

# CMPE-257-02 PROJECT: CLASSIFICATION OF CLICKBAIT HEADLINES

Team-1

Anjali Sreeja Kadiyala  
*Computer Engineering Department*  
*San Jose State University*  
anjalisreeja.kadiyala@sjsu.edu

Ashutosh Ojha  
*Computer Engineering Department*  
*San Jose State University*  
ashutosh.ojha@sjsu.edu

Samyukta Shetty  
*Computer Engineering Department*  
*San Jose State University*  
samyukta.shetty@sjsu.edu

Harisha Korapati  
*Computer Engineering Department*  
*San Jose State University*  
harisha.korapati@sjsu.edu

Kishore Kumar Natarajan  
*Computer Engineering Department*  
*San Jose State University*  
kishorekumar.natarajan@sjsu.edu

**Abstract**—Most online news media businesses rely significantly on money produced by reader clicks, and because there are so many of them, they must fight for reader attention. To entice readers to click on an article and then visit the media site, outlets frequently use appealing headlines alongside article links, enticing readers to click on the link. Clickbait headlines are what they are called. While these baits may tempt readers to click, in the long term, clickbait rarely lives up to the readers' expectations, leaving them disillusioned. In this paper, we detect if the headline is a clickbait or non-clickbait by using different Machine Learning Classification algorithms. Different text vectorization methods are used to convert text into vector format to proceed with the classification algorithm approach. We conclude that by using Kernel SVM model with the best parameters after hyperparameter tuning we get an accuracy of 95.05%. Our code can be found at <https://github.com/Anjali-Kadiyala/headline-classification>.

**Index Terms**—support vector machine, headlines, clickbait, classification

## I. INTRODUCTION

The media landscape is changing dramatically as people increasingly consume news online. This shift can be attributed to two main factors. First, unlike traditional offline media, where readers' allegiances to a particular newspaper were almost static, online media provides readers with a diverse range of options, ranging from local, national, and international media outlets to a variety of niche blogs focusing on specific topics of interest. Second, most online media sites do not charge a subscription fee, and the majority of their money comes from adverts on their web pages. In the modern era, every media organization must compete with a slew of other media outlets for reader attention, and readers' clicks are how they make money. As a result, they adopt numerous strategies to entice readers to visit the media site and click on a story, such as creating enticing headlines to accompany article links and enticing readers to click on the links.

Clickbait headlines are what they are called. "(On the Internet) information whose principal objective is to draw attention and

persuade users to click on a link to a particular web page," according to the Oxford English Dictionary.

Clickbait take advantage of the cognitive phenomena known as the Curiosity Gap [1], in which headlines provide forward referencing clues to pique readers' interest enough for them to feel driven to click on the link to fill the knowledge gap. While these baits may fool readers into clicking, in the long term, clickbait rarely lives up to the readers' expectations, leaving them unhappy. According to cognitive studies (such as [2]), clickbait facilitates attention diversion. As readers switch to new articles after being enticed by headlines, the attention residue from these frequent transitions causes cognitive overload, discouraging them from reading more exciting and in-depth news pieces. There are also concerns regarding journalistic gatekeeping in the changing media landscape with the prevalence of clickbait [3].

Despite the uproar over clickbait's negative impacts, there has been no effort to develop a systematic approach to a comprehensive remedy. Many clickbait can be found in the 'Promoted Stories' section at the end of articles on the websites of 'The Guardian' and 'The Washington Post.' according to the click-to-share ratio and the amount of time spent on these topics, Facebook announced in 2014 that they would eliminate clickbait stories from users' news feeds. Nonetheless, Facebook users continue to complain about clickbait, and there is a new effort to combat clickbait. In a recent study, Potthast et al. [4] sought to detect clickbaity posts on Twitter. The difficulty with such isolated approaches is that clickbait can be found on specific social media platforms and other well-known websites. As a result, we need a complete solution that works across the web. Some ad-hoc approaches have been developed, such as 'Downworthy' [5], which detects clickbait headlines using a fixed set of common clickbait phrases and then converts the headlines into something more garbage-ish, or 'Clickbait Remover for Facebook' [6], which prevents links to a fixed set of domains from appearing in users' news feeds.

The issue with having a fixed rule set is that it is not scalable and may require regular adjusting as new clickbait terms arise. Similarly, blocking links to a specific set of domains will also prohibit non-clickbait article links.

We take the first step towards building a comprehensive solution that can work across the web in this work. We first build a classifier that automatically detects whether a headline is clickbait or not.

