Analysis Report

Clickbait Identification

Language: Hindi

Members:

Adith John Rajeev 2020114010

Eshika Khandelwal 2020114018

Harshit Gupta 2020114017

GITHUB LINK: https://github.com/esh04/Clickbait-Identification

Introduction:

In this project, we have generated suitable parameters and compared their occurrence, count, properties etc., of Clickbait titles vs Non-Clickbait titles.

Clickbait is a pejorative term describing web content that aims to generate online advertising revenue, especially at the expense of quality or accuracy, relying on sensationalist headlines or eye-catching thumbnail pictures to attract click-throughs to encourage forwarding of the material over online social networks. Clickbait headlines typically aim to exploit the "curiosity gap", providing just enough information to make readers curious but not enough to satisfy their curiosity without clicking through to the linked content.

We propose various structural, linguistic, word and sentence level features for developing a classification model. We have then classified the data using simple classifier models such as SVMs, Logistic Regression and XGBoost.

The dataset we used contained 41800 Hindi sentences labelled either 0(Non-Clickbait) or 1(Clickbait).

Pre-processing:

The Hindi sentences in the dataset required cleaning up and replacing several ASCII character codes into their actual punctuations. We observed that each sentence terminated with '\n'. Also, in several sentences, we observed ASCII characters such as ''' and 'quot;' representing quotation marks in the text. We replace them with suitable characters accordingly. We also remove necessary punctuations, brackets, and

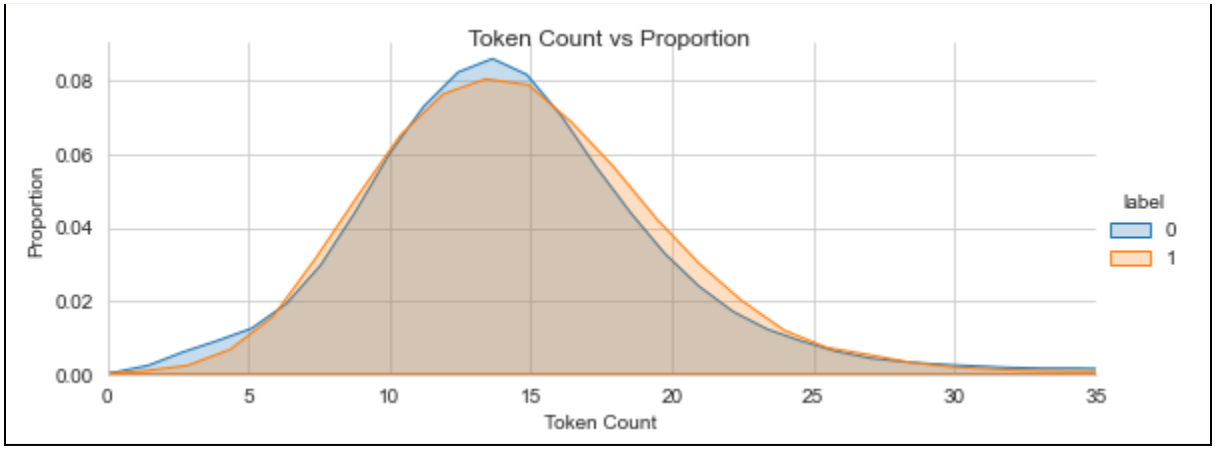
dots (sequences such as '…' or '..') that show some continuity in the title. We let the numbers and English characters remain for analysis since it affects token count and sentence structure. Furthermore, it helps us with unique and accurate analysis since there is a possibility that the sentences given to us in the test case might contain the same.

Analysis:

We generated our very own parameters based on various structural, linguistic, word and sentence level features for analysis and classification purposes.

