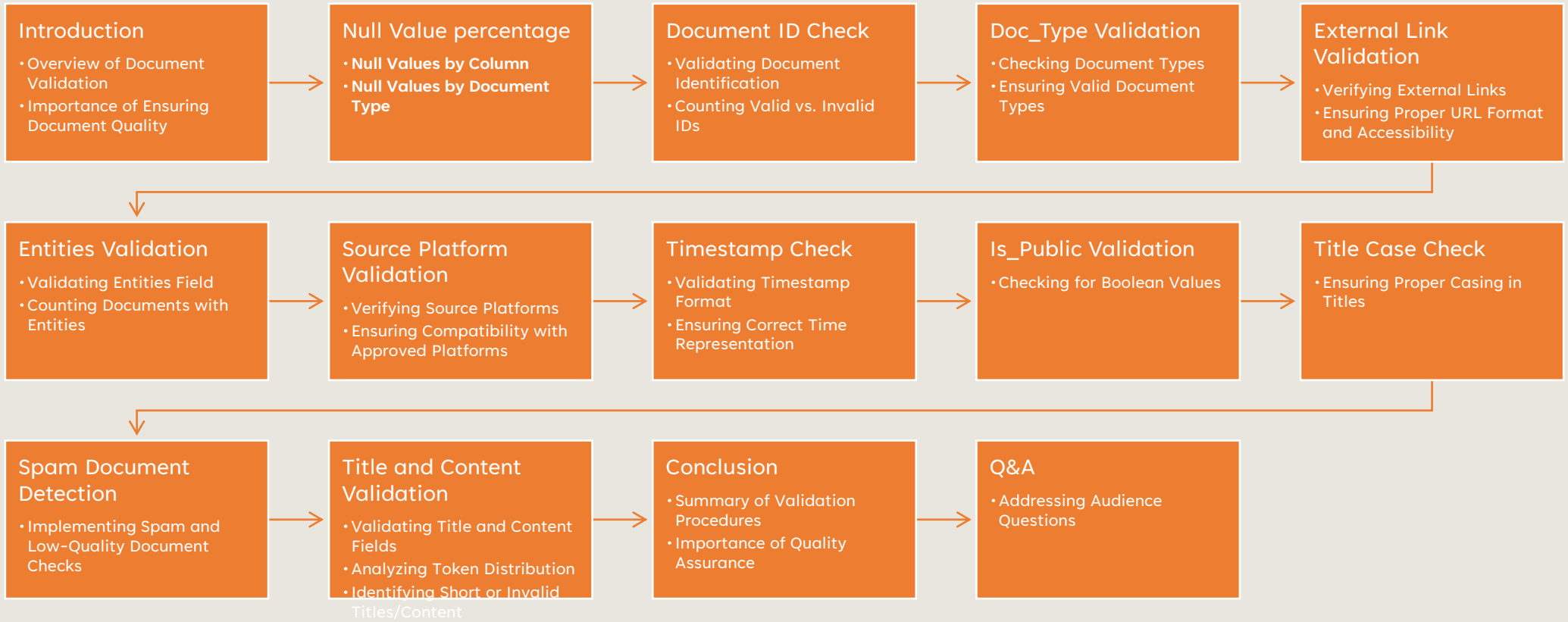
# DOCUMENT QUALITY ASSURANCE AND VALIDATION PROCEDURES

Shashank Shukla

# AGENDA


Introduction
• Overview of Document Validation
• Importance of Ensuring Document Quality
Null Value percentage
• Null Values by Column
• Null Values by Document Type
Document ID Check
• Validating Document Identification
• Counting Valid vs. Invalid IDs
Doc_Type Validation
• Checking Document Types
• Ensuring Valid Document Types
External Link Validation
• Verifying External Links
• Ensuring Proper URL Format and Accessibility
Entities Validation
• Validating Entities Field
• Counting Documents with Entities
Source Platform Validation
• Verifying Source Platforms
• Ensuring Compatibility with Approved Platforms
Timestamp Check
• Validating Timestamp Format
• Ensuring Correct Time Representation
Is_Public Validation
• Checking for Boolean Values
Title Case Check
• Ensuring Proper Casing in Titles
Spam Document Detection
• Implementing Spam and Low-Quality Document Checks
Title and Content Validation
• Validating Title and Content Fields
• Analyzing Token Distribution
• Identifying Short or Invalid Titles/Content
Conclusion
• Summary of Validation Procedures
• Importance of Quality Assurance
Q&A
• Addressing Audience Questions

# INTRODUCTION

In this presentation, we will delve into the meticulous process of data validation and anomaly detection. With a dataset comprising approximately 2 million records, our objective is to ensure data accuracy and integrity before feeding it to the Vespa. We'll explore various validation checks and methodologies, shedding light on key insights and findings. Join us on this journey as we unravel the intricacies of data quality assurance and anomaly identification.