## II. DATASET

We gathered much information in both the clickbait and non-clickbait categories.

**Non-clickbait:** We used NewsReader to extract headlines from a corpus of 18,513 Wikinews articles [7]. A community of contributors on Wikinews creates articles, and the community must vet each news piece before being published. Style rules specify how certain events should be documented and presented to readers. For example, when writing a story's title, the author must adhere to guidelines<sup>4</sup>. We consider the headlines of these stories to be the gold standard for non-clickbait because of Wikinews' stringent vetting.

**Clickbait:** We discovered the following domains that publish many clickbait articles: 'BuzzFeed,' 'Upworthy,' 'ViralNova,' 'Scoopwhoop,' and 'ViralStories.' During September 2015, we crawled 8,069 web items from these domains. To avoid false negatives (i.e., articles in these domains that are not clickbait), we enlisted the help of six volunteers who were asked to classify the headlines of these articles as clickbait or non-clickbait. We split the articles among the volunteers so that at least three people labeled each one.

With a Fleiss of 0.79, we achieved a substantial inter-annotator agreement. 7,623 articles were classified as clickbait based on the majority vote. The stories under BuzzFeed's 'news' section, most reported like traditional news, are notable examples of articles the volunteers marked as non-clickbait.

Finally, we randomly chose 7,500 items from both categories to ensure an equal sample of clickbait and non-clickbait content when comparing and developing the classifier.

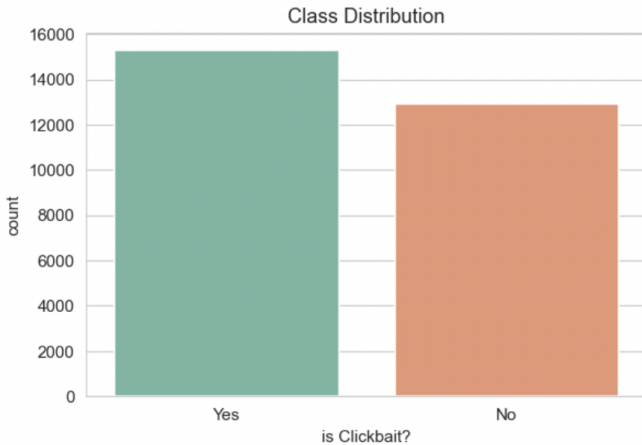


Fig. 1. Dataset class distribution

## III. RELATED WORK

The origin of clickbaits can be traced back to the advent of tabloid journalism. tabloid journalism, type of popular, largely sensationalistic journalism that takes its name from the format of a small newspaper, roughly half the size of an ordinary broadsheet. Tabloid journalism is not, however, found only in newspapers, and not every newspaper that is printed in tabloid format is a tabloid in content and style. Notably, many free local publications historically have been printed in tabloid format, and in the early 21st century several traditional British broadsheet newspapers, such as The Independent, The Times, and The Scotsman, changed to the smaller size, preferring, however, to call it "compact" format. On the other hand, one of the most-popular tabloids in Europe, the German Bild-Zeitung, was long printed as a broadsheet before shifting, as did many German newspapers, to a format that was smaller than a broadsheet but bigger than the standard tabloid. The origins of the term tabloid are disputed. According to the most-plausible explanation, the name derives from tablet, the product of compressed pharmaceuticals.

Tabloid journalism, which started focusing on 'soft news' compared to 'hard news'. soft news, also called market-centred journalism, journalistic style and genre that blurs the line between information and entertainment. Although the term soft news was originally synonymous with feature stories placed in newspapers or television newscasts for human interest, the concept expanded to include a wide range of media outlets that present more personality-centred stories. Traditionally, so-called hard news relates the circumstances of a recent event or incident considered to be of general local, regional, national, or international significance. By contrast, soft news usually centres on the lives of individuals and has little, if any, perceived urgency. Hard news generally concerns issues, politics, economics, international relations, welfare, and scientific developments, whereas soft news focuses on human-interest stories and celebrity.

