Strength of Hadith Classifier

- Data Science Project -

{ إِنْ قَامَتِ السَّاعَةُ وَفِي يَدِ أَحَدِكُمْ فَسِيلَةٌ، فَلْيَغْرِسْهَا}

-لا يمنعكم قرب قيام الساعة من السعي في الأرض وعمارتها-

Waad Alhejali 4110091
Maryam Aqel 4111815

INTODUCTION

The "Strength of Hadith Classifier" project aims to develop a reliable system for classifying Hadith as either one of the four classes authentic, good and weak and non-authentic. Hadith, the sayings and actions of the Prophet Muhammad (peace be upon him), hold great significance in the Islamic tradition. However, verifying the authenticity of Hadith can be a challenging and time-consuming task, often requiring extensive expertise.

To address this challenge, our project follows a systematic approach. We began by gathering a substantial amount of Hadith data through web scraping. This data was then manually labeled, distinguishing the 4 classes. We then performed preprocessing and visualization techniques on the labeled data to gain insights and identify patterns.

Data transformation techniques were applied to prepare the dataset for subsequent modeling stages. This involved feature engineering and selection processes to optimize the performance of the classifier. We also explored multiple machine learning models, evaluating their performance metrics to identify the most effective one.

Our objective is to provide a reliable tool for scholars, researchers, and the general public to assess the strength of Hadith. By automating the classification process, we aim to simplify the task of identifying authentic Hadith, thereby facilitating accurate information and references.

In this report, we will provide a detailed documentation of our project, including the data collection process, manual labeling, preprocessing techniques, data transformation methods, and the evaluation of machine learning models. We will discuss the results obtained, along with insights into the strengths and limitations of different approaches.

By developing an effective Hadith classifier, we hope to contribute to the field of Islamic studies, promote research, and foster a deeper understanding of the Islamic tradition and may allah make it in the balance of our deeds and everyone who reads and educates themselves on hadiths,

In the following sections we will discuss the process of our whole project and what it went through step by step.


298 - مَثَلُ الْمُنْفِقِ والْمُتَصَدِّقِ، كَمَثَلِ رَجُلٍ عليه جُبَّتَانِ، أوْ جُنَّتَانِ، مِن لَدُنْ ثُدِيِّهِما إلى تَرَاقِيهِما، فإذا ...

الراوي : أبو هريرة | المحدث : مسلم | المصدر : صحيح مسلم

الصفحة أو الرقم : 1021 | خلاصة حكم المحدث : [صحيح] | أحاديث مشابهة | شرح حديث مشابه

Figure 1: The data we want to scrap from the website

1. WEB SCRAPING

We used BeautifulSoup which is a Python library that provides tools for web scraping HTML and XML files. It creates a parse tree from a page source code that can be used to extract data in a hierarchical and more readable manner. Our data was scraped from ALDORAR ALSANIYYAH website which is a website for various alshareia sciences: (tafsir, hadith, aqidah, fiqh, and other sciences). In order to scrap the data from the website we build a method that takes desired URL and pattern which is the structure of data, then we used it to scrap 20 pages each time we change the URL and the pattern in order to scrap all the data we want which is the Hadith that has pattern 'h5', the sources, narrators, and muhaddiths which has 'primary-text-color' pattern

WORK PROCESS

First, we built the method that scraps the hadiths and features in organized matter. Then, we divided the work among ourselves, with each person scraping 10 pages, resulting in a total of 300 rows. We ensured that the data was correct and organized as it appeared on the webpage. Afterwards, we manually performed the labeling and concatenated the two datasets together, resulting in a total of 600 rows.

	Hadith	Narrator	Speaker	Source	Labels
30	...كُلَّ يَوْمٍ هُوَ فِي شَأْنٍ} قال : من شأن {	أبو الدرداء	ابن حجر العسقلاني	تغليق التعليق	فيه اضطراب وله شاهد بإسناد ضعيف
31	أبغضُ الحلال إلى الله الطَّلاقُ	عبدالله بن عمر	الألباني	التعليقات الرضية	صحيح
32	...أبو بكرٍ الصِّدِّيقُ تاجُ الإسلامِ وعُمرُ بن	عبدالله بن مسعود	ابن حجر العسقلاني	تسديد القوس	صحيح
33	...أنَّ أبيَّ بن كعبٍ أمَّهم - يَعني - في رمضان	بعض أصحاب محمد	الألباني	ضعيف أبي داود	ضعيف
34	... ألَّي [يعني سعرَ بنَ دينم] كنتُ في شِعْبٍ	مصدقا رسول الله	الألباني	ضعيف أبي داود	ضعيف
...	...	...	...	...	...
325	...إنَّ النَّبيَّ صلَّى اللَّهُ عليهِ وآلِه وسلَّ	عبدالله بن عمر	الشوكاني	الفوائد المجموعة	في إسناده كذاب
326	...أوَّلُ حُبِّ في الإسلامِ، حُبُّ النَّبيِّ صلَ	أنس بن مالك	الشوكاني	الفوائد المجموعة	في إسناده كذابان
327	...خيرُ أمَّتي الَّذي إذا أساءوا استغفروا ، وإ	جابر بن عبدالله	ابن حجر العسقلاني	بلوغ المرام	إسناده ضعيف
328	...لا تُسكِّنوهنَّ الغرف، ولا تعلِّموهنَّ الكتاب	عائشة أم المؤمنين	الشوكاني	الفوائد المجموعة	في إسناده محمد بن إبراهيم الشامي كان يضع الحديث
329	...سلوا اللَّه من فضلِه ، فإنَّ اللَّه يحبُّ	عبدالله بن مسعود	السخاوي	المقاصد الحسنة	« طرق