# PERCENTAGE OF NULL VALUES FOR EACH COLUMN

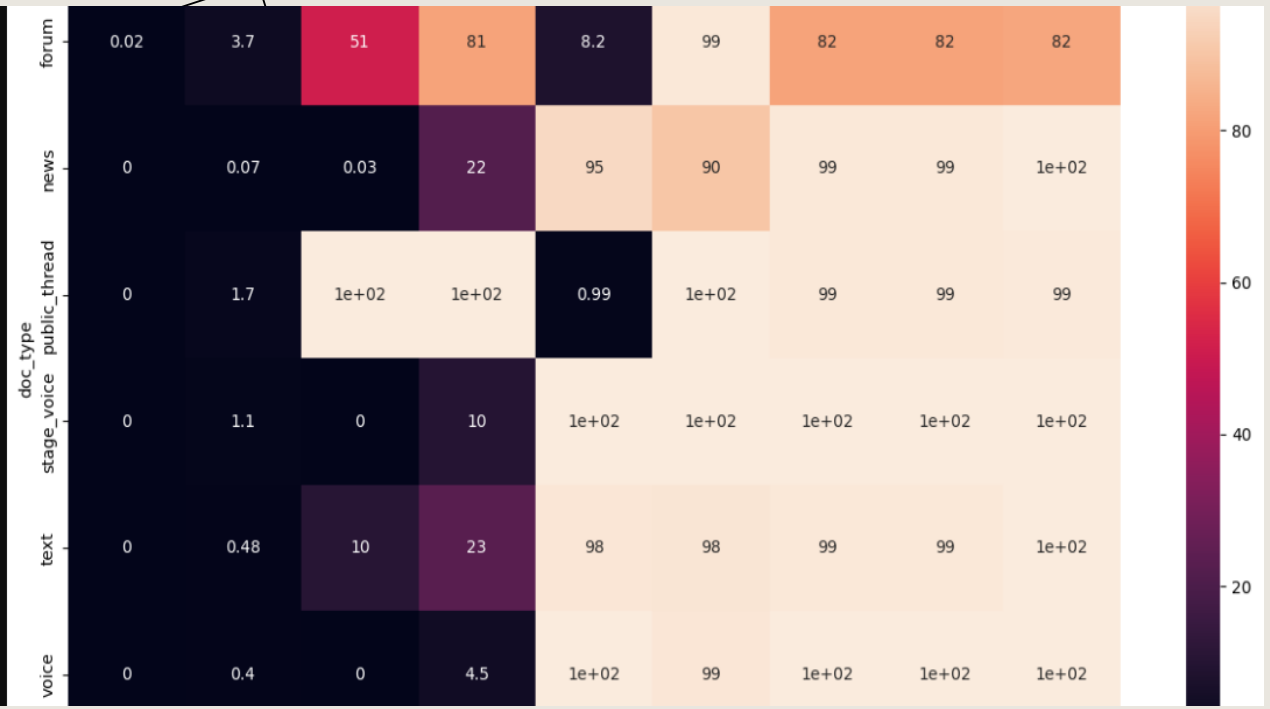
| Column | Null Value percentage |
|---|---|
| attachments | 98.53 |
| entities | 96.80 |
| thread_id | 95.80 |
| title | 95.80 |
| parent_doc_id | 82.39 |
| user_roles | 33.30 |
| user_avatar | 17.75 |
| content | 1.04 |
| user_display_name | 0.05 |
| doc_type | 0.05 |
| external_link | 0.05 |
| channel_name | 0.05 |
| community_name | 0.05 |
| user_id | 0.05 |
| user_other_name | 0.002 |
| user_name | 0.002 |
| channel_id | 0.000051 |