There have been many researches in media studies highlighting the problems with tabloidization. In the paper "Obituary for the newspaper? tracking the tabloid,"<sup>5</sup>, Discussing newspapers in the 21st century commonly entails a narrative of impending extinction arising from technological, demographic, and cultural change. This article reports on research into three Australian newspapers (two broadsheets, one tabloid) that is concerned, in the first instance, with the concept of 'tabloidization', and the proposition that identifiable tabloid properties, such as the simplification and spectacularization of news, are increasingly characteristic of contemporary newspapers. Adaptive changes to newspaper design, style, and content in the interests of survival and renewal are addressed through quantitative content analysis in tracking formal changes to newspapers, and qualitative research through interviews with journalists in exploring their everyday negotiation of the role and trajectory of newspapers. These questions of industrial context, textual form, occupational practice, professional ideology, and politico-cultural judgment are raised in seeking to

understand the dynamics of the shifting forms and contested readings of contemporary newspapers through a critically reflexive analysis of tabloidization discourse and process[7]. The browser extension ‘Downworthy’ detects clickbait headlines using a fixed set of common clickbait phrases, and then converts them to meaningless garbage text. The problems with the above approaches are that they either work on a single domain, or the fixed ruleset does not capture the nuances employed across different websites. The clickbait headlines that websites like BuzzFeed, ViralNova and Upworthy use to drive traffic, especially through social networks. Even Huffington Post has jumped on the bandwagon of endless recycled listicles and bombastic titles [8].

Downworthy replaces hyperbolic headlines from bombastic viral websites with a slightly more realistic version. For example:

- “Literally” becomes “Figuratively”
- “Will Blow Your Mind” become “Might Perhaps Mildly Entertain You For a Moment”
- “One Weird Trick” becomes “One Piece of Completely Anecdotal Horseshit”
- “Go Viral” becomes “Be Overused So Much That You’ll Silently Pray for the Sweet Release of Death to Make it Stop”
- “Can’t Even Handle” become “Can Totally Handle Without Any Significant Issue”
- “Incredible” becomes “Painfully Ordinary”
- “You Won’t Believe” becomes “In All Likelihood, You’ll Believe

The paper “clickbait Detection” proposes a new model for the detection of clickbait, i.e., short messages that lure readers to click a link. Clickbait is primarily used by online content publishers to increase their readership, whereas its automatic detection will give readers a way of filtering their news stream. We contribute by compiling the first clickbait corpus of 2992 Twitter tweets, 767 of which are clickbait, and, by developing a clickbait model based on 215 features that enables a random forest classifier to achieve 0.79 ROC-AUC at 0.76 precision and 0.76 recall [9].

In the article, how different headlines influence individuals has been a long-term concern among journalists and academics. Past research has found that headlines can change perceptions of a criminal suspect’s supposed guilt, influence how individuals assess political candidates, and affect comprehension and memory of news articles. How headlines are worded and the issues headlines discuss also can affect individuals. In an earlier Center for Media Engagement white paper on headlines, Research Associate Alex Curry and Director Talia Stroud investigated how solutions-oriented headlines are received by audiences. These headlines mention the possible solutions to the problems facing communities. In general, headlines that discuss “solutions” lead to more digital clicks compared to non-solutions-based headlines.

The influence of headlines on perceptions of news events also can vary based on the issue discussed. For instance, prior

research has found that headlines can influence feelings about political candidates as well as change perceptions of racial violence. In science contexts, however, headlines have been found to have little effect on attitudes related to genetic determinism (e.g. “gene causes high blood pressure”). Although headlines can influence subsequent attitudes and beliefs, effects also can be highly contextualized based on the news issue.

The majors finding in this article are Question-based headlines lead to more negative attitudes about the headline and more negative expectations for the associated news story compared to traditional headlines, Forward-reference headlines did not lead to different reactions, expectations, or anticipated engagement compared to traditional headlines, Types of headlines and policy issues should be paired carefully; question-based headlines about Congress yielded the most negative reactions and News source brand (USA Today, BuzzFeed, and Fox News/MSNBC) matters for headline and story perceptions[10].

#### IV. COMPARING CLICKBAITS AND NON-CLICKBAITS

There are many factors involving word choices and sentence formation that play an important role in defining a heading as clickbait or non-clickbait. Linguistic analysis on these headlines can be carried on to “derive insights about the semantic and syntactic nuances that occur more frequently in clickbait headlines compared to the traditional non-clickbait headlines.” [11] Stanford coreNLP [12] is one such tool that performs agreeable analysis. The following section describes the various factors influencing the headlines and the historical analysis found on the dataset.

##### A. Sentence structure

1) *length of the headlines*: Its is found that the length of the headlines differs significantly in clickbait and non-clickbait headlines. Figure 4 shows the distribution of length in both categories.

We see that the conventional non-clickbait headlines are shorter than clickbait headlines. It is seen that the average length of clickbait and non-clickbait headings in 10 and 7 respectively.