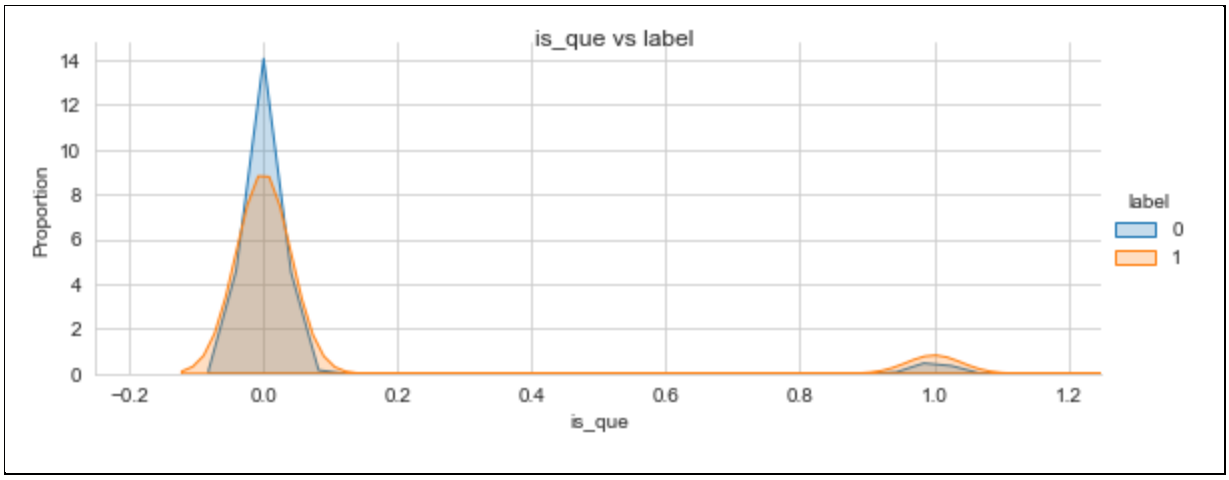
Analysis is done on the basis of-

1. Number of Tokens



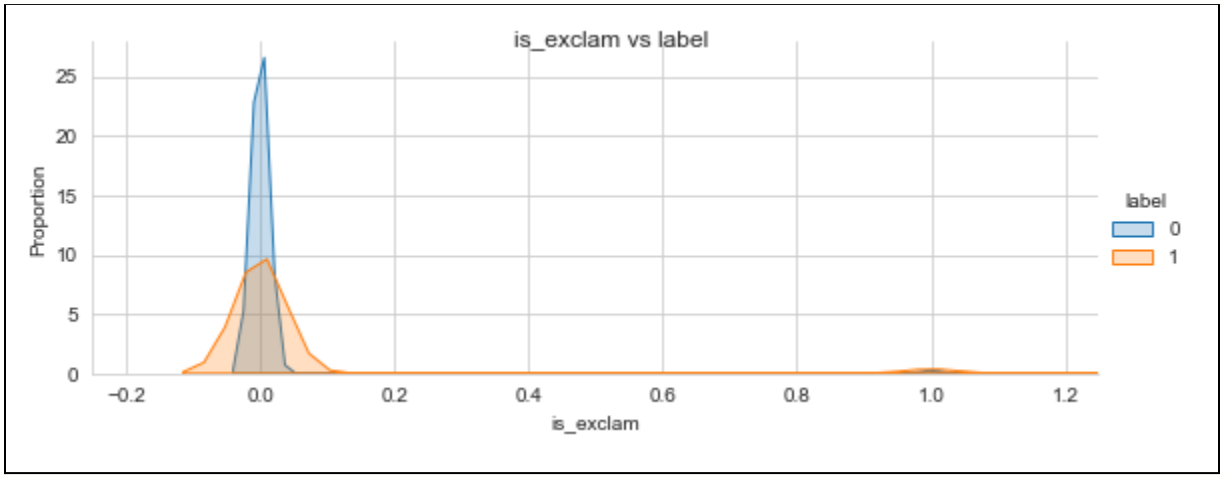
We analyse whether the number of words in a sentence impacts the binary classification. It is interesting to see that in the dataset, both Clickbait and Non-Clickbait sentences are NOT heavily impacted by the number of tokens in the sentence. Majority of the titles have around 13-15 tokens in them. So, while the graph shows minor differences, the impact of 'Number Of Tokens' is not that severe, probably because hindi sentences require more tokens to express verbs which contributes significantly to the token weightage.

2. Presence of Question Marks



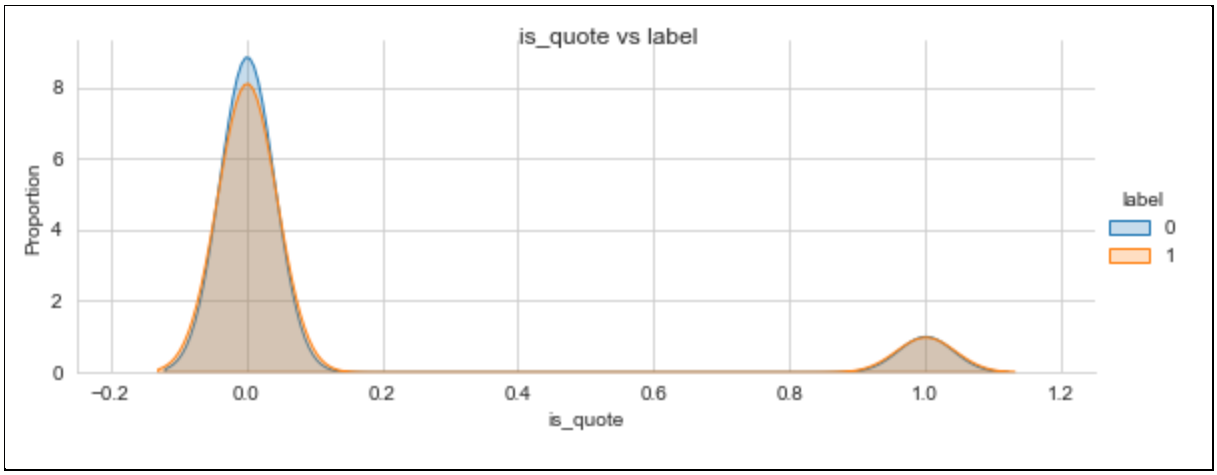
We analyse whether the occurrence of a question mark(?) in a sentence impacts the binary classification. It is interesting to see that in the dataset that clickbait sentences have a higher tendency to have question marks in them as compared to non-clickbait sentences. However, the occurrence of a question mark is significantly lower in both the binary classes as seen in the graph.

3. Presence of Exclamation Marks



We analyse whether the occurrence of an exclamation mark(!) in a sentence impacts the binary classification. It is interesting to see that in the dataset that clickbait sentences have a higher tendency to have exclamation marks in them as compared to non-clickbait sentences. However, the occurrence of an exclamation mark is significantly lower in both the binary classes as seen in the graph.

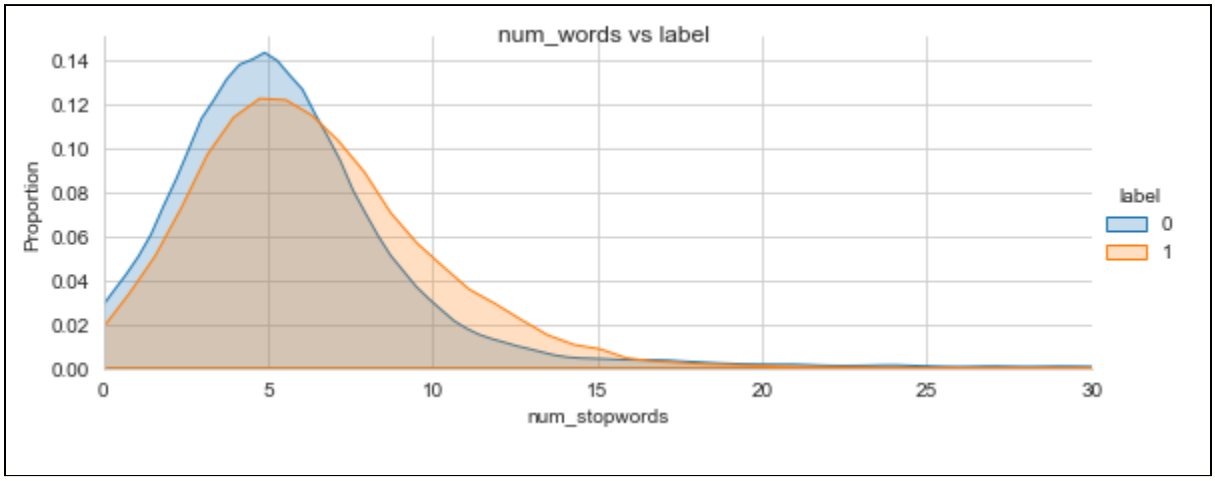
4. Presence of Quotations



The presence of quotation marks does NOT help us greatly in the process of binary classification. The only interesting thing to note is that several clickbait and non-clickbait sentences have quotations, phrases and other such aspects in them.