300 rows × 5 columns

Data 2

	Hadith	Narrator	Source	Speaker	Page Number	Labels
0	...أنَّ أبيَّ بن كعبٍ أمَّهم - يَعني - في رمضان	بعض أصحاب محمد	ضعيف أبي داود	الألباني	1428	ضعيف
1	...ألَّي [يعني سعرَ بنَ دينم] كنتُ في شِعْبٍ م	مصدقا رسول الله	ضعيف أبي داود	الألباني	1581	ضعيف
2	... عن أبي حمزةَ الثُّماليِّ قال : يعني بالذَّاب	أبو حمزة الثمالي	العجاب في بيان الأسباب	ابن حجر العسقلاني	2/889	إسناده ضعيف
3	...أخَذَ ابْنُ مسعودٍ قَوْمًا ازْتَدُّوا عَنِ ال	عبدالله بن عتبة	الصارم المسلول	ابن تيمية	3/605	إسناده صحيح
4	...إذا مات العبدُ والله عزَّ وجلَّ يعلم منه شرَّ	عامر بن ربيعة	تسديد القوس	ابن حجر العسقلاني	1/348	إسناده ضعيف
...	...	...	...	...	...	...
295	...مَثَلُ البَخِيلِ والمُتَصَدِّقِ مَثَلُ رَجُلٍ	أبو هريرة	صحيح مسلم	مسلم	177/2	صحيح
296	...ضَرَبَ رَسولُ اللهِ صَلَّى اللَّهُ عليه وسلَّم	أبو هريرة	صحيح مسلم	مسلم	3/465	صحيح
297	...مَثَلُ المُنْفِقِ والمُتَصَدِّقِ، كَمَثَلِ رَ	أبو هريرة	صحيح مسلم	مسلم	2/48	صحيح
298	...قالَ رَجُلٌ لَأَتَصَدَّقَنَّ اللَّيْلَةَ بصَدَقَ	أبو هريرة	صحيح مسلم	مسلم	1/428	صحيح
299	...إنَّ الخازِنَ المُسْلِمَ الأمِينَ الذي يُنْفِذ	أبو موسى الأشعري	صحيح مسلم	مسلم	2/228	صحيح

300 rows × 6 columns

Data 1

	Hadith	Narrator	Source	Speaker	Page Number	Labels
0	...أنَّ أبيَّ بن كعبٍ أمَّهم - يَعني - في رمضان	بعض أصحاب محمد	ضعيف أبي داود	الألباني	1428	ضعيف
1	...ألَّي [يعني سعرَ بنَ دينم] كنتُ في شِعْبٍ م	مصدقا رسول الله	ضعيف أبي داود	الألباني	1581	ضعيف
2	... عن أبي حمزةَ الثُّماليِّ قال : يعني بالذَّاب	أبو حمزة الثمالي	العجاب في بيان الأسباب	ابن حجر العسقلاني	2/889	إسناده ضعيف
3	...أخَذَ ابْنُ مسعودٍ قَوْمًا ازْتَدُّوا عَنِ ال	عبدالله بن عتبة	الصارم المسلول	ابن تيمية	3/605	إسناده صحيح
4	...إذا مات العبدُ والله عزَّ وجلَّ يعلم منه شرَّ	عامر بن ربيعة	تسديد القوس	ابن حجر العسقلاني	1/348	إسناده ضعيف
...	...	...	...	...	...	...
595	...إنَّ النَّبيَّ صلَّى اللَّهُ عليهِ وآلِه وسلَّ	عبدالله بن عمر	الفوائد المجموعة	الشوكاني	NaN	في إسناده كذاب
596	...أوَّلُ حُبِّ في الإسلامِ، حُبُّ النَّبيِّ صلَ	أنس بن مالك	الفوائد المجموعة	الشوكاني	NaN	في إسناده كذابان
597	...خيرُ أمَّتي الَّذي إذا أساءوا استغفروا ، وإ	جابر بن عبدالله	بلوغ المرام	ابن حجر العسقلاني	NaN	إسناده ضعيف
598	...لا تُسكِّنوهنَّ الغرف، ولا تعلِّموهنَّ الكتاب	عائشة أم المؤمنين	الفوائد المجموعة	الشوكاني	NaN	في إسناده محمد بن إبراهيم الشامي كان يضع الحديث
599	...سلوا اللَّه من فضلِه ، فإنَّ اللَّه يحبُّ	عبدالله بن مسعود	المقاصد الحسنة	السخاوي	NaN	« طرق