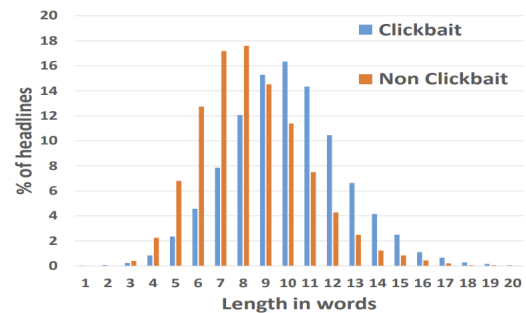


Fig. 2. Distribution of the length of both clickbait and non-clickbait headlines [11]

2) *Length of the words*: The average length of words is shorter in clickbait headings, even though the number of words is more in comparison to non-clickbait headings. For the dataset, it is found that the average word length of clickbait headlines is found to be 4.5 characters, and the average word length of non-clickbait headlines is 6. [1] The shorter word length is mainly due to frequent use of shorter function words and word shortenings like you'll, we're, we'd in clickbait headings. [11]

In Figure 5(a), about 0.6% of the traditional news headlines contain word shortenings, and 22% of clickbait headlines have such shortened words. [11]

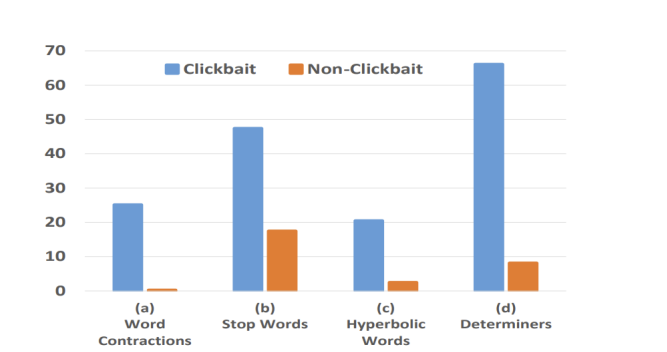


Fig. 3. Percentage of clickbait and non-clickbait headlines, which include (a) Word Contractions, (c) Hyperbolic Words, and (d) Determiners. (b) Percentage of words identified as stop words in both clickbait and non-click headlines.

3) *Length of the syntactic dependencies*: Stanford collapsed-coprocessor dependency parser [13] can be used to identify the syntactic dependencies between all pairs of words in the headlines as found by existing analysis on the dataset. It is then used to compute the distance between the governing and the dependent words in terms of the number of words separating them. It is found that clickbait headlines have longer dependencies' than non-clickbaits, due to the presence of complex phrases. [11]

#### B. Stop Words, Hyperbolic and Common Phrases

1) *Stop words*: The set of most common words used in any language are called stop words. They are very important in analysis because if we remove the stop-words, then we can focus on the important words present in any content. Figure 5(b) shows the percentage of stop words (in English) present in both categories of headlines. We can infer from the figure that stop words are used more frequently in clickbait headings (e.g. 45% compared to 18% in non-clickbaits) to complete the structure of the headlines [11]. In non-clickbaits, the usage of content words is high and stop word inference is done by the reader.

2) *Hyperbolic words*: Sentiment analysis on the words in headings can be done using the Stanford Sentiment Analysis tool [14].

Words that indicate very positive sentiments (e.g., Awe-inspiring, breathtakingly, gut-wrenching, soul-stirring, etc.) are called hyperbolic words. Such words are found in abundance in clickbait headings and are quite non-existent in non-clickbait headlines. The percentage of hyperbolic words present in both categories of headlines is shown in Figure 5(c).

3) *Internet slangs*: Internet slang words like WOW, LOL, LMAO etc are also found to be predominantly used in clickbait headings. These words, along with the use of hyperbolic words play an important role in capturing the user's interest.

4) *Punctuation patterns*: The informal punctuation patterns such as !?, ..., \*\*\*, !!! – are mostly commonly found in clickbait headings. Such patterns are rarely used in non-clickbait headings. [11]

5) *Common bait phrases*: Another common observation made in the click-bait headings is the use of phrases like s “Will Blow Your Mind”, “You Won’t Believe” which lures the reader to open the article.

#### C. Subjects, Determiners and Possessives

1) *Sentence subjects*: Subject words in headlines can be determined using the Stanford syntactic dependency parser [13]. Table 1 shows the most commonly occurring subject words in both clickbait and non-clickbait headlines.