# PERCENTAGE OF NULL VALUES BY DOCUMENT TYPE

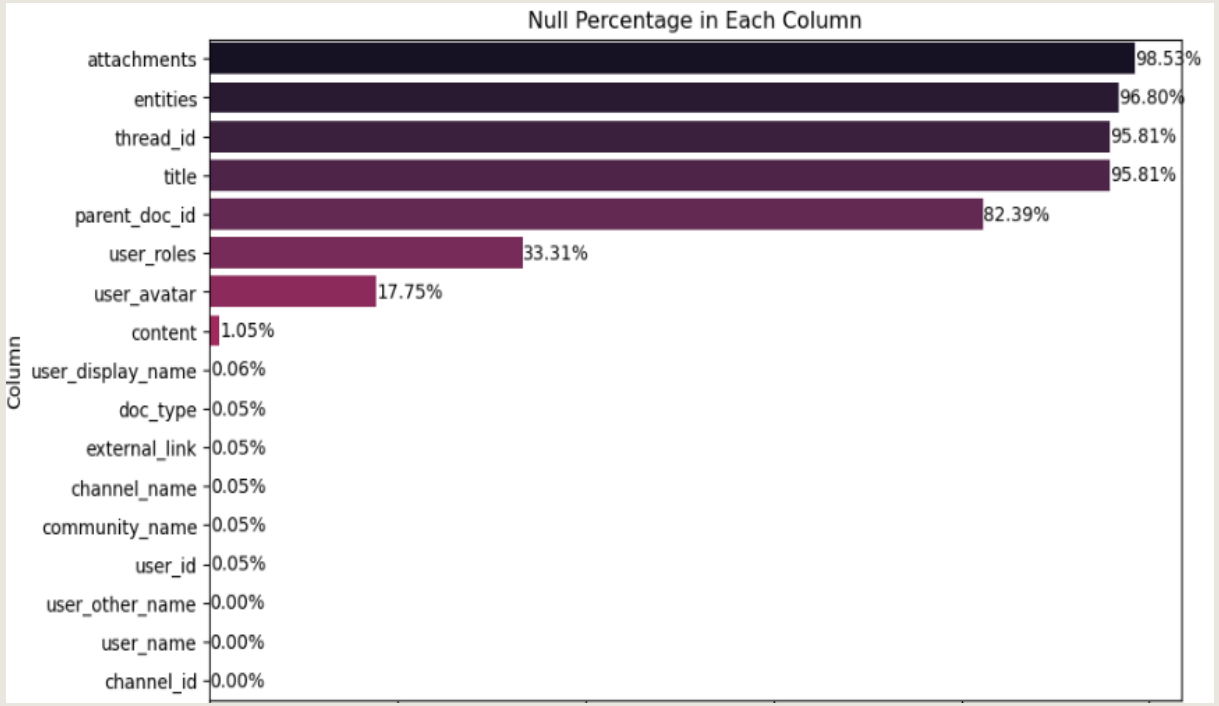
| doctype | username | content | user avatar | user roles | Parent Doc Id | Attachments | title | thread_id | entities |
|---|---|---|---|---|---|---|---|---|---|
| **forum** | 0.02 | 3.68 | 51.48 | 81.35 | 8.19 | 99.20 | 81.57 | 81.57 | 82.05 |
| **news** | 0.00 | 0.07 | 0.03 | 21.51 | 95.31 | 90.16 | 98.98 | 98.98 | 100.00 |
| **public thread** | 0.00 | 1.73 | 100.00 | 100.00 | 0.99 | 100.00 | 98.88 | 98.88 | 99.38 |
| **stage voice** | 0.00 | 1.11 | 0.00 | 10.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| **text** | 0.00 | 0.48 | 10.44 | 22.84 | 98.50 | 98.41 | 98.89 | 98.89 | 100.00 |
| **voice** | 0.00 | 0.40 | 0.00 | 4.53 | 100.00 | 98.80 | 100.00 | 100.00 | 100.00 |