600 rows × 6 columns

The concatenated data

	Hadith	Narrator	Source	Speaker	Labels
0	... أن أبي بن كعب أسهم يعني في رمضان وكان يقنت في	بعض أصحاب محمد	ضعيف أبي داود	الألباني	1
1	... أني يعني سعر بن ديسم كنت في شعب من هذه الشعاب	مصدقا رسول الله	ضعيف أبي داود	الألباني	1
2	...عن أبي حمزة الثمالي قال يعني بالناس في هذه الآية	أبو حمزة الثمالي	العجاب في بيان الأسباب	ابن حجر العسقلاني	1
3	...أخذ ابن مسعود قوما ارتدوا عن الإسلام من أهل العر	عبدالله بن عتبة	الصارم المسلول	ابن تيمية	3
4	...إذا مات العبد والله عز وجل يعلم منه شرا وقال الثا	عامر بن ربيعة	تسديد القوس	ابن حجر العسقلاني	1
...	...	...	...	...	...
501	... يا أيها الناس من عمل منكم لنا على عمل فكتمنا	عدي بن عميرة الكندي	صحيح أبي داود	الألباني	3
502	... بعثني رسول الله صلى الله عليه وسلم إلى اليمن قاضيا	علي بن أبي طالب	صحيح أبي داود	الألباني	2
503	...أن النبي صلى الله عليه وآله وسلم اجتلى عائشة عند	عبدالله بن عمر	الفوائد المجموعة	الشوكاني	1
504	...أول حب في الإسلام حب النبي صلى الله عليه وآله وسلم	أنس بن مالك	الفوائد المجموعة	الشوكاني	1
505	...خير أمتي الذي إذا أساءوا استغفروا وإذا سافروا	جابر بن عبدالله	بلوغ المرام	ابن حجر العسقلاني	1

506 rows × 5 columns

Final Data

2. DATA DESCRIBTION

We started by Dataset1 (300 rows) and concatenate it with Dataset2 (300 rows) when we scraped the features to get the concatenated Dataset with 600 rows. then we started Doing Data preprocessing and transformations such that will be shown in the following sections and lastly we end up with this final dataset. we scraped and collected important features to classify if the Hadith is authentic or not, which are the:

- Hadith: content of hadith itself
- Narrator: the person who narrated the hadith like: عائشة رضي الله عنها، أبو هريرة
- Speaker: religious cholars that studies and investigates Hadiths and determines their correctness and authenitcity or vulnerability.
- Source: Source of the specific Hadith
- Label of the Hadith which is one of four classes: صحيح، حسن ، ضعيف، لا يصح

Then we transformed and mapped these labels to numeric values from 0-3 as shown in the final dataset. Moreover, we scraped the

- Page number feature: which is the page number of the source of the Hadith as well but it was dropped for model performance reasons at the end.

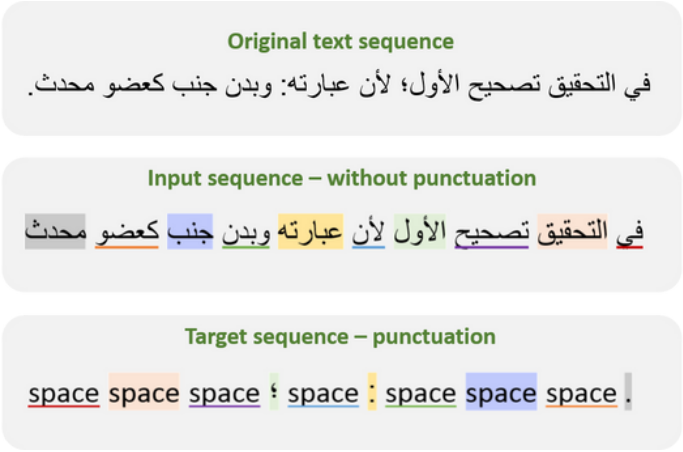


Figure 2: Methods for data preprocessing performance

3. DATA PREPROCESSING

For data preprocessing, we developed a method to remove numbering, commas, and dots from the text. This step was crucial to ensure that the data is clean and consistent. Additionally, we implemented another method to remove diacritical marks, which are commonly used in Arabic text to indicate vowel sounds. By removing diacritical marks, we aimed to simplify the text for further analysis.