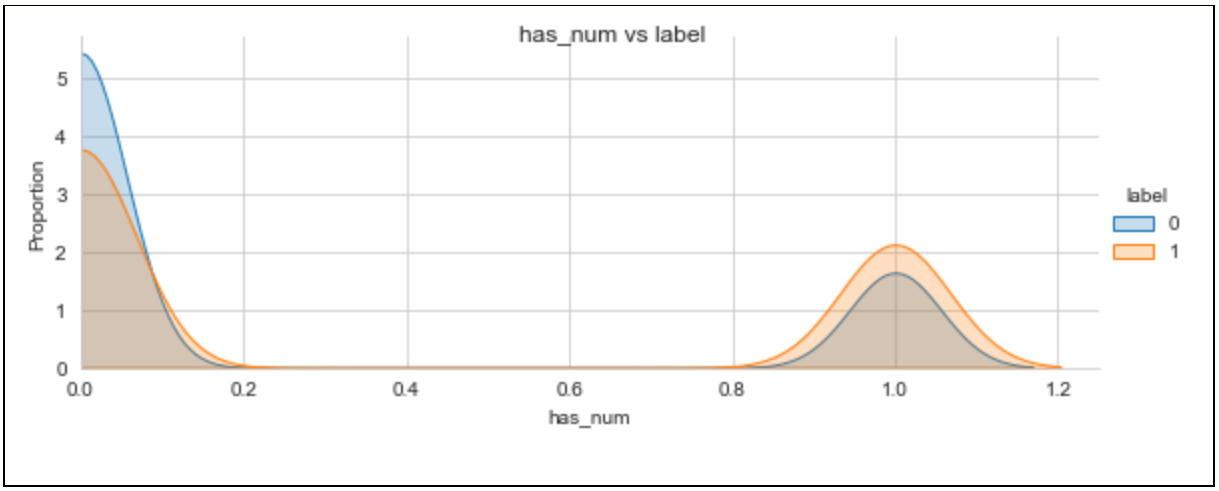
5. Number of Stopwords

A custom stopword.txt file was used in order to locate the stopwords. The analysis graph is as follows:



The average number of stop words in Clickbait news headlines is 4.85 words whereas the average number in Non-clickbait headlines was 5.73 words.

6. Presence of Numerals

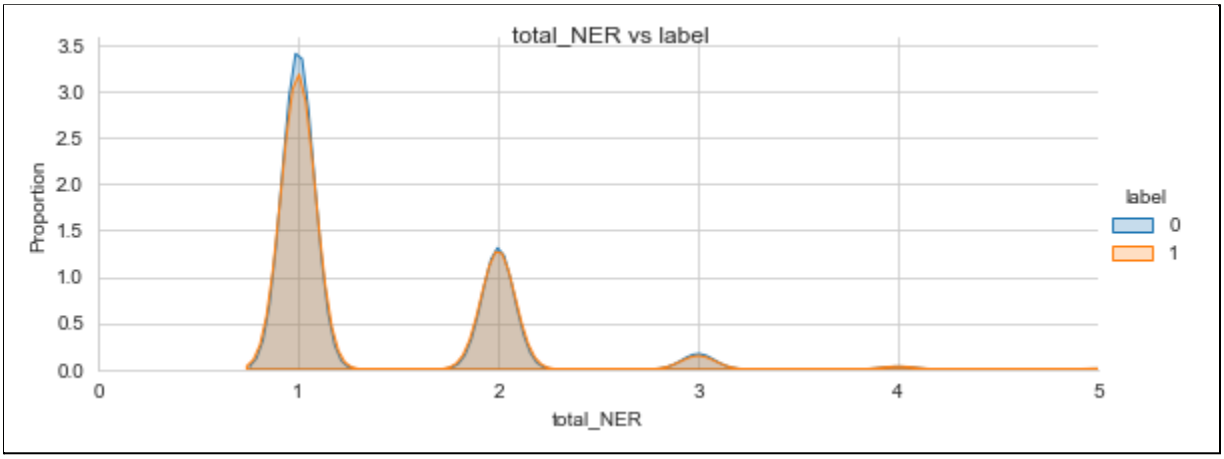


We make several interesting observations here. Around 40% of the clickbait sentences have numerals in them. This shows that having numerics in a sentence can impact the binary classification process. The number of sentences having numerals in them is fairly higher for clickbait sentences in comparison to non-clickbait sentences. This makes it a fairly necessary parameter for the process of binary classification.

7. The Entities Present

We used the Polyglot libraries Named Entity Recognition to do this section of the analysis.

Below is the graph of the NERs present in Clickbait titles Vs. Non-Clickbait-



The NER comparison graph turned out to be very similar for this particular dataset.

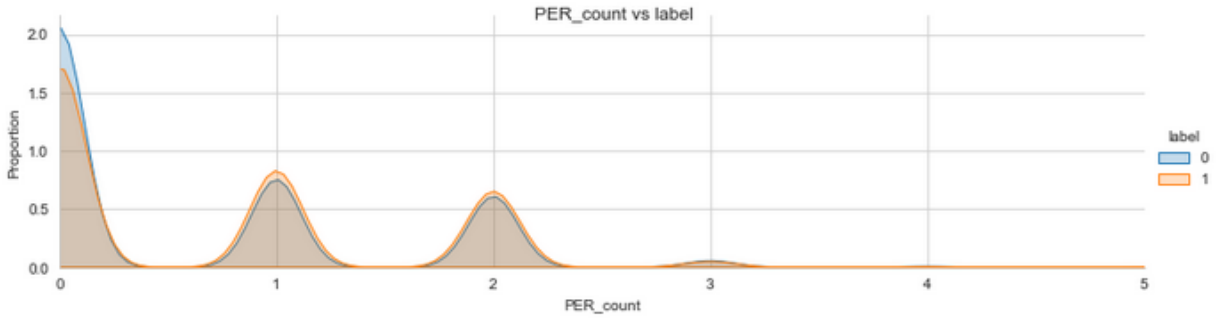
The average number of NER's in:

Non-Clickbait titles: 1.3118987341772153

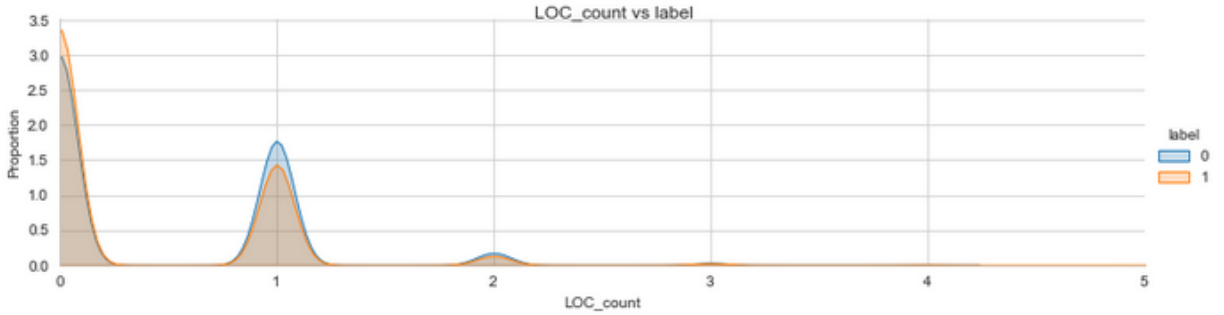
Clickbait titles: 1.3179005524861878

Hence we divided the NER's into three categories in order to analyse further - Personal NERs, Organisation NERs, Location NERs.