# VISUAL REPRESENTATION OF NULL VALUES

## Document Type




## Column Wise



# DOCUMENT ID VALIDATION

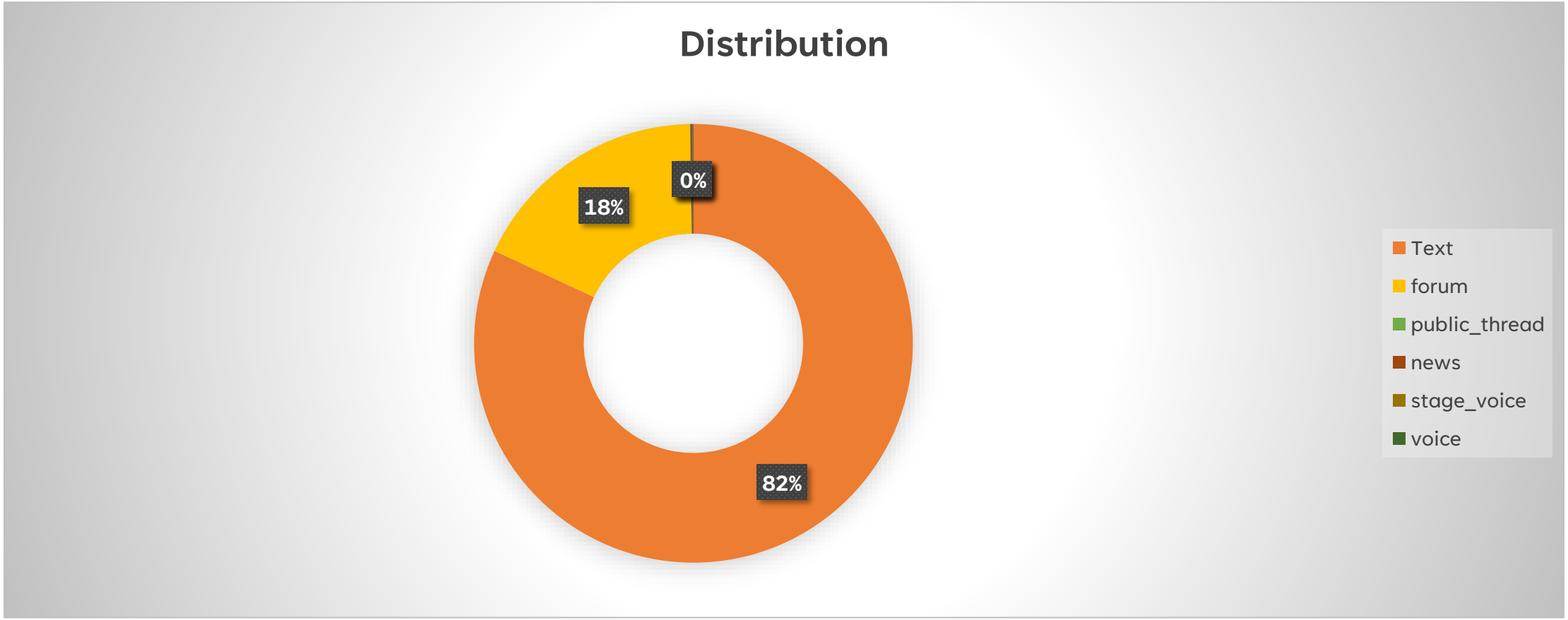
| Total Documents | Documents with Valid ids | Documents with invalid ids |
| --- | --- | --- |
| 19,67,974 | 19,67,974 | 0 |

# DOCUMENT TYPE VALIDATION

## Set Of Valid Document Types

- News
- text
- public thread
- stage voice
- voice
- Forum

# DOC TYPE PERCENTAGE DISTRIBUTION



# ENTITIES VALIDATION

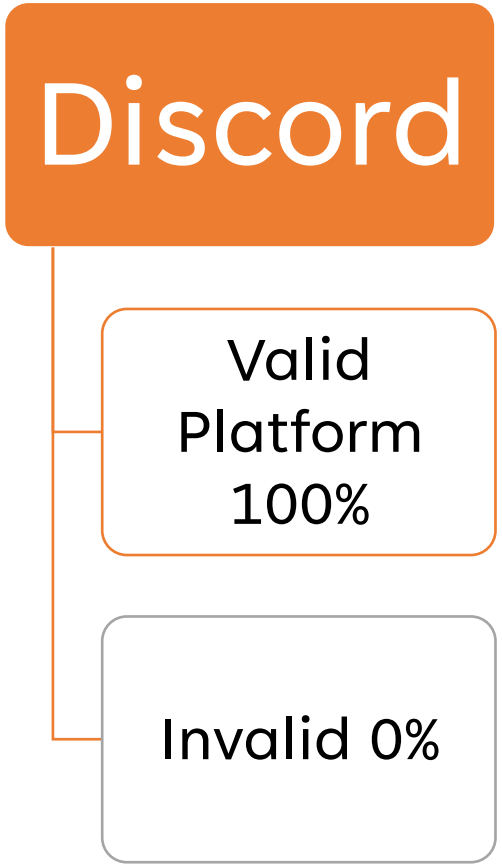
In this column 96.80% of values are Null

The total Invalid Entities is 97.87%

Total Valid Entities are 3.2%

# SOURCE PLATFORM VALIDATION

- There is only 1 Platform Type Discord



# TIMESTAMP VALIDATION

01

In this specific field, there are no null values.

02

In this validation, we have ensured that the dates in both the "created_at" and "created_at_str"

fields match and that the dates are not greater than today's date.