Clickbait	I, you, dog, everyone, girls, guys, he, here, it, kids, men, mom, one, parent, photos, reasons, she, something, that, they
Non-clickbait	bomb, court, crash, earthquake, explosion, fire, government, group, house, U.S., China, India, Iran, Israel, Korea, leader, Obama, police, president, senate

TABLE I  
20 most commonly occurring subject words in both clickbait and non-clickbait headlines. [11]

The 20 most commonly occurring subject words found in both clickbait and non-clickbait headlines are listed in Table 1. There is a repetition of popular subjects in clickbait headings. “Nearly 62% of the clickbait headlines contained one of the 40 most common clickbait subject words. On the other hand, only 16% of the non-clickbait headlines contained the top 40 non-clickbait subject words.” [11]

2) *Determiners*: Determiners like their, my, which, these are often found to be employed in clickbait headings to reference someone or something. These words direct the user to a particular subject of interest and hence are more found in clickbaits headings than non clickbaits.

Figure 5(d) shows the percentage of headlines in both clickbait and non-clickbait headlines, where determiners are present. [11]

3) *Possessive case*: The reader is often addressed in the first and second person with the use of subject words I, We, You in clickbait headings. The third person references are common nouns like he, she, it, they, man, dog rather than specific proper nouns, which is in stark contrast with the non-clickbait headlines, as the reporting is always done in third person. [11]

## V. CLASSIFYING HEADLINES AS CLICKBAITS: OUR APPROACH

A comparative analysis of the headlines done in the related paper indicates prominent linguistic and structural differences between the clickbait and non-clickbait headlines. Based on this analysis, we align our approach to identify which characteristic varies the most.

Overall, our approach involves five primary phases:

1. Text Normalization
2. Text Vectorization
3. Model Selection
4. Hyperparameter Tuning
5. Final Modelling

### A. Text Normalization

As described in the earlier section, our dataset contains only two columns. The first column contains all the clickbait and the non-clickbait headlines. This is the only column which forms the independent feature set. The second column is the one that contains the labels for each corresponding headline as clickbait or not. This is the dependent variable and is binary in nature.

Before, the data can be fed into classification models, a standard cleaning and pre-processing is performed on the headline column. It is essential to mention that the dataset contained no missing values or duplicates and hence no datapoints of the original dataset were dropped.

The various other cleaning and pre-processing steps undertaken include:

1. Convert to lower case: We converted all the texts to lowercase with an intention to avoid case sensitivity.
2. Clean text: We further cleaned the headlines to remove

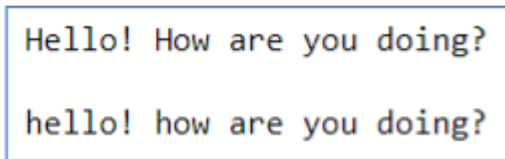


Fig. 4. Converting entire text to lower case.

numbers, punctuations, links link jargons (http, com, in, www) and other special characters.

3. Tokenization: Tokenization is the process of breaking

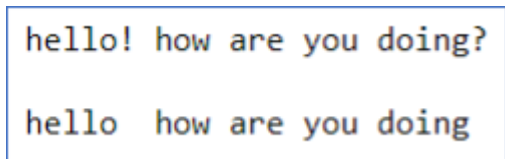


Fig. 5. Clean text

the raw text into small chunks: sentences or words. We have performed word tokenization to split the raw texts into

constituent words.

4. Stop words removal: Stop words are a set of commonly

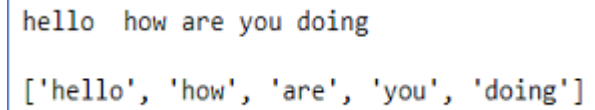


Fig. 6. Tokenization

used words in a language. In the context of Natural Language Processing and Text Mining, words like “a”, “the”, “is”, “are”, etc., can be easily ignored as they carry very little information. We used nltk library to remove the stop words from the headlines.

5. Part of Speech Tagging: We performed this step as a prerequisite to the next step. POS tagging refers to a process of grammatical tagging of a word in a text depending on the definition of the word and its context. In a nutshell, it takes in each word of a text, identifies its context in the text and tags it accordingly as a noun, verb, adjective, or adverb.

6. Lemmatization: Lemmatization is the process of grouping together words with the same root or lemma but with different inflections or derivatives of meaning. The objective is to reduce sparsity later by reducing the number of words for a particular root word. For example, for a root word: connect, all the other derivatives like connected, connects, connection, connecting would be replaced with the base word connect in all the texts. We have used nltk.stem.WordNetLemmatizer to achieve this objective and this lemmatizer requires the words to be post-tagged appropriately to be successfully lemmatized.