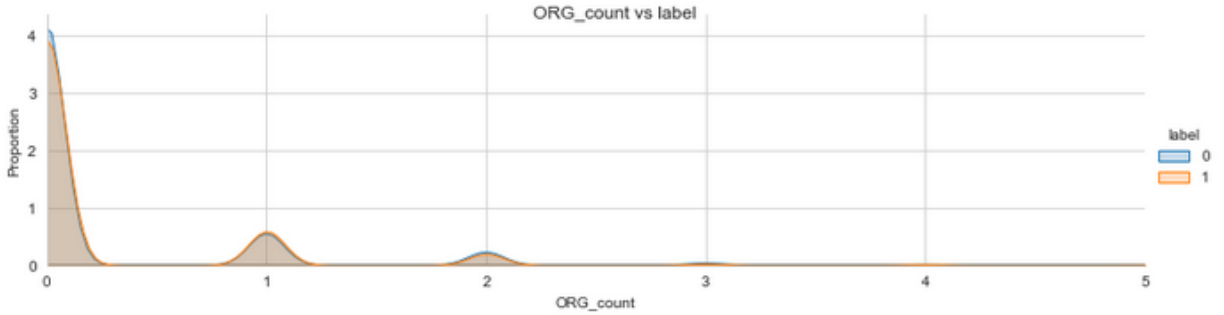
Personal NER:




```
Average number of Personal NER's in:
Non-Clickbait titles: 0.629915611814346
Clickbait titles: 0.7158563535911602
```




```
Average number of Location NER's in:
Non-Clickbait titles: 0.44814345991561183
Clickbait titles: 0.360939226519337
```




```
Average number of NER's in:
Non-Clickbait titles: 0.23383966244725737
Clickbait titles: 0.24110497237569062
```

The average words in the three cases were also indistinguishable

To get a clearer view on the differences, we plotted word clouds only of those words that were unique to each set. For example, the Non-clickbait Personal NER cloud has only those Personal NERs that are present in the Non-clickbait titles and absent in the Clickbait titles.

Personal NER

1. ***Non-clickbait***

2. ***Clickbait***

[('पाल्ट्रो', 13), ('स्टारबक्स', 11), ('ओझा', 9), ('ग्वेनेथ', 9), ('डेमोक्रेट्स', 7), ('मैरिड', 7), ('रिएक्शन', 7), ('दाऊद', 6), ('फ्रैंकलिन', 6), ('डिक', 6), ('ग्रिंट', 6), ('ब्रेट', 6), ('जेडन', 6), ('स्पीलबर्ग', 6), ('प्लेड', 6), ('जेम्सन', 6), ('रेगी', 5), ('वाट्स', 5), ('कारमेलो', 5), ('डेस', 5), ('स्वानसन', 5), ('एगुइलेरा', 5), ('सोफी', 5), ('ड्वेन', 5), ('माया', 5), ('होम', 5), ('गैम्बियास', 5), ('ट्रेश्ड', 5), ('मेमोरियल', 5), ('पेज', 5), ('मिसेज', 5), ('कैन', 5), ('विद', 5), ('हावर्ड', 5), ('रूकी', 4), ('हैनिगन', 4), ('गर्ल्स', 4), ('स्पेल', 4), ('बॉयज़', 4), ('डेथ', 4), ('विदरस्पून', 4), ('गोर्का', 4), ('मेमे', 4), ('पांडा', 4), ('क्राउड', 4), ('महोने', 4), ('जवाब', 4), ('ग्रोवर', 4), ('रेजिना', 4), ('फाउंडिंग', 4), ('रुइन', 4), ('फैक्ट', 4), ('मिलेनियर', 4), ('मौर्य', 4), ('फेलिक्स', 3), ('गोर्सुख', 3), ('वेगा', 3), ('एटनबरो', 3), ('हरमायनी', 3), ('फ्लोरिडा', 3), ('होली', 3), ('हेगन', 3), ('नेवर', 3), ('टेकी', 3), ('एडवेंचर्स', 3), ('गेटिंग', 3), ('मैकमोहन', 3), ('नाउ', 3), ('मैकायला', 3), ('असांजे', 3), ('हाईएंड', 3), ('इज़ी', 3), ('लोन', 3), ('गेटी', 3), ('चोई', 3), ('बेलोंग', 3), ('सेबस्टियन', 3), ('हू', 3), ('वाया', 3), ('वेड', 3), ('ओपीओइड', 3), ('ओडी', 3), ('सिरी', 3), ('क्लियोपेट्रा', 3), ('कॉनन', 3), ('हिप्स्टर्स', 3), ('चीफ', 3), ('टाईस', 3), ('कॉस्मो', 3), ('क्रेमर', 3), ('ओमारो', 3), ('एंगर्स', 3), ('लायन', 3), ('गायकवाड़', 3), ('रिडले', 3), ('जोकुन', 3), ('मिर्का', 3), ('अल्बा', 3), ('श्मिट', 3), ('सुज़ैन', 3)]

Location NER:

