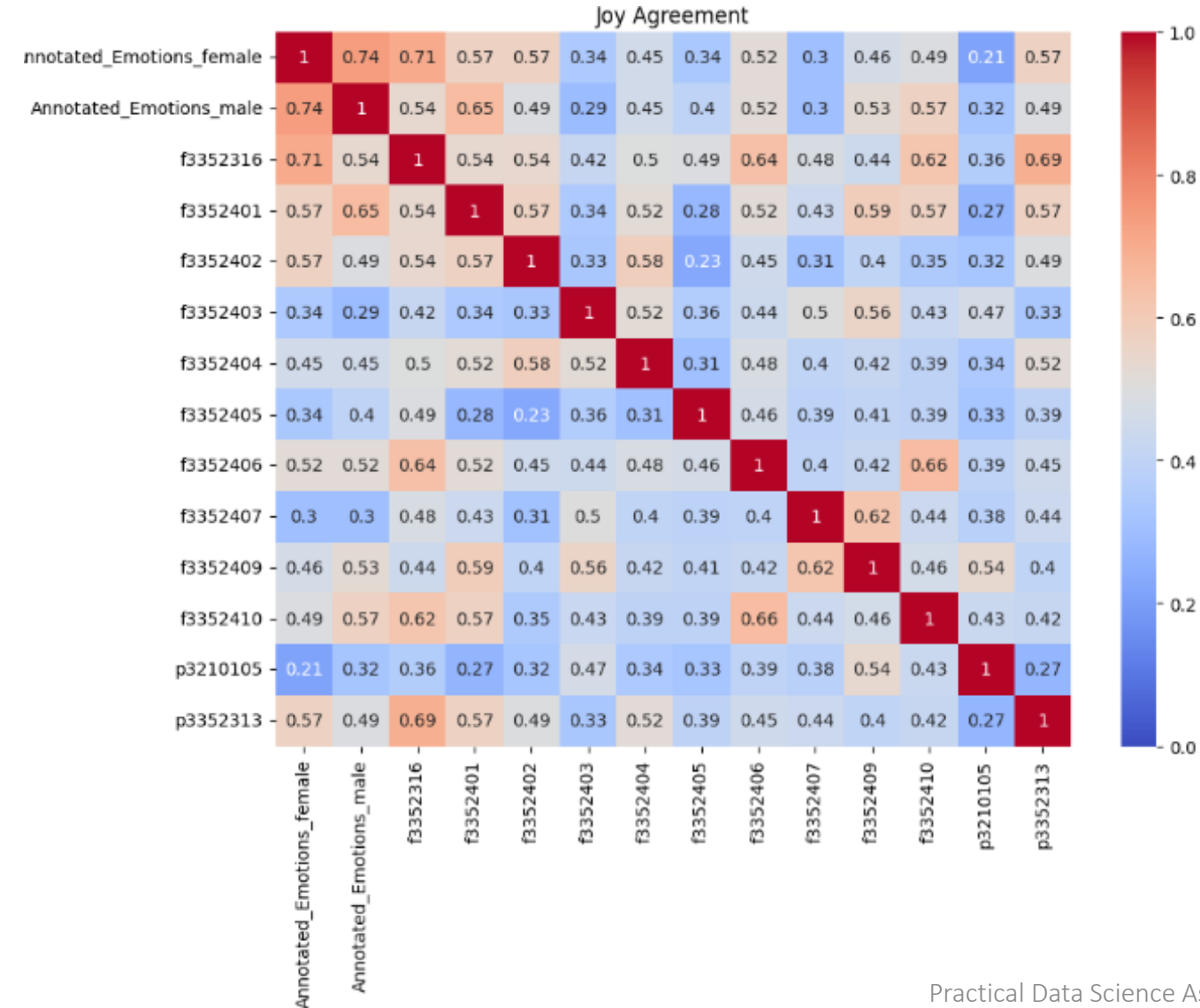
### Anger:

- Lowest Kappa Scores:  
f3352407 vs Male Annotator= 0.15,  
f3352402 vs Female Annotator= 0.097
- Indicates low alignment with gendered annotations in Anger.