### B. Text Vectorization

Text Vectorization is the process of converting text into a numerical representation. There are numerous existing techniques of performing vectorization of text. Although, normalization and vectorization both have the same objective of transforming text into a format accepted by the models, we still have separated these into two different steps because we wanted to do a comparative study of a couple of vectorization techniques in the context of this classification task.

We have used three popular methods of text vectorization:

1. Bag of Words: This is a technique that converts text into finite length vectors where each column of a vector represents a word. The values in each cell of a row or text vector shows the number of occurrences of a word in the sentence. Bag of words is the most trivial representation of text into vectors that considers only the frequency of the words in each text.

2. TF-IDF: This technique is based on bag of words, which considers the relevance of words along with their frequency in a text. TF stands for Term Frequency, which in other words is the frequency of a word in a text. IDF stands for Inverse Document Frequency, which measures how frequently does a word occur in the entire corpus/dataset. The IDF value assigns a weight to each word where a less frequent word is given more relevance and vice versa.



```
In [13]: stop_words = stopwords.words('english')
print(stop_words)

['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've", "you'll", "you'd", 'your', 'yours', 'y',
'yourself', 'yourselves', 'he', 'him', 'his', 'himself', 'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself',
'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', "that'll", 'these', 'those',
'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do', 'does', 'did', 'doing', 'a', 'an',
'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about', 'against', 'b
etween', 'into', 'through', 'during', 'before', 'after', 'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'o
f', 'over', 'under', 'again', 'further', 'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both',
'each', 'few', 'more', 'most', 'other', 'some', 'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very',
's', 't', 'can', 'will', 'just', 'don', "don't", 'should', "should've", 'now', 'd', 'll', 'm', 'o', 're', 've', 'y', 'ain', 'ar
en', 'aren't', 'couldn', "couldn't", 'didn', "didn't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn', "hasn't", 'haven', "have
n't", 'ish', 'ish't', 'ma', 'mightn', "mightn't", 'mustn', "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "should
n't", 'wasn', "wasn't", 'weren', "weren't", 'won', "won't", 'wouldn', "wouldn't"]
```

Fig. 7. Stop words

Vectorization Used	Naïve Bayes					Linear SVM					Kernel SVM					XGBoost				
	Acc.	Pre.	Rec.	F1	roc-auc	Acc.	Pre.	Rec.	F1	roc-auc	Acc.	Pre.	Rec.	F1	roc-auc	Acc.	Pre.	Rec.	F1	roc-auc
Bag of Words	0.91	0.87	0.97	0.91	0.90	0.93	0.93	0.94	0.94	0.93	0.94	0.95	0.94	0.95	0.94	0.82	0.93	0.73	0.81	0.83
TF-IDF	0.91	0.88	0.96	0.92	0.91	0.94	0.95	0.94	0.95	0.94	0.95	0.96	0.95	0.96	0.95	0.82	0.92	0.72	0.80	0.82
Word2Vec	0.67	0.76	0.54	0.63	0.67	0.84	0.91	0.79	0.84	0.85	0.78	0.89	0.68	0.77	0.79	0.83	0.87	0.80	0.83	0.83

TABLE II

Comparison between TF-IDF, Word2Vec, Bag of Words representation of data.

```
['hello', 'how', 'are', 'you', 'doing']
['hello', 'how', 'be', 'you', 'do']
```

Fig. 8. Lemmatization

3. Word2Vec: This is an advanced technique of text vectorization. The other two techniques discussed above fail to capture the context or semantic meaning of words in the texts. They focus primarily on the words and frequency of their occurrences. Word2Vec on the other hand tries to capture the semantic meaning of words with respect to the context of the corpus. Word2Vec achieves this by using a neural network to find vector representations for each word instead of just a single value. These word vectors define a word based on a set of 100 features. The working of Word2Vec in depth is well beyond the scope of this paper.

To have a good comparison between these three vectorization techniques, we performed each one of these method on the normalized corpus/dataset to produce three different datasets. Each of the datasets were then split into 3 sets: train, validation, and test sets in the ratio 60:20:20. In the next phase, we would be training models on each of the three vectorized datasets and compare performance.