1. ***Non-Clickbait***

[('नरेगा', 9), ('विस्कॉन्सिन', 7), ('अहमदाबाद', 7), ('मद्रास', 6), ('पुष्कर', 6), ('कश्मीरी', 6), ('इस्लाम', 6), ('जिम्बाब्वे', 5), ('लेबनान', 5), ('हेनरी', 5), ('तीन', 5), ('जामनगर', 5), ('सेंटो', 5), ('रवांडा', 5), ('खापें', 5), ('वर्ल्ड', 5), ('गवर्नमेंट', 5), ('एकदिवसीय', 4), ('क्रिकेट', 4), ('इबेरिया', 4), ('रामगढ़', 4), ('अलेप्पो', 4), ('माल्टा', 4), ('हैकर्स', 4), ('दिमित्री', 4), ('पेइट', 4), ('हैम', 4), ('अफ़्रीकी', 4), ('बर्मिंघम', 4), ('प्रेसीडेंसी', 4), ('राइट', 4), ('आयरिश', 4), ('व्हाइट', 4), ('क्लार्क', 4), ('गुआंतनामो', 4), ('सर्वे', 4), ('एनसीपी', 3), ('औरंगाबाद', 3), ('हिल्स', 3), ('नॉरफ़ॉक', 3), ('बरेली', 3), ('ब्रिस्टल', 3), ('बेरूत', 3), ('रियाद', 3), ('एमडी', 3), ('शेफ़ील्ड', 3), ('डेड', 3), ('कजाकिस्तान', 3), ('नाजी', 3), ('ब्रिजपोर्ट', 3), ('आसनसोल', 3), ('काहिरा', 3), ('न्यूज़क्वेस्ट', 3), ('इक्वेटोरियल', 3), ('इराकी', 3), ('कोस्टा', 3), ('माली', 3), ('जीई', 3), ('मालाबार', 3), ('ओडिशा', 3), ('उत्तरजीवितावादी', 3), ('तमिल', 3), ('इंगलैंड', 3), ('किशोर', 3), ('फैंसी', 3), ('धाबी', 3), ('मथुरा', 3), ('सियोल', 3), ('आर्लिंगटन', 3), ('टैक्सी', 3), ('नेशनलिस्ट', 3), ('डेरियस', 3), ('रहीमी', 3), ('उत्तराखंड', 3), ('नॉट', 3), ('पोरबंदर', 2), ('कार्लोस', 2), ('स्लिम', 2), ('केन', 2), ('व्याट', 2), ('कैथोलिक', 2), ('नेस्ले', 2), ('पुतिन', 2), ('फेरारी', 2), ('मॉरीशस', 2), ('एफसी', 2), ('काकाडू', 2), ('नासिक', 2), ('रोमानियाई', 2), ('लाजपत', 2), ('अंतर्राष्ट्रीय', 2), ('कंजर्वेटिव', 2), ('उपनिवेश', 2), ('शेटलैंड', 2), ('पड़ोसी', 2), ('सोनीपत', 2), ('पीबीएस', 2), ('बैराज', 2), ('तंजानिया', 2), ('डेनिमगढ़', 2)]

2. Clickbait

[('बीवर', 6), ('लेह', 6), ('जॉन', 5), ('रूडी', 4), ('स्टेडियम', 4), ('नेकां', 4), ('सननिडेल', 4), ('तिमोर', 4), ('लीजेंड', 4), ('रशमोर', 3), ('पुर्तगाल', 3), ('द्वितीय', 3), ('लैपलैंड', 3), ('ओबामास', 3), ('बीच', 3), ('ट्रंप', 3), ('कॉमी', 3), ('फिदेल', 3), ('कास्त्रो', 3), ('ऑर्गनाइज़र', 3), ('मुग़ल', 3), ('पेंटहाउस', 3), ('लंड', 3), ('काठमांडू', 3), ('स्टेट', 3), ('पुन्स', 2), ('इटालियन', 2), ('किर्गिस्तान', 2), ('भारतीयों', 2), ('क्रिसमस', 2), ('विदेश', 2), ('डाउनटाउन', 2), ('समरसेट', 2), ('स्लोवाकिया', 2), ('देशभर', 2), ('रिकर्स', 2), ('उप्र', 2), ('ताहिती', 2), ('डेट्रायट', 2), ('जॉर्ज', 2), ('स्टूडेंट', 2), ('इलेक्शन', 2), ('ग्रॉग्स', 2), ('नैरोबी', 2), ('वेयरहाउस', 2), ('जोंगउन', 2), ('हेट', 2), ('होक्स', 2), ('ट्रम्पविरोधी', 2), ('पार्लियामेंट', 2), ('लबौर', 2), ('कान्सास', 2), ('प्राचीन', 2), ('मणिपुर', 1), ('जोसेफ', 1), ('सेल्फी', 1), ('संता', 1), ('आर्म्स', 1), ('राजधानी', 1), ('शहरों', 1), ('इंडियानापोलिस', 1), ('डीएनसी', 1), ('लौवर', 1), ('क्लैम्सन', 1), ('स्टुअन्स', 1), ('एक्सेटर', 1), ('नेब्रास्का', 1), ('ग्रेसर्ड', 1), ('रॉकेट्स', 1), ('पीटर्सबर्ग', 1), ('आलियाह', 1), ('सातवें', 1), ('डॉक्सिंग', 1), ('चेसापिक', 1), ('बेसबॉल', 1), ('लैग्स', 1), ('ट्रैवल', 1), ('बैन', 1), ('यूरोजोन', 1), ('एमिरेट्स', 1), ('ब्रूकलिन', 1), ('फेड्स', 1), ('आइवरी', 1), ('एरिजोना', 1), ('सारलैंड', 1), ('उबेर', 1), ('हेंज', 1), ('हाइकु', 1), ('इज़ी', 1), ('लोए', 1), ('पोप', 1), ('फ्रांसिस', 1), ('रेमन', 1), ('जोय', 1), ('कोलोन', 1), ('फ़्लोरिडा', 1), ('पड़ोस', 1), ('न्यूफ़ाउंडलैंड', 1), ('स्टॉप्स', 1), ('रशियन', 1)]