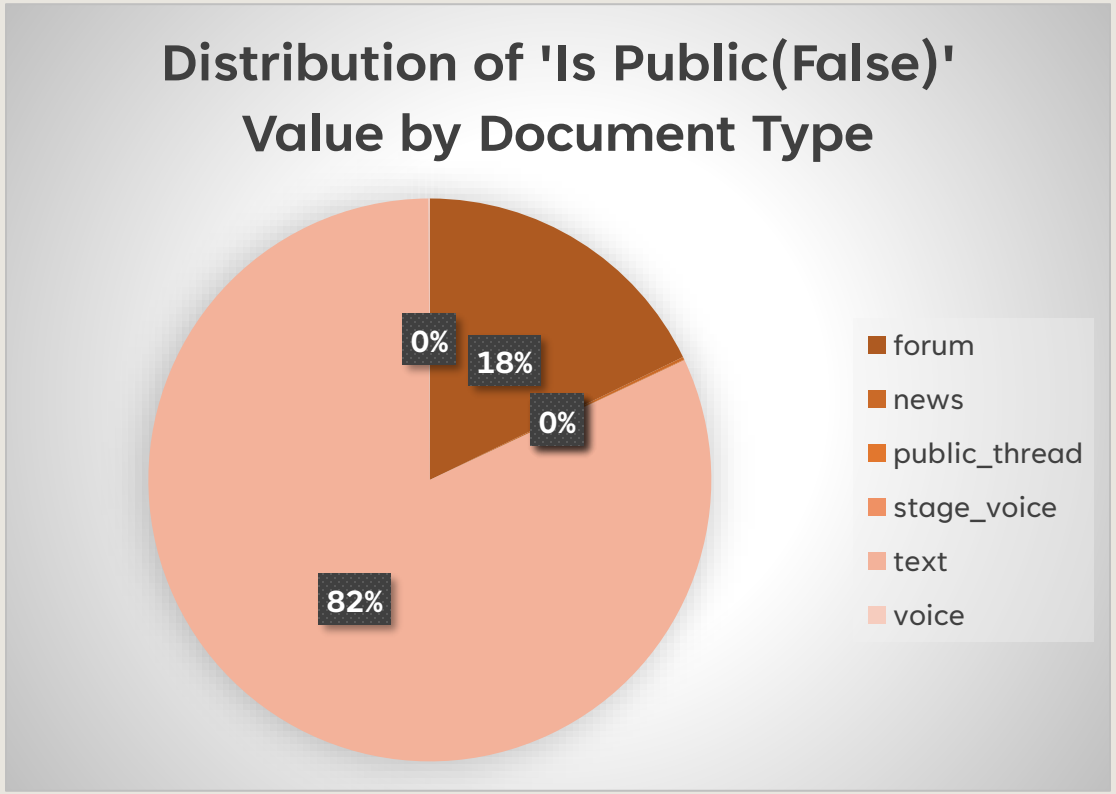
03

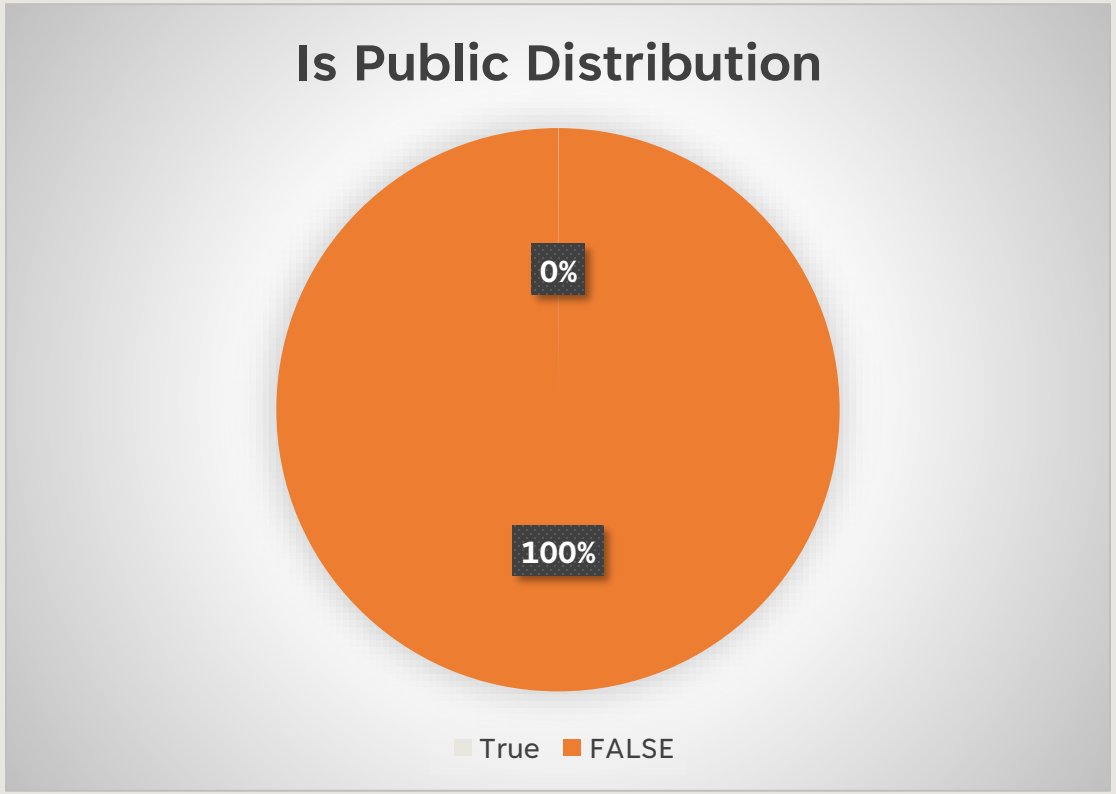
The total number of documents is 1,967,974, and all of them are considered valid.