Furthermore, we utilized a punctuation remover tool to eliminate any non-Arabic punctuation marks. This step helped us maintain the integrity of the Arabic text while removing unnecessary symbols that could potentially affect the classification process.

These preprocessing techniques were essential in preparing the data for subsequent stages, such as feature extraction and model training. By eliminating unwanted characters and symbols, we aimed to enhance the quality and accuracy of the data, which ultimately contributes to the effectiveness of the Hadith classifier.

The original Hadith

```
'1] - أنَّ أبيَّ بنَ كعبٍ أمْهم - يَعني - في رمضانَ وَكانَ يقنُتُ في النِّصفِ الآخِرِ ...',
'2' - أنِّي [ يعني سعرَ بنَ ديسم ] كنتُ في شِعْبٍ من هذه الشِّعابِ على عهدِ ...',
'3' - عن أبي حمزةَ الثُّماليِّ قال : يعني بالنَّاسِ في هذه الآيةِ نبيُّ اللهِ صلَّى اللهُ ...',
```

After removing prefix and suffix

```
أنَّ أبيَّ بنَ كعبٍ أمْهم - يَعني - في رمضانَ وَكانَ يقنُتُ في النِّصفِ الآخِرِ
أنِّي [ يعني سعرَ بنَ ديسم ] كنتُ في شِعْبٍ من هذه الشِّعابِ على عهدِ
عن أبي حمزةَ الثُّماليِّ قال : يعني بالنَّاسِ في هذه الآيةِ نبيُّ اللهِ صلَّى اللهُ
```

After removing punctuation and harakat

```
أن أبي بن كعب أمهم يعني في رمضان وكان يقنت في النصف الآخر
أني يعني سعر بن ديسم كنت في شعب من هذه الشعاب على عهد
عن أبي حمزة الثمالي قال يعني بالناس في هذه الآية نبي الله صلى الله
```

```python
#Method to preprocess Hadiths and remove unnecessary prefix and suffix
def preprocess_hadith(hadith):
    # Remove numbering and hyphen, then strip leading/trailing spaces
    cleaned_hadith = re.sub(r'^\d+\s*-\s*', '', hadith).strip()
    # Remove commas
    cleaned_hadith = cleaned_hadith.replace(',', '')
    # Remove ellipsis
    cleaned_hadith = cleaned_hadith.replace('...', '')
    return cleaned_hadith
```

```python
#Method to remove all non arabic letters from hadiths wether its the numbering, pu
def punctuation_remover(hadith):
 # Remove punctuation and symbols
    translator = str.maketrans("", "", string.punctuation + string.digits)
    hadith = hadith.translate(translator)

    # Remove non-Arabic letters
    hadith = ''.join(char for char in hadith if char.isalpha() or char.isspace())
    #remove wrong double spaces into one
    hadith = re.sub(r'\s+', ' ', hadith)

    return hadith.upper()
```

```python
#Method that removes arabic harakat such as
def harakat_remover(text):
    reshaped_text = arabic_reshaper.reshape(text)
    harakat_pattern = re.compile(r'[\u064b-\u0652]+')
    text_without_harakat = re.sub(harakat_pattern, '', reshaped_text)
    return text_without_harakat
```

Figure 3: Methods for datapreprocessing

After concateninting the dataset and having 600 row we manually label our target variables but with multiple different classes which we then need to preprocess.

We firstly Drop the page number column since we it does not add any information to our classifier.

```
df_og
```

	Hadith	Narrator	Source	Speaker	Labels
0	...أنّ أبيّ بنَ كعب أتمهم - يعني - في رمضان	بعض أصحاب محمد	ضعيف أبي داود	الألباني	ضعيف
1	...أني [يعني سعر بن ديسم] كنتُ في شعبٍ م	مصدقا رسول الله	ضعيف أبي داود	الألباني	ضعيف
2	... عن أبي حمزة الثُّماليّ قال : يعني بالنّاس	أبو حمزة الثمالي	العجاب في بيان الأسباب	ابن حجر العسقلاني	إسناده ضعيف
3	...أخذ ابنُ مسعودٍ قومًا ارتدُّوا عن ال	عبدالله بن عتبة	الصارم المسلول	ابن تيمية	إسناده صحيح
4	...إذا مات العبدُ واللهُ عزّ وجلّ يعلم منه شرًّ	عامر بن ربيعة	تسديد القوس	ابن حجر العسقلاني	إسناده ضعيف
...	...	...	...	...	...
595	...أنّ النّبيّ صلّى الله عليه وآله وسلّ	عبدالله بن عمر	الفوائد المجموعة	الشوكاني	في إسناده كذاب
596	...أوّلُ خُبثٍ في الإسلامِ، حُبُّ النّبيّ صلّ	أنس بن مالك	الفوائد المجموعة	الشوكاني	في إسناده كذابان
597	...خيرُ أمّتي الّذي إذا أساءوا استغفروا ، وإ	جابر بن عبدالله	بلوغ المرام	ابن حجر العسقلاني	إسناده ضعيف
598	...لا تُسكِنُوهنّ الغرف، ولا تعلّموهنّ الكتاب	عائشة أم المؤمنين	الفوائد المجموعة	الشوكاني	في إسناده محمد بن إبراهيم الشامي كان يضع الحديث
599	...سلوا الله من فضلِه ، فإنّ الله يحبّ	عبدالله بن مسعود	المقاصد الحسنة	السخاوي	ه طرق