Organisation NER

1. ***Non-clickbait***

```
[('मर्चेंट', 9), ('जुर्वेटस', 7), ('न्याय', 7), ('सुपर', 7), ('संरक्षण', 6), ('लोकपाल', 6), ('वायु', 5), ('पर्यावरण', 5), ('सेंट', 5), ('पीटर्सबर्ग', 5), ('ऑफ़', 5), ('ताइवान', 4), ('प्रतिनिधिमंडल', 4), ('हमास', 4), ('इंग्लिश', 4), ('बैंकॉक', 4), ('टोटेनहैम', 4), ('नीति', 4), ('क्वींस', 4), ('पूर्णकालिक', 4), ('देयर', 4), ('हेनरिक', 4), ('स्टेंसन', 4), ('बांग्लादेश', 4), ('संस', 4), ('एडिसन', 4), ('ग्रीनपीस', 4), ('एफिल', 4), ('रेलवे', 4), ('प्रूडेंशियल', 4), ('जॉन', 4), ('कैडबरी', 4), ('ग्रेग', 3), ('प्रशिक्षित', 3), ('कमांडो', 3), ('कैसीनो', 3), ('सत्ताधारी', 3), ('हार्ट्स', 3), ('डे', 3), ('मेयो', 3), ('ठाणे', 3), ('कैवानी', 3), ('ब्रेस', 3), ('हिजबुल्ला', 3), ('समर्थक', 3), ('ईयू', 3), ('हिज', 3), ('आईबीएम', 3), ('जापान', 3), ('मोहन', 3), ('बागान', 3), ('चेन्नई', 3), ('हार्ड', 3), ('दावोस', 3), ('काइल', 3), ('एडमंड', 3), ('वर्जिन', 3), ('ब्लू', 3), ('ब्रेकफास्ट', 3), ('पैरोडीडिफेंडिंग', 3), ('पोपोविच', 2), ('फॉरेस्ट', 2), ('कोस्ट', 2), ('नगरपालिका', 2), ('वर्जीनिया', 2), ('पूसा', 2), ('जेनर', 2), ('ट्रैवल', 2), ('बर्लिन', 2), ('पेंटागन', 2), ('पीडब्ल्यूसी', 2), ('यूनिटी', 2), ('रॉयल', 2), ('रिप्ले', 2), ('टाइगर', 2), ('दलीप', 2), ('ताहिल', 2), ('फैट', 2), ('बीपीओ', 2), ('ब्रिज', 2), ('मार्स', 2), ('यहूदी', 2), ('निकाय', 2), ('पैनल', 2), ('केयर', 2), ('सीज', 2), ('मॉस्को', 2), ('बेटन', 2), ('ओरियन', 2), ('बायोग्राफी', 2), ('स्नूकर', 2), ('फाइनेंस', 2), ('स्वीट', 2), ('मेसी', 2), ('ईरान', 2), ('पूर्वी', 2), ('एमवीपी', 2), ('कोवेंट्री', 2), ('एलजी', 2), ('मनोरंजन', 2)]
```

2. Clickbait

[('जर्मन', 7), ('पिक्सार', 6), ('विंडोज', 6), ('वोक्सवैगन', 5), ('बान', 4), ('टिम', 4), ('जेफ़', 4), ('रिजर्व', 4), ('पोर्ट्रेट', 4), ('गोथम', 4), ('बीपी', 4), ('केमैन', 4), ('इंटरनेट', 4), ('कीस्टोन', 4), ('डिज्नी', 3), ('नाथन', 3), ('बेडफोर्ड', 3), ('रशमोर', 3), ('क्रिस्मस', 3), ('केली', 3), ('आयोटे', 3), ('कॉनस', 3), ('हैम्पशायर', 3), ('वोग', 3), ('ज्योग्राफिक', 3), ('ब्लैंड', 3), ('बॉडीगार्ड', 3), ('सदर्न', 3), ('ओरेगन', 3), ('दुबई', 3), ('हमलावर', 3), ('एरिक', 3), ('होल्डर', 3), ('रिप्स', 3), ('पुनर्जागरण', 3), ('नाइस', 3), ('डिप्टी', 3), ('एक्सेल', 2), ('वेटिकन', 2), ('मिनेसोटा', 2), ('वॉल्ट', 2), ('जेसी', 2), ('पेनी', 2), ('इलस्ट्रेटेड', 2), ('जर्सी', 2), ('एबे', 2), ('यूनिसेफ', 2), ('कैलिफ़ोर्निया', 2), ('जस्टिस', 2), ('होलीवुड', 2), ('एमएसएनबीसी', 2), ('टीवी', 2), ('एंटोनियो', 2), ('लैंग्वेज', 2), ('म्यूजिकल', 2), ('साइन', 2), ('जॉर्डन', 2), ('क्रूज़', 2), ('पोल', 2), ('टोयोटा', 2), ('अलबामा', 2), ('जापानी', 2), ('प्रधानमंत्री', 2), ('अंडरवाटर', 2), ('लिंकिन', 2), ('स्पेनिश', 2), ('सियाना', 2), ('मिलर', 2), ('बाथ', 2), ('टार्जन', 2), ('स्मोक', 2), ('पांडा', 2), ('बेंज', 2), ('सैक', 2), ('सोंग', 2), ('होमलैंड', 2), ('डेनिस', 2), ('अर्लिंग्टन', 1), ('गॉट', 1), ('दा', 1), ('क्राउन', 1), ('लॉरी', 1), ('एंडरसन', 1), ('ऑस्कर', 1), ('यूट्यूब', 1), ('विद', 1), ('ब्लूम', 1), ('लिबरल', 1), ('पराग्वे', 1), ('कंस्ट्रक्शन', 1), ('एंजेल्स', 1), ('लुफ्थांसा', 1), ('ओल्ड', 1), ('वेव', 1), ('फायरिंग', 1), ('यूसी', 1), ('बर्कले', 1), ('प्यूजो', 1), ('लाइज', 1), ('एक्टिविस्ट', 1)]

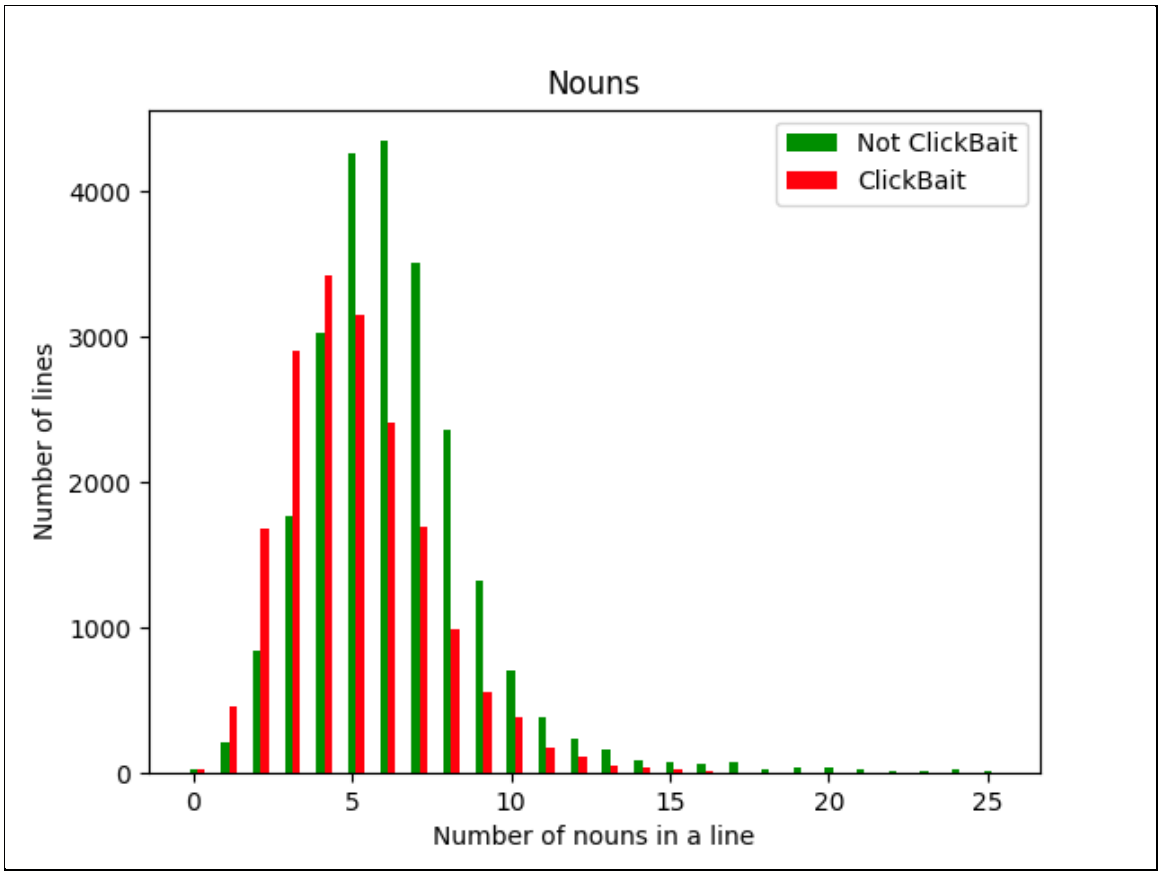
Observations Based On Above Word Clouds-