# IS PUBLIC VALIDATION

| | |
|---|---|
| Total document | •19,67,974 |
| True | •626 |
| False | •19,67,348 |

# THE DISTRIBUTION OF THE "IS PUBLIC" ATTRIBUTE BASED ON DOCUMENT TYPE?



# TITLE CASE VALIDATION

| | |
|---|---|
| Total Document | 19,67,974 |
| Valid Title | 2.47% |
| Invalid Title | 97.57% |

# SPAM DOCUMENT DETECTION CONTENT

## Total Document

19,67,974

## Valid Title count

13,09,995 (67%)

## Invalid Tile count

6,57,979 (33%)

# SPAM DOCUMENT DETECTION TITLE

## Total Document

82470 (non-null)

## Valid Title count

67365 (81%)

## Invalid Tile count

15105 (19%)

# TITLE AND CONTENT VALIDATION

## Missing or Empty Values

- Title Missing or Empty: 1,885,504(95.81%) documents
- Content Missing or Empty: 20,602(1.05%) documents

## Title Less Than 3 Tokens

- Title non-null values: 82,470
  - 1 token: 14,148 (17%) documents
  - 2 tokens: 11,793(14%) documents
- Content
  - 1 token: 555,182(28%) documents
  - 2 tokens: 255,502(13%) documents