### C. Model Selection

After normalizing and vectorizing the dataset, we ended up 3 representations of the dataset based on the vectorization technique used. We trained 4 classification models on each of

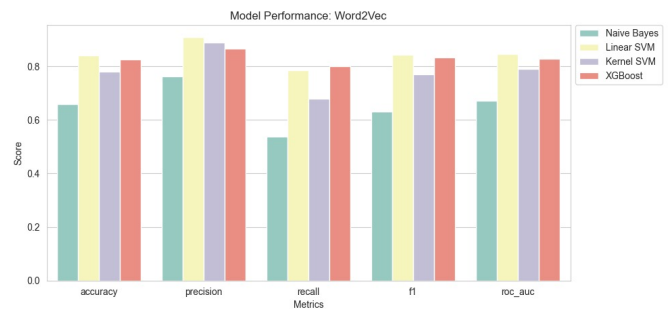


Fig. 9. Word2Vec model performance

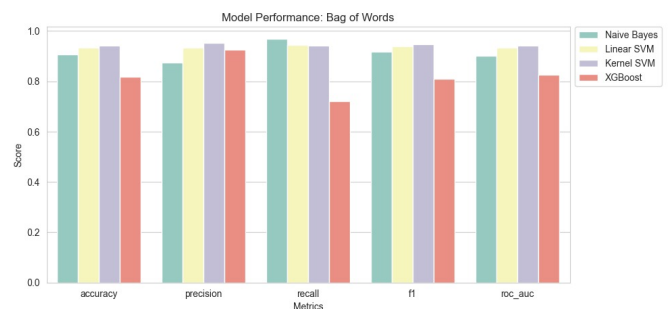


Fig. 10. Bag of words model performance

the 3 vectorized dataset resulting in a total of 12 combinations. Each of these combinations was trained on the training set and tested using the validation set.

The models used are:

1. Gaussian Naïve Bayes

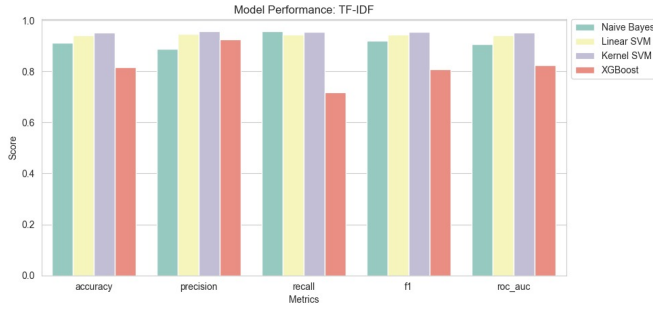


Fig. 11. TF-IDF model performance

2. Linear SVM
3. Kernel SVM
4. XGBoost

Based on the performance of the various models and vectorized dataset combinations, as evident from Table 1, Kernel SVM and TF-IDF combination proves to be the best performing combination.

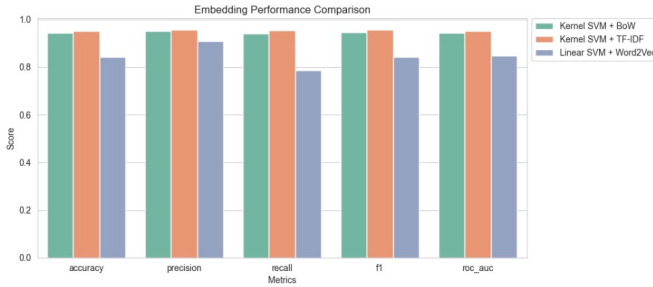


Fig. 12. Performance comparison

#### D. Hyperparameter Tuning

It is essential to note that the models used in the model selection phase were initialised with default parameters. This means that there might still be a scope of improvement in the models' performance.

Based on the model selection phase, we observed that Kernel SVM + TF-IDF combination performed the best among all the 12 combinations. Hence, in this phase, we try to boost up the performance of the combination. We achieve this by finding out the best set of parameters for the model using GridSearchCV.

GridSearchCV is a library function that is a member of sklearn's model\_selection package. It helps to loop through a set of possible hyperparameters and determine the best sets of values for the hyperparameters based on the training set provided. We picked roc\_auc score for determining the best parameters for the model. The cross-validation parameter for GridSearchCV was set as 5 to ensure sufficient validation in the search process.

We chose to determine two primary hyperparameters for the model: 1. C: Also known as the regularization parameter. We wanted to experiment with a couple of regularization values

to find the best bias-variance trade-off. The default value is 1.0.

2. kernel: The default kernel used is rbf. We wanted to test if any other kernels like poly or sigmoid would work better in this scenario. After GridSearchCV, the best parameters turned out to be the default values themselves, i.e.,  $C=1.0$ , `kernel='rbf'` with a roc\_auc score of 98.8% on the training set.