1. We quickly observe and analyse that, in all three NER classes, the words which contribute to a title being clickbait or not-clickbait are unique. There are a few minor similarities, but the majority of the words are common words are different.

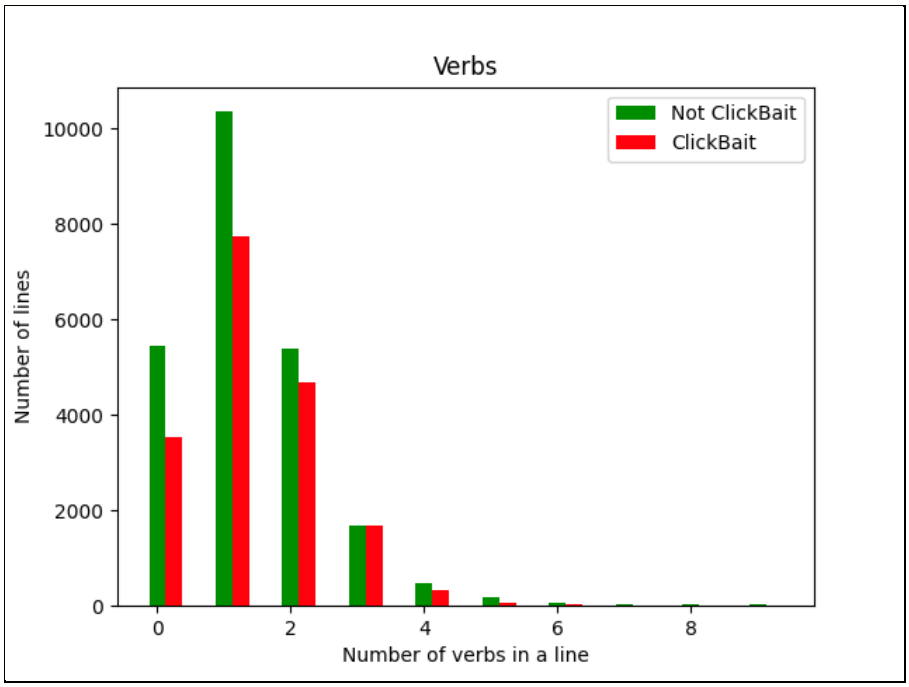
2. In consideration of PERSON based NER tags, we observe that Clickbait titles have more proper nouns in comparison to non-clickbait titles. Non-Clickbait titles use more common nouns in them.
3. This shows that people tend to be clickbaited by selecting and viewing those titles which have names of their favourite personalities or anything which can generate and give gossip to them.
4. In ORGANISATIONAL tags, Clickbait titles have a higher tendency to refer to professions or job seeking audiences, which proves that several clickbait titles are in fact those which bait people to study abroad or apply for jobs and such. Non-Clickbait titles shift towards nationality and in fact cover several more fields and organisational aspects in comparison to the former.

8. POS Tags

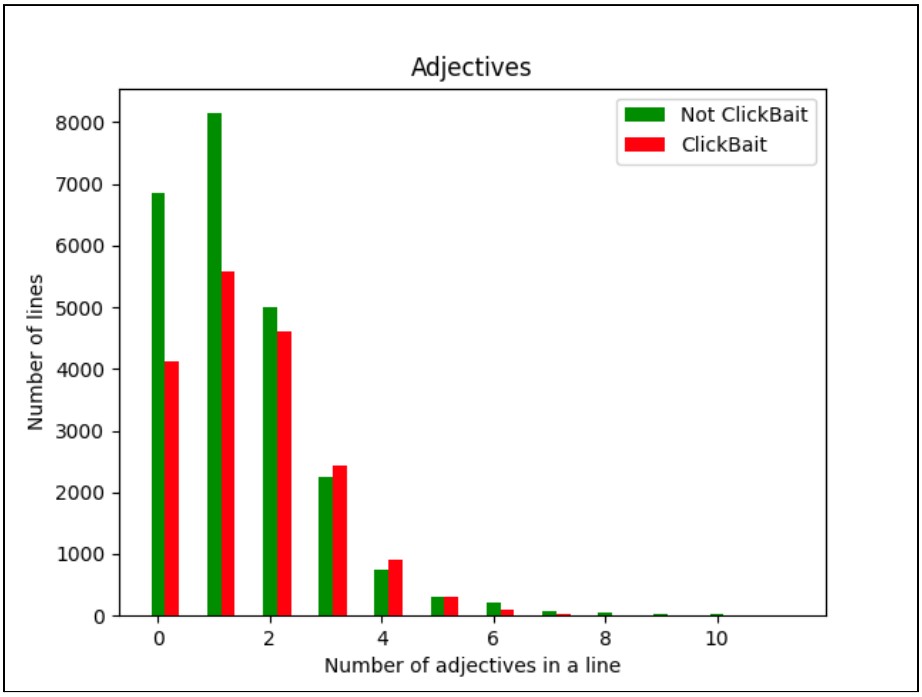
Under POS tags we focused on nouns, verbs and adjectives and bar graphs were plotted for each of them separately. The stanza library was used to POS tag all the data that was collected and the data obtained was simultaneously used to plot graphs and was added to a csv file for further classification.