600 rows × 5 columns

When analyzing the labels we still have unique values so we replace some labels to have unified 4 class labels only

```
fdf['Labels'] = fdf['Labels'].replace('لا يصح' ,'إسناده غير صحيح')
fdf['Labels'] = fdf['Labels'].replace('حسن' ,'إسناده الحسن')
```

```
<ipython-input-38-813fccee5d54>:1: SettingWithCopyWarning:
```

Lastly after acquiring only 4 unified labels on the whole dataset, we transform them by mapping into numerical values since its easier for classifier to learn the pattern

```
# Mapping of labels to numeric values for better performance when classi
label_mapping = {'0 :'لا يصح' ,1 :'ضعيف' ,2 :'حسن' ,3 :'صحيح'}

# Replace values in the 'Labels' column
fdf['Labels'] = fdf['Labels'].replace(label_mapping)
```

And like that we acquire our final dataset Described above

	Hadith	Narrator	Source	Speaker	Labels
0	... أن أبي بن كعب أمهم يعني في رمضان وكان يقنت في	بعض أصحاب محمد	ضعيف أبي داود	الألباني	1
1	... أني يعني سعر بن ديسم كنت في شعب من هذه الشعاب	مصدقا رسول الله	ضعيف أبي داود	الألباني	1
2	...عن أبي حمزة الثمالي قال يعني بالناس في هذه الآية	أبو حمزة الثمالي	العجاب في بيان الأسباب	ابن حجر العسقلاني	1
3	...أخذ ابن مسعود قوما ارتدوا عن الإسلام من أهل العر	عبدالله بن عتبة	الصارم المسلول	ابن تيمية	3
4	...إذا مات العبد والله عز وجل يعلم منه شرا وقال الثا	عامر بن ربيعة	تسديد القوس	ابن حجر العسقلاني	1
...	...	...	...	...	...
501	... يا أيها الناس من عمل منكم لنا على عمل فكتمنا	عدي بن عميرة الكندي	صحيح أبي داود	الألباني	3
502	... بعثني رسول الله صلى الله عليه وسلم إلى اليمن قاضيا	علي بن أبي طالب	صحيح أبي داود	الألباني	2
503	...أن النبي صلى الله عليه وآله وسلم اجتلى عائشة عند	عبدالله بن عمر	الفوائد المجموعة	الشوكاني	1
504	...أول حب في الإسلام حب النبي صلى الله عليه وآله وسلم	أنس بن مالك	الفوائد المجموعة	الشوكاني	1
505	...خير أمتي الذي إذا أساءوا استغفروا وإذا سافروا	جابر بن عبدالله	بلوغ المرام	ابن حجر العسقلاني	1

506 rows × 5 columns

Figure 4: Visualization of Frequent words in most Hadiths

4. DATA VISUALIZATION

We utilized various visualization techniques to gain insights into the collected data and understand its characteristics. These visualizations provided meaningful information about the frequency of words, occurrence of labels, distribution of Hadith speakers, and correlations between labels and sources.

1- Frequent words in the Hadith

This helped identify commonly occurring terms such as

"اللّٰه"
"الرسول صلى اللّٰه عليه وسلم"
"قال الرسول"
"عليه وسلم"
"رسول اللّٰه"
الجنة

The prevalence of these words aligns with the nature of Hadith and provides valuable insights into the content of the dataset

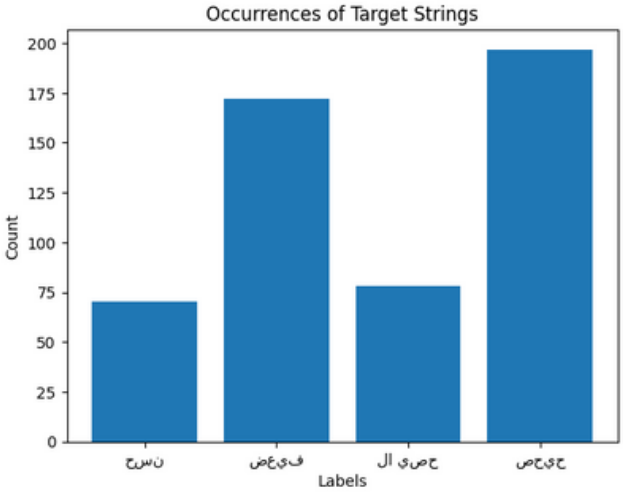


Figure 5: Visualization of label occurrences

2- Visualization of label occurrences

We also visualized the occurrence of each label, enabling us to understand the distribution of authentic and non-authentic Hadith in our dataset. This analysis allowed us to assess the balance of the dataset and gain a better understanding of the prevalence of different label categories.