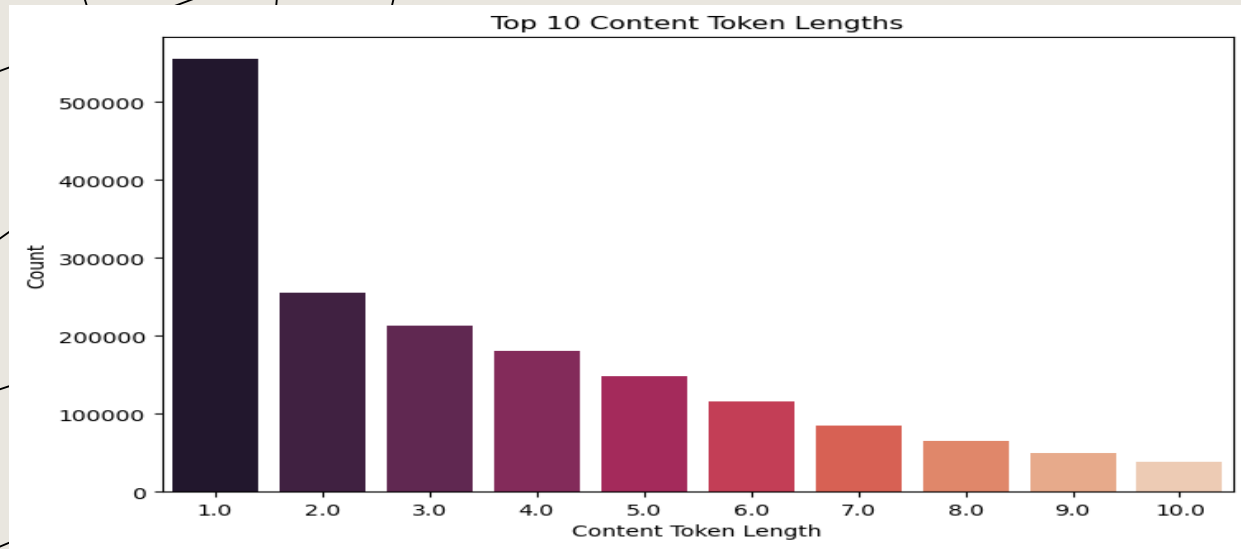
## Has Only Numbers

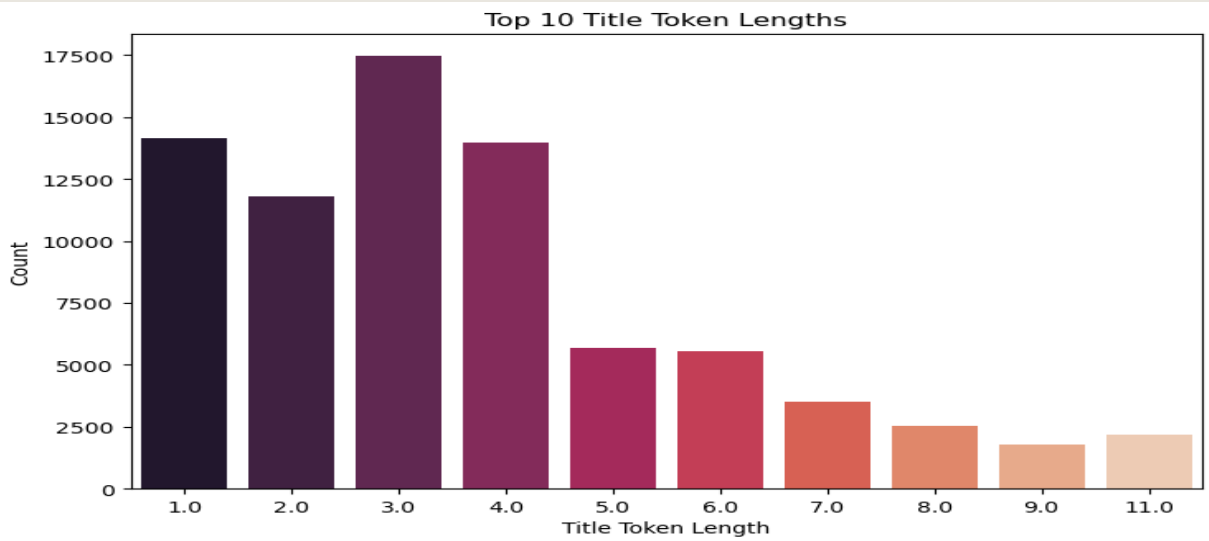
- Title Has Only Numbers: 5 documents
- Content Has Only Numbers: 108 documents