In the graphs from plotting nouns, the X-axis plots the number of words in one line to the number of such lines in the Y-axis. As we can observe, in the lines with 4 or fewer nouns, there are more clickbait titles present. For lines with 5 or greater nouns, they are almost majorly Non-Clickbait titles. The largest set we get is the set of lines having 6 nouns in a line. There are almost double Non-Clickbait titles as Clickbait titles having 6 nouns per line. In the range 15-25 nouns in a line, they are almost exclusively Non-Clickbait titles.



As for the Verb graph, there are much fewer verbs in a line as compared to Nouns, with the maximum at around 9 verbs in a line. The X-axis represents the number of verbs in a line and the Y-axis represents the number of such lines. Here, the key observation is that Non-Clickbait titles almost always have more verbs as compared to Clickbait titles. There are also very few titles past 6 (almost negligible).



Coming to the adjectives graph, again, the X-axis represents the number of adjectives in a line and the Y-axis represents the number of such lines. Here, the max number of adjectives to a line comes to about 10. We have therefore limited the graph to this number on the X-axis. The result obtained shows us that under 3 adjectives in a line has Non-Clickbait titles as the majority. There are varied results in the 4-10 range with a small difference between the 2 classes.

Classifying The Data:

To analyse and observe the impact of the parameters we generated above, we feed them into various classifier models for training purposes and allow the model to make suitable predictions on other sentences. For generating accuracy, we take 80% of the training data for training the classifier models and the remaining 20% for verification purposes to check the accuracy of the model.

We divide our parameters generated columns into 3 classes:

A. Structural: is_que, num_tokens, is_exclam, is_quote, num_stopwords, has_num
B. NER: PER_count, ORG_count, LOC_count, total_NER
C. POS: num_noun, num_verb, num_adj

Classifier 1: SVMs

1. A+B+C
    a. Model accuracy score : 0.6988
    b. Precision Score : 0.7095
    c. Recall Score : 0.6727
    d. F1 Score : 0.6722
2. A+C
    a. Model accuracy score : 0.7068
    b. Precision Score : 0.7245
    c. Recall Score : 0.6790
    d. F1 Score : 0.6780
3. A+B
    a. Model accuracy score : 0.6813
    b. Precision Score : 0.6873
    c. Recall Score : 0.6559
    d. F1 Score : 0.6544
4. B+C

a. Model accuracy score : 0.6537
b. Precision Score : 0.6568
c. Recall Score : 0.6253
d. F1 Score : 0.6200

Classifier 2: Logistic Regression

1. A+B+C
    a. Logistic Regression Accuracy Score : 0.6783
    b. Logistic Regression Precision Score : 0.6679
    c. Logistic Regression Recall Score : 0.6587
    d. Logistic Regression F1 Score : 0.6596
2. A+C
    a. Logistic Regression Accuracy Score : 0.6874
    b. Logistic Regression Precision Score : 0.6842
    c. Logistic Regression Recall Score : 0.6676
    d. Logistic Regression F1 Score : 0.6686
3. A+B
    a. Logistic Regression Accuracy Score : 0.6711
    b. Logistic Regression Precision Score : 0.6688
    c. Logistic Regression Recall Score : 0.6488
    d. Logistic Regression F1 Score : 0.6483
4. B+C
    a. Logistic Regression Accuracy Score : 0.6451
    b. Logistic Regression Precision Score : 0.6230
    c. Logistic Regression Recall Score : 0.6237
    d. Logistic Regression F1 Score : 0.6225

Classifier 3: XGBoost

1. A+B+C
    a. XGBoost Accuracy Score: 0.6980
    b. XGBoost Precision Score: 0.7141
    c. XGBoost Recall Score : 0.6752
    d. XGBoost F1 Score : 0.6760
2. A+C

    a. XGBoost Accuracy Score: 0.7001
    b. XGBoost Precision Score: 0.7237
    c. XGBoost Recall Score : 0.6765
    d. XGBoost F1 Score : 0.6769
3. A+B
    a. XGBoost Accuracy Score: 0.6919
    b. XGBoost Precision Score: 0.7114
    c. XGBoost Recall Score : 0.6682
    d. XGBoost F1 Score : 0.6680
4. B+C
    a. XGBoost Accuracy Score: 0.6556
    b. XGBoost Precision Score: 0.6622
    c. XGBoost Recall Score : 0.6283
    d. XGBoost F1 Score : 0.6240

Observations:

1. The model accuracy is significantly less when we use only NERs and POS Tags as parameters for the classification process. This means that parameters that are based on number of stopwords, tokens and occurrence of numerals, question marks, exclamation marks and quotation marks are extremely important for the binary classification process.
2. Parameters based on structural components and POS Tags give better accuracy results in classification compared to the involvement of NER tags along with it. This is clearly seen in the slight improvement in the accuracy across ALL the 3 models.
3. Across all the 3 models used for binary classification, the order of parameters which gives the best accuracy are: A+C > A+B+C > A+B > B+C.

	SVM				Logistic Regression				XVG Boost			
	A+B+C	A+C	A+B	B+C	A+B+C	A+C	A+B	B+C	A+B+C	A+C	A+B	B+C
Accuracy	0.6988	0.7068	0.6813	0.6537	0.6783	0.6874	0.6711	0.6451	0.6980	0.7001	0.6919	0.6556
Precision	0.7095	0.7245	0.6873	0.6568	0.6679	0.6842	0.6688	0.6230	0.7141	0.7237	0.7114	0.6622
Recall	0.6727	0.6790	0.6559	0.6253	0.6587	0.6676	0.6488	0.6237	0.6752	0.6765	0.6682	0.6283
F1 Score	0.6722	0.6780	0.6544	0.6200	0.6596	0.6686	0.6483	0.6225	0.6760	0.6769	0.6680	0.6240

Conclusion And Future Work

In this paper, we compared the properties of Clickbait and Non-Clickbait titles and highlighted the differences between these two categories. We used these to design relevant features and parameters to classify the titles according to the above mentioned classes. We have made the data and source codes publicly available at https://github.com/esh04/Clickbait-Identification.git, so that others can view our contribution towards this field for the **Hindi language** and collectively improve the work. A natural extension to our work would be to reason out why a specific title is Clickbait or not. It also gives us a clear insight into what type of headlines are clickbait and how we can draw common ground inferences to not fall prey to such titles.

Reference: **https://www.researchgate.net/publication/323156528_Believe_it_or_not_identifying_bizarre_news_in_online_news_media**

-END-