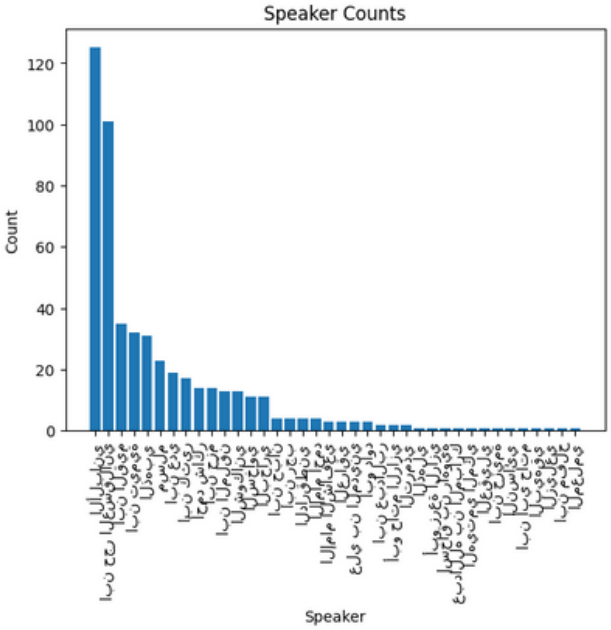
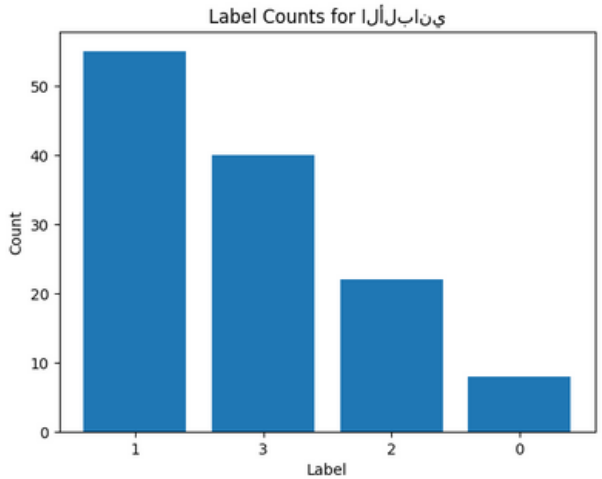


Figure 6: Visualization of hadith speaker

3- Frequency of each Hadith Speaker

we explored the frequency of different Hadith speakers through visualization. This analysis revealed was the most prominent "الألباني" that speaker in our dataset. We further delved into this specific speaker, visualizing the frequency of their Hadith labels. This provided insights into the distribution and characteristics of the Hadiths attributed to this particular speaker



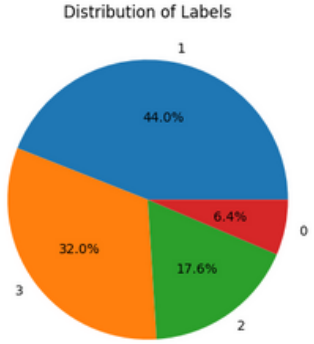


Figure 7: Two visualizations of Alalbani hadith classes

4- Frequent distribution of sources

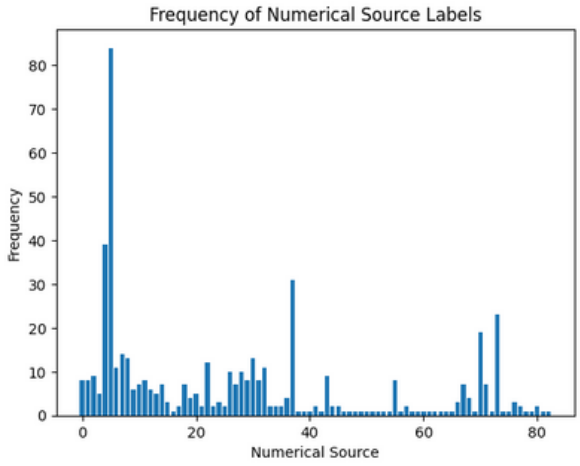


Figure 8: Visualization of source distributions

This visualization helped identify if specific sources were more frequent than others, shedding light on potential biases or variations in the data.