#### E. Final Modeling

It's crucial to mention that until now, only 60% of the original dataset had been used for training in the model selection phase. Another 20% of the dataset was being used for validation purposes. The remaining 20% of the dataset was extracted and kept untouched to serve as test set for generalization testing.

For the final phase, the training and validation sets were combined, and the hyperparameter-optimized model was trained on this 80% of the dataset.

Post training the model, the test set was used for the first time to evaluate the performance of the model, thus ensuring a more generalized evaluation.

### VI. RESULTS

One of the main motivations to pick this project was to perform a comparative study between different vectorization techniques, i.e., techniques to convert from text to vectors or numbers. The different techniques compared are bag of words, TF-IDF, and Word2Vec. Our initial hypothesis was that Word2Vec would perform the best. However, Word2Vec performed the worst.

Another result that can be drawn from this project is the difference between the paper's approach and our approach. The paper's approach has defined an extra set of engineering features in order to do the classification, whereas we did not. Our method is better as it gives us better accuracy. (Paper accuracy 93%, ours 95.05%)

Acc.	Pre.	Rec.	F1	roc_auc
0.95	0.96	0.95	0.96	0.95

TABLE III  
Results of the final model

### VII. CONCLUSION

This project classified clickbait and non-clickbait headlines using support vector machines. We analyzed the clickbait and non-clickbait headlines and found many fascinating distinctions between the two. We discuss the major steps of classification headlines as clickbait: text normalization, text vectorization, model selection, hyperparameter tuning, and final modeling. We delve deeper into the details of each of these steps, which indicate the work done prior to having our final model ready. With the best model, the kernel support vector

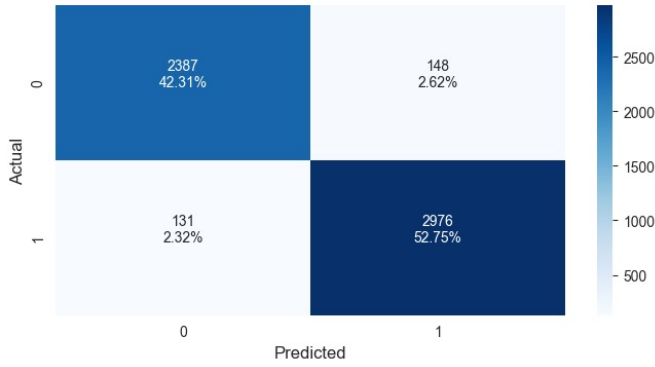


Fig. 13. Confusion matrix

machine, after hyperparameter tuning, we get an accuracy of 95.05% and F1 score of 95.52%.

### VIII. FUTURE SCOPE

We see the potential of using state-of-the-art deep learning architectures to perform the classification. With models like BERT, this task would likely be much faster and more efficient. Another possible improvement could be with using more advanced vectorization techniques.

### REFERENCES

- [1] G. Loewenstein, "The psychology of curiosity: A review and reinterpretation." Psychological bulletin, vol. 116, 1994.
- [2] G. Mark, "Click bait is a distracting affront to our focus," [nytimes.com/roomfordebate/2014/11/24/you-wont-believe-whatthese-people-say-about-click-bait/click-bait-is-a-distracting-affrontto-our-focus](https://nytimes.com/roomfordebate/2014/11/24/you-wont-believe-whatthese-people-say-about-click-bait/click-bait-is-a-distracting-affrontto-our-focus).
- [3] J. Dvorkin, "Column: Why click-bait will be the death of journalism," [pbs.org/newshour/making-sense/what-you-dont-know-aboutclick-bait-journalism-could-kill-you/](https://pbs.org/newshour/making-sense/what-you-dont-know-aboutclick-bait-journalism-could-kill-you/).
- [4] M. Potthast, S. Kopsel, B. Stein, and M. Hagen, "Clickbait detection," in Advances in Information Retrieval. Springer, 2016.
- [5] A. Gianotto, "Downworthy: A browser plugin to turn hyperbolic viral headlines into what they really mean," [downworthy.snipe.net/](https://downworthy.snipe.net/).
- [6] W. Markus, "Clickbait remover for facebook," [chrome.google.com/webstore/detail/clickbait-remover-forfacebook/hkbhmlgcpmneffdammbemapiiiniagj](https://chrome.google.com/webstore/detail/clickbait-remover-forfacebook/hkbhmlgcpmneffdammbemapiiiniagj).
- [7] D. Rowe, "Obituary for the newspaper? tracking the tabloid," Journalism, 