### Fear:

- Notable Low Scores:  
f3352405 vs Male Annotator: 0.14,  
f3352405 vs Female Annotator: 0.12
- Suggests greater disagreement with gendered annotations in identifying Fear.



### Joy:

- Annotators align well

Female Annotator vs Male Annotator: 0.74

- Inconsistent pairs

Female annotator vs p3210105: 0.21



- *"Kappa scores highlight a trend of lower agreement in Anger and Fear, with Joy showing relatively stronger alignment across annotators."*
- *"These findings suggest potential discrepancies in interpreting Anger and Fear that may benefit from focused guideline improvement."*

- Male and Female Annotators: Show strong alignment in emotional interpretation, especially for Joy (highest correlation).
- Lower Correlations for Fear and Anger: Indicates possible sensitivity differences; annotators may have higher thresholds for detecting negative emotions.

A thin vertical black line is positioned on the left side of the slide, extending from the top to the bottom.

# Section B

## “Greek Proverbs Analysis”

# | 1. Basic

# Loading Dataset and redundant text reporting

	Unnamed: 0	greek_proverb	place	uri	collector	area	lat	lon	english_proverb
0	105697	Γέλια σαν κομπολόγια	Ήπειρος, Ζαγόρι, Βίτσα	<a href="http://hdl.handle.net/20.500.11853/168435">http://hdl.handle.net/20.500.11853/168435</a>	Σάρρος, Δημήτριος Μ.	Ήπειρος	37.998253	23.737867	Laughter is like a rosary. (A rosary is a string
1	8413	Καρδίαν καθαρὰν θέλ' ο Θεός	Ήπειρος	<a href="http://hdl.handle.net/20.500.11853/167032">http://hdl.handle.net/20.500.11853/167032</a>	Γόνιος, Α.	Ήπειρος	37.998253	23.737867	God asks for a clean heart. (Proverb) < eot_id
2	7684	Ου Θεός κι ου γείτονας	Ήπειρος	<a href="http://hdl.handle.net/20.500.11853/168991">http://hdl.handle.net/20.500.11853/168991</a>	Γαλδέμης, Αναστάσιος Δ.	Ήπειρος	37.998253	23.737867	No God and no neighbor. (meaning that when you...
3	18546	Θέλει να κρυφθή πίσω από το δάχτυλό του	Ήπειρος	<a href="http://hdl.handle.net/20.500.11853/273352">http://hdl.handle.net/20.500.11853/273352</a>	Ζηκίδης, Γεώργιος Δ.	Ήπειρος	37.998253	23.737867	He wants to hide behind his finger. (Proverb: ...
4	94001	Όλοι κλαίν' τα χάλια τ'ς κι ο μυλωνάς τη δέσι	Ήπειρος	<a href="http://hdl.handle.net/20.500.11853/204755">http://hdl.handle.net/20.500.11853/204755</a>	Παπαγεωργίου, Ιωάννης	Ήπειρος	37.998253	23.737867	All cry over the mess, but the miller is the o...