5- Correlation Heatmap between labels and sources

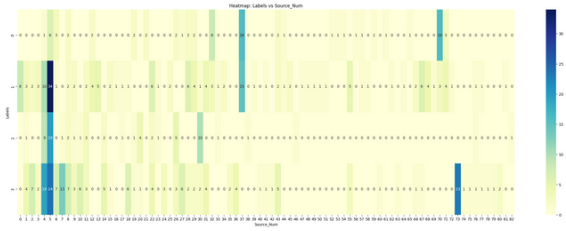
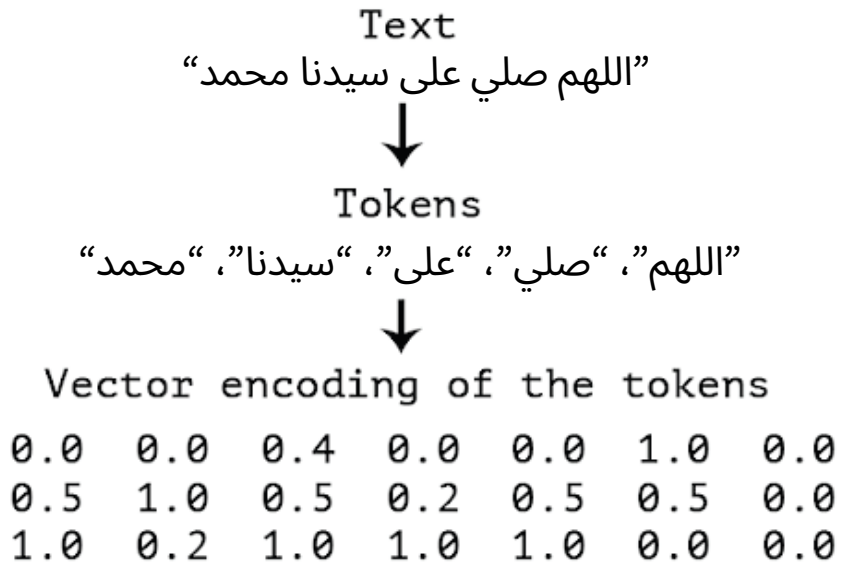


Figure 9: Visualization of source and label correlations

This visualization provided a comprehensive view of the interconnections between various elements in our dataset, allowing for a deeper understanding of patterns and associations.



5. DATA TRANSFORMATION

For the data transformation, we utilized a label encoder to convert the text values in the 'Speaker' and 'Source' columns into numerical representations. This was done by assigning unique numbers to each speaker. The transformed data was stored back into the 'Speaker' and 'Source' columns of the dataframe, and we checked the distribution of the transformed values using the `value_counts()` function.

To process the hadith data and prepare it for model input, we employed a tokenizer. The tokenizer was fitted on the text data (`X_text`) to learn the vocabulary and assign numerical indices to each unique word. Then, we converted the hadith texts into sequences of these numerical representations using the `texts_to_sequences()` function. Finally, we padded the sequences to ensure they have the same length using the `pad_sequences()` function. The resulting `X_padded` variable contains the transformed hadith data in a format suitable for model input.

```
# Create an instance of LabelEncoder
label_encoder = LabelEncoder()

# Convert the text values in 'Source' column to numbers
fdf['Source'] = label_encoder.fit_transform(fdf['Source'])

# Check the updated 'Source' column
fdf['Source'].value_counts()
```

```
17    84
48    39
50    31
67    23
33    19
      ..
14     1
63     1
79     1
26     1
60     1
Name: Source, Length: 83, dtype: int64
```

```
# Convert the text values in 'Speaker' column to numbers
fdf['Speaker'] = label_encoder.fit_transform(fdf['Speaker'])

# Check the updated 'Speaker' column
fdf['Speaker'].value_counts()
```

```
18    125
10    101
6      35
8      32
25     31
37     23
15     19
16     17
3      14
11     14
7      13
29     13
28     11
21     11
9       4
```

Figure 10: label encoding in data transformation in code

6. MODEL EVALUATION

For the model, we began by splitting the data into training and testing sets. Then we use two evaluation methods **Accuracy score** and for more details the **confusion matrix** since we have small dataset and partially imbalanced.

We then apply two supervised classifier algorithms. The first being a **Decision Tree Classifier**, when applying it to the training data it resulted in an accuracy of **0.54**. Additionally, we used a **Random Forest Classifier** on the training data, which yielded an accuracy of **0.59** as shown in table 1.

To address the issue of unbalanced data, we performed **oversampling**. This technique helps to balance the classes by generating synthetic samples from the minority class. After oversampling, we trained the Decision Tree Classifier again, and this time the accuracy improved to **0.79**. Similarly, we trained the Random Forest Classifier after oversampling, and the accuracy increased to **0.81**.

We also used **confusion matrix** for performance evaluation of the models to make sure it has a good performance with a good balanced representation of the dataset.