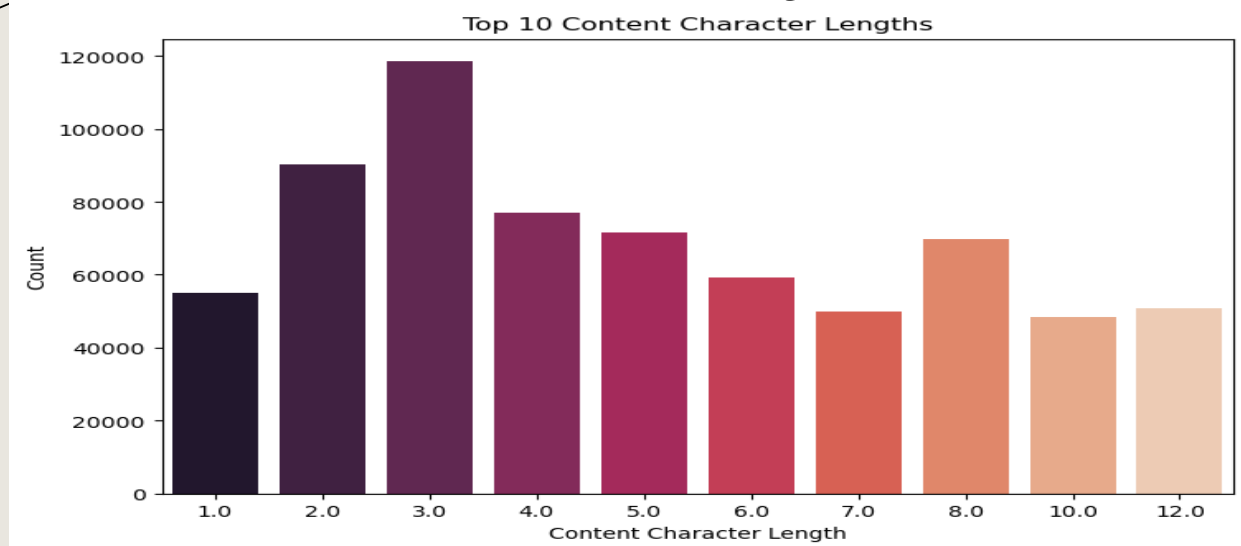
## Has Only Special Characters

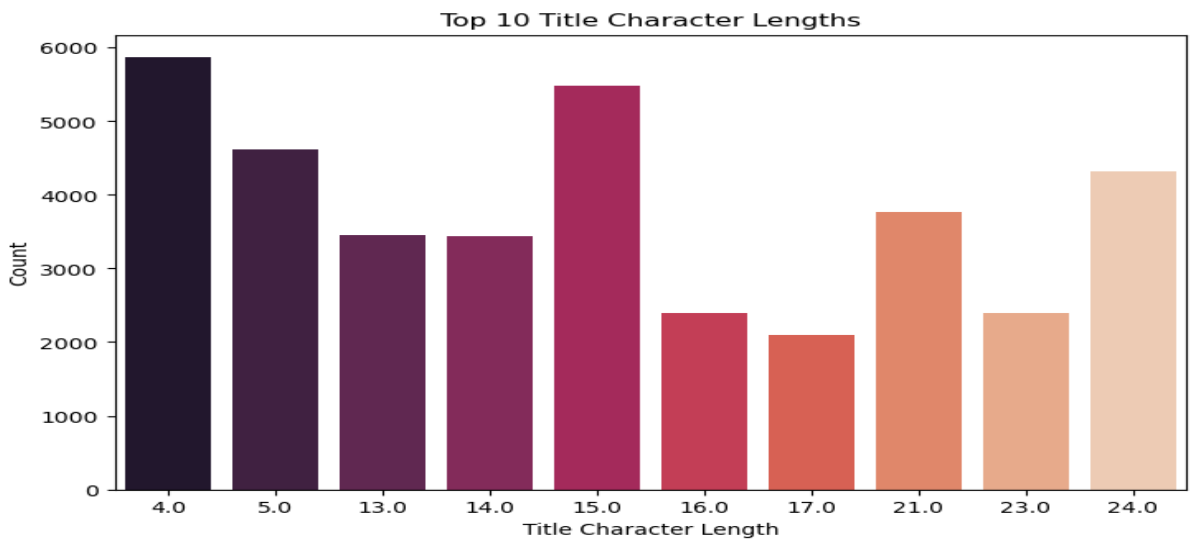
- Title Has Only Special Characters: 158 documents
- Content Has Only Special Characters: 992 documents

# TOP 10 TOKENS AND CHARACTER LENGTH



# SUMMARY

**In our comprehensive data analysis, we have identified and examined several critical aspects of the dataset, which is comprised of a total of 1,967,974 documents. Here are the key findings and insights:**

**Null Value Analysis**

We observed varying degrees of null values across different columns, with some columns having a high null percentage. For instance, 'attachments,' 'entities,' and 'thread_id' columns have null percentages exceeding 95%.

**Document Type Distribution**

The dataset contains a variety of document types. Most prevalent are 'text' and 'forum' types, accounting for 81.92% and 17.80% of the dataset, respectively. There is also a small percentage of 'news,' 'voice,' 'public_thread,' and 'stage_voice' document types.

**Document Validation**

We validated documents based on their document type, identifying discrepancies in null value percentages among different types. Notably, 'forum' and 'news' types exhibit significantly different null value patterns.

**Entity Analysis**

A portion of the documents (3.2%) contains valid entities. Additionally, 2.13% of the documents have at least two entities, while the majority (97.87%) do not contain any entities.

**Platform and Timestamp Validation**

All documents have valid source platforms and timestamps, ensuring data integrity.

**Is Public Attribute**

The 'is_public' attribute has a true value for 626 documents, indicating public accessibility, while 1,967,348 documents have a false value, indicating non-public content.

**Title and Content Quality**

The dataset has 2.47% of documents with valid titles, while the majority (97.53%) have missing or inadequate titles. Additionally, some documents have content with fewer than three tokens, and a few contain titles or content with numbers or special characters.

**Spam Detection**

Among the documents, 657,979 are identified as potential spam.

# THANK YOU

Shashank Shukla

shashanks.ven@splore.com