Using Regex pattern: <<r'r'\s\*(.?)\s|\s\*<.??>\s|\s\*(.|\s<.\*|\sW" >>

	greek_proverb	english_proverb	redundant_text
0	Γέλια σαν κομπολόγια	Laughter is like a rosary. (A rosary is a string)	[ , , , , , , (A rosary is a string)]
1	Καρδίαν καθαράν θέλ' ο Θεός	God asks for a clean heart. (Proverb) < eot_id	[ , , , , , , (Proverb) , < eot_id]
2	Ου Θεός κι ου γείτονας	No God and no neighbor. (meaning that when you...	[ , , , , , , (meaning that when you are in...
3	Θέλει να κρυφθή πίσω από το δάχτυλό του	He wants to hide behind his finger. (Proverb: ...	[ , , , , , , , (Proverb: To hide behin...
4	'Όλοι κλαίν' τα χάλια τ'ς κι ο μυλωνάς τη δέσι	All cry over the mess, but the miller is the o...	[ , , , , , , , , , ]
...	...	...	...
11495	Παστρζικό τσανα τσ' ένα!	A snake in the grass is not seen! or A snake i...	[ , , , , , , ! , , , , , ]
11496	Κάλλιο γεναίκα κάμισσα, πέρζι πολυπρικούσα	It is better to have a small profit than a lar...	[ , , , , , , , , , ]
11497	'Όγοιος κάθετα στη στερζά τσαί θάλασσα γερεύει...	The devil of one's backside is always quicker ...	[ , , , ' , , , , , , ]
11498	Πήε στ' δαιμόνου τ' μάννα	Take the devil's mother. (Take the devil's mot...	[ , , , ' , , , (Take the devil's mother to y...
11499	Θα σε κάνου να κατουρζήσεις αίμα	You will make someone drink blood. (Translatio...	[ , , , , , , , (Translation of the Greek ..

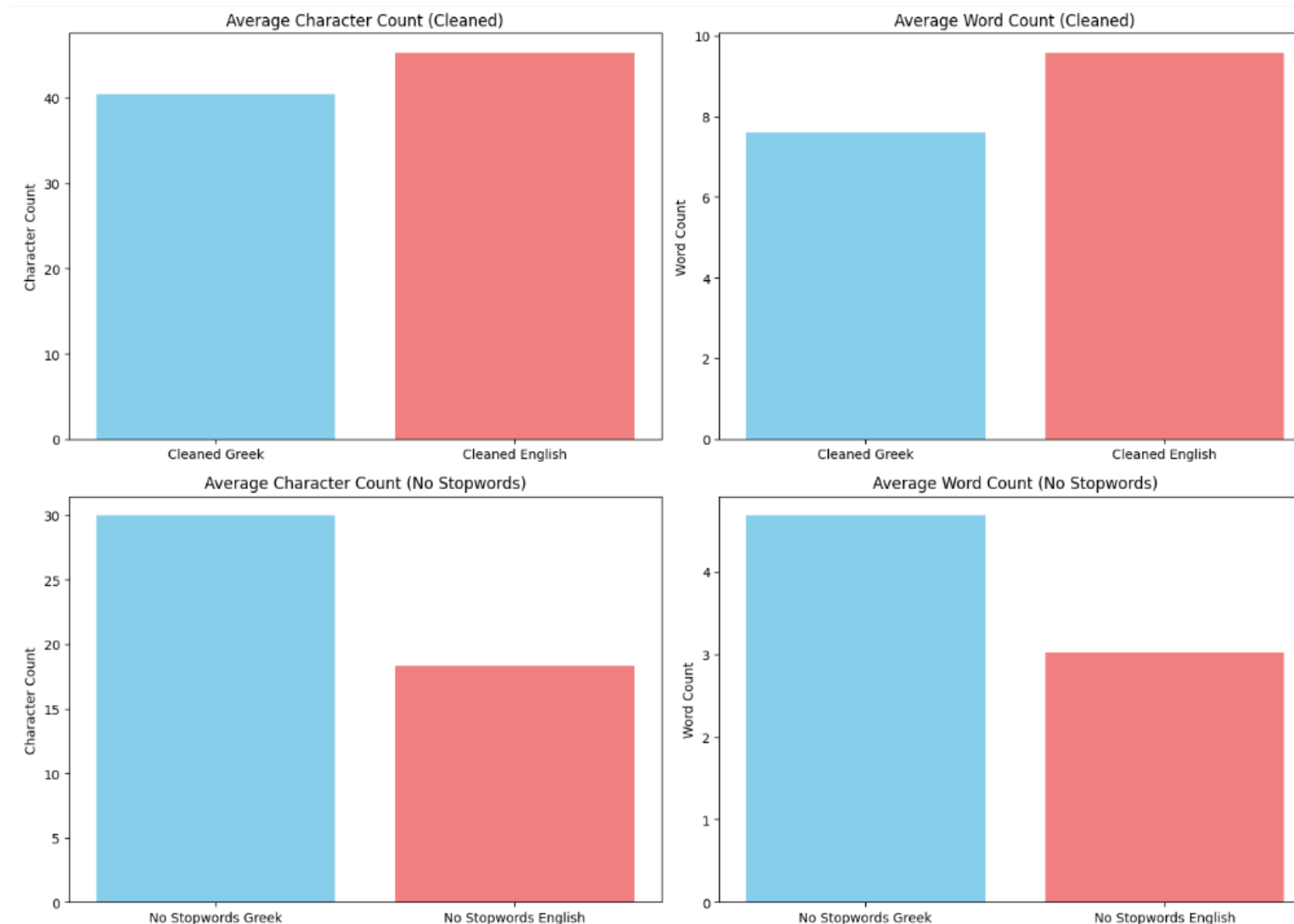
## | 2.Preprocessing

# Processed Dataframe

	<b>greek_proverb</b>	<b>english_proverb</b>	<b>cleaned_english_proverb</b>	<b>english_proverb_no_stopwords</b>	<b>cleaned_greek_proverb</b>	<b>greek_proverb_no_stopwords</b>
<b>0</b>	Γέλια σαν κομπολόγια	Laughter is like a rosary. (A rosary is a string	Laughter is like a rosary	Laughter rosary	Γέλια σαν κομπολόγια	Γέλια κομπολόγια
<b>1</b>	Καρδίαν καθαράν θέλ' ο Θεός	God asks for a clean heart. (Proverb) < eot_id	God asks for a clean heart	God clean heart	Καρδίαν καθαράν θέλ ο Θεός	Καρδίαν καθαράν θέλ Θεός
<b>2</b>	Ου Θεός κι ου γείτονας	No God and no neighbor. (meaning that when you...	No God and no neighbor	God neighbor	Ου Θεός κι ου γείτονας	Θεός γείτονας
<b>3</b>	Θέλει να κρυφθή πίσω από το δάχτυλό του	He wants to hide behind his finger. (Proverb: ...	He wants to hide behind his finger	hide finger	Θέλει να κρυφθή πίσω από το δάχτυλό του	Θέλει κρυφθή πίσω δάχτυλό
<b>4</b>	Όλοι κλαίν' τα χάλια τ'ς κι ο μυλωνάς τη δέσι	All cry over the mess, but the miller is the o...	All cry over the mess but the miller is the o...	mess miller	Όλοι κλαίν τα χάλια τ ς κι ο μυλωνάς τη δέσι	Όλοι κλαίν χάλια μυλωνάς δέσι
...	...	...	...	...	...	...



# | 3.Exploration



## Average Character Count

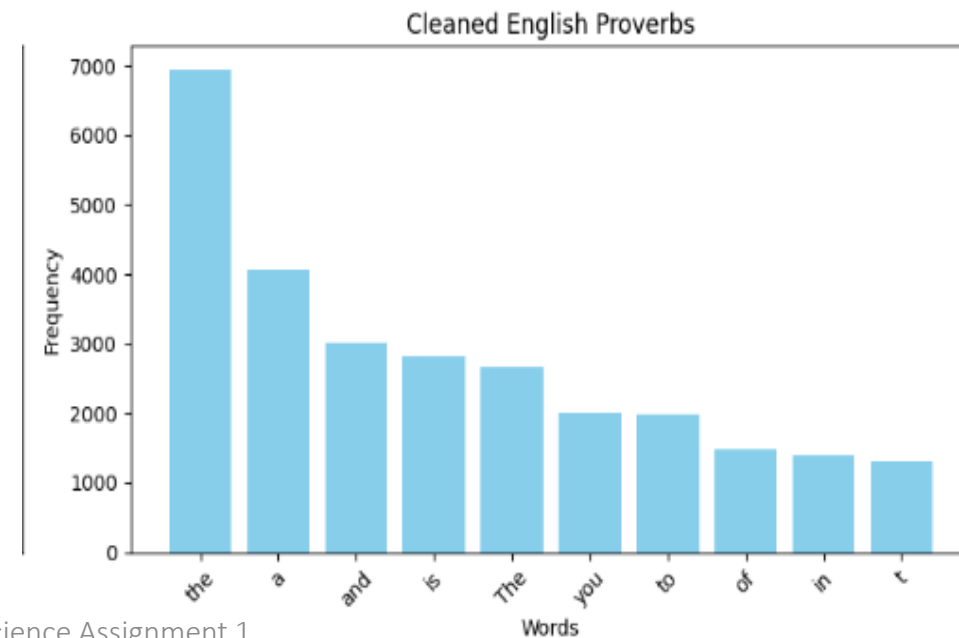
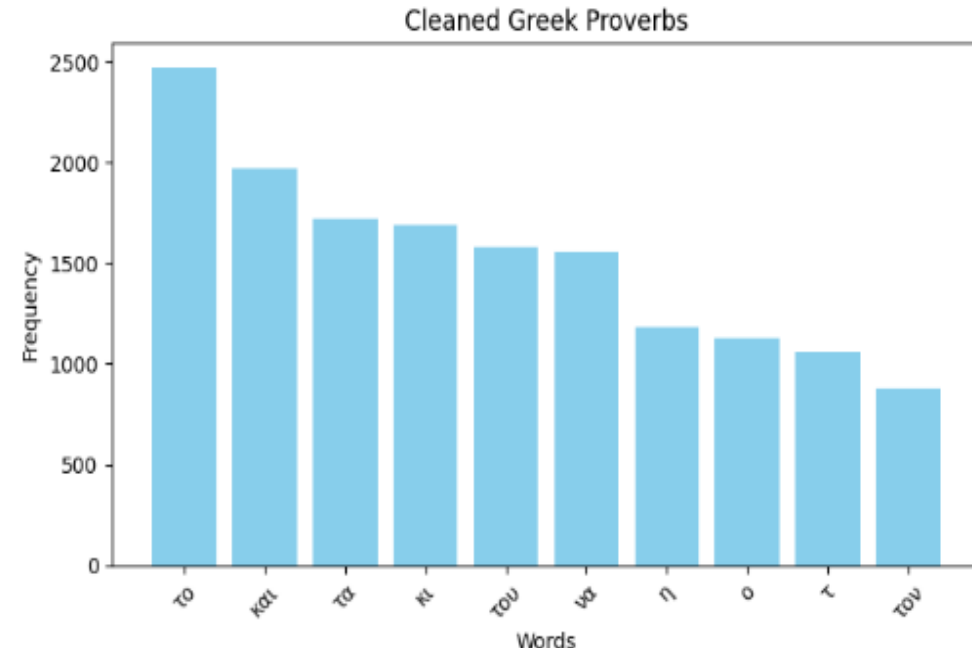
### •Greek:

- Cleaned text: **42.37** characters
- No-stopwords text: **29.96** characters
- Reduction due to stopwords: **10.42** characters (25.8%)

### •English:

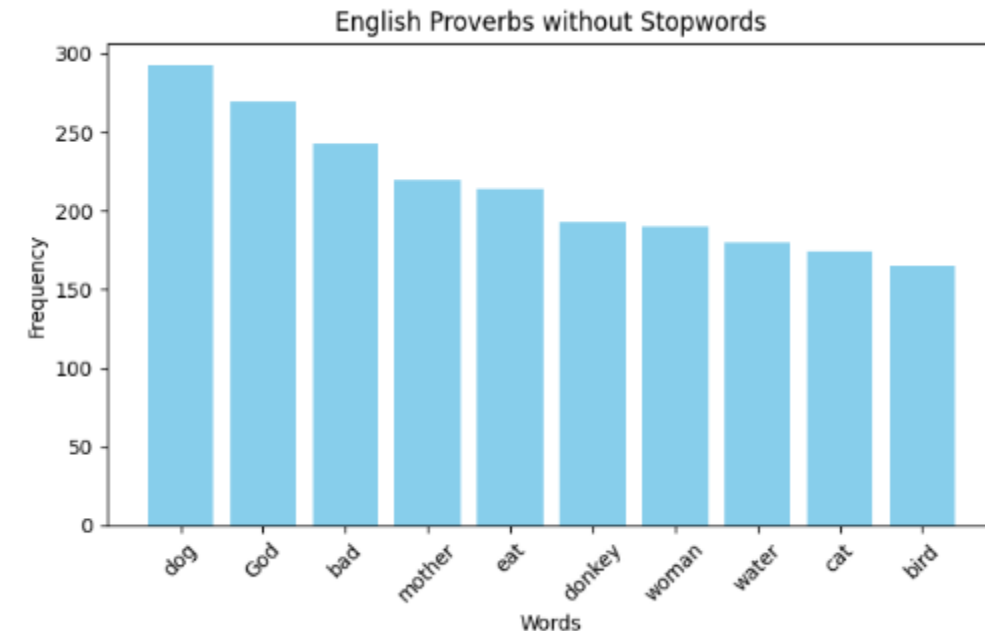
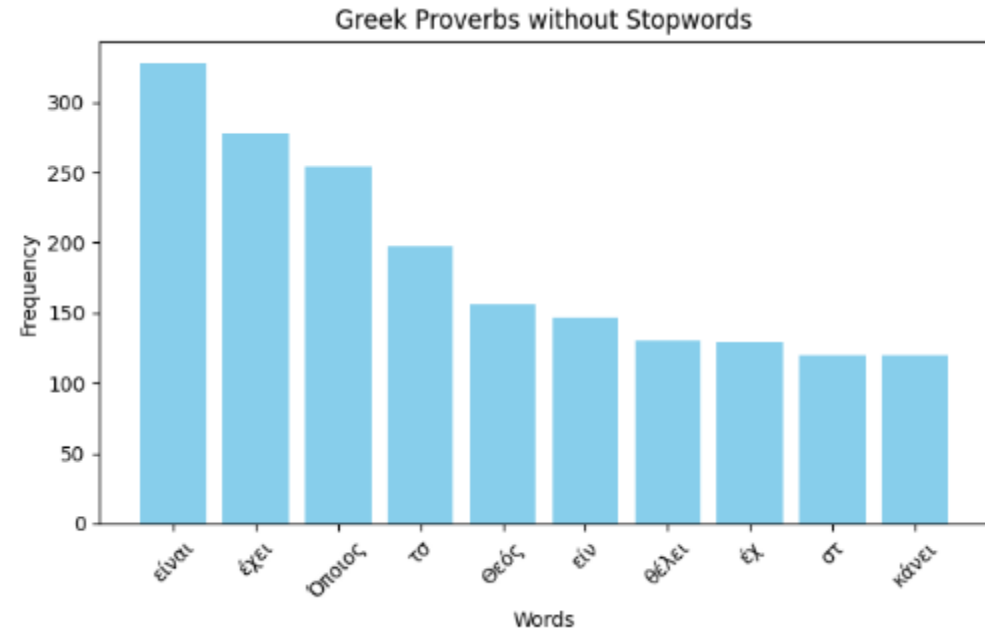
- Cleaned text: **45.28** characters
- No-stopwords text: **18.32** characters
- Reduction due to stopwords: **26.96** characters (59.6%)

- Average Word Count
- Greek:
  - Cleaned text: **7.58** words
  - No-stopwords text: **4.68** words
  - Reduction due to stopwords: **2.9** words (**38.3%**)
- English:
  - Cleaned text: **9.58** words
  - No-stopwords text: **3.02** words
  - Reduction due to stopwords: **6.56** words (**68.5%**)



## Analysis of Most Common Words

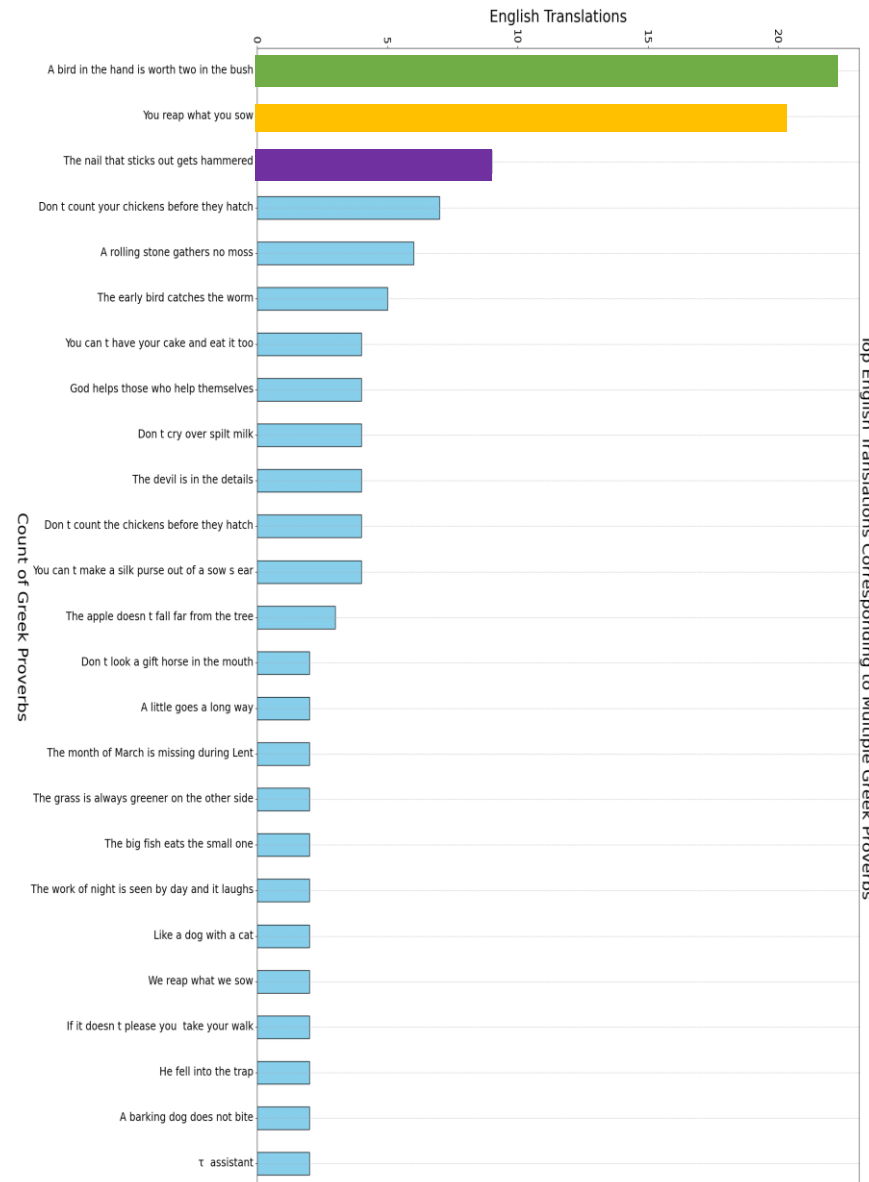
- **Greek Proverbs:**
  - Common stopwords: "το" (the), "και" (and)
  - Key words post-removal: "Θεός" (God), "μάτια" (eyes), "καλό" (good), "κακό" (bad), "Οποιος" (whoever)
- **English Proverbs:**
  - Common stopwords: "the," "a," "and," "is"
  - Key words post-removal: "dog," "cat," "wolf," "water," "mother," "bad," "God"



## | 4. Visualisation




# | 5. Normalization



“A bird in the hand is worth two in the bush”  
 “You reap what you sow”  
 “The nail that sticks out gets hammered”





# Q&A

A thin vertical black line is positioned on the left side of the slide.

# Thank You!