	Decision Tree	**Random Forest**
Without oversampling	0.54	0.59
With oversampling	0.79	0.81

Table1 : accuracy score evaluation performance

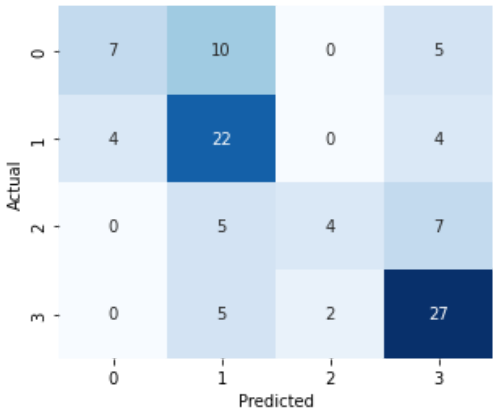


Figure 11: Random forest performance before oversampling

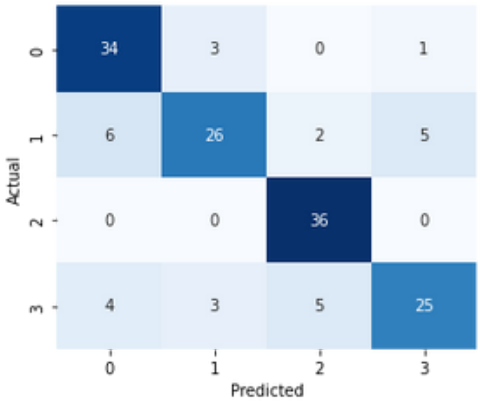


Figure 12: Random forest performance after oversampling

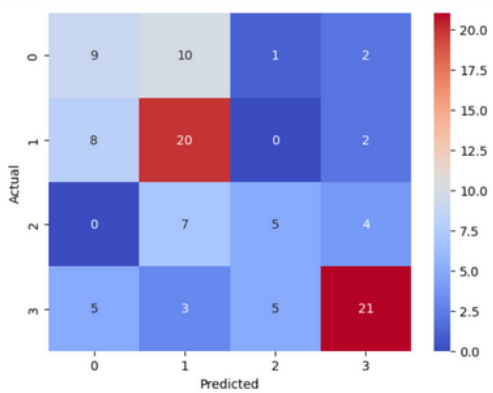


Figure 13 : Decision Tree performance before oversampling

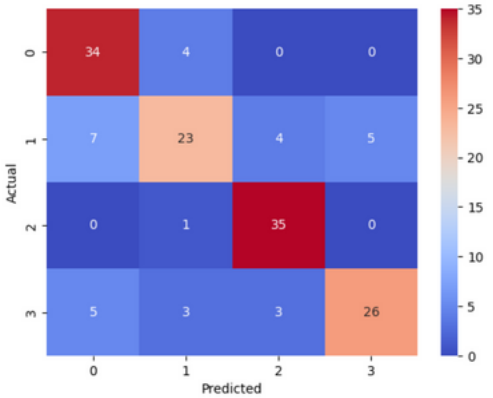


Figure 14 : Decision Tree performance after oversampling

We notice how the model had poor performance and accuracy at the begging and looking at the confusion matrix of both random forest and decision tree. **7** out of **22** are correctly classified as **0** in random forest while **9** out of **22** correctly classified as class **0** for example.
for label **1** in random forest had **22** out of **30** while decision tree had only **20**. we notice how without oversampling and since data was imbalanced random forest and decision tree were having little similar performance. but with oversampling we can notice how random forest has slightly better performance but they are both similar. for label **0** they have the same result with **34** correct **0** labels out of **38**
while in label **1, 26** out **30** correctly classified in random forest while **23** out of **30** in decision tree. In label **2** we can see the how random forrest correctly classified all **36** samples while decision tree missed one sample.

CONCLUSION

The project is a system for classifying Hadith as Good, weak, true, not true. It combines web scraping, data processing, visualization, transformation and model building to create a comprehensive approach to understanding and analyzing the dataset. The web scraping process collected a large dataset of Hadith texts from online sources, while data processing techniques ensured consistency and cleanliness. Visualization techniques provided insights into the dataset, allowing for deeper understanding of word frequency, label distributions, speaker occurrences, and correlations between labels and sources. The model development phase involved splitting the data into training and testing sets and applying decision tree and random forest classifiers. then evaluating their performance with proper evaluation metrices that ensured balance and quality results like confusion matrix and accuracy score. The project contributes to Islamic studies by providing a reliable tool for accurately and efficiently assessing the authenticity of Hadith te

REFRENCES

الدرر السنية - الموسوعة الحديثية - نتائج البحث
dorar.net

python Package Index
beautifulsoup4
Screen-scraping library
pypi.org

20حديثا ضعيف السند قوى التأثير.. نصوص غير صحيحة يحبها الناس
youm7.com

https://hdith.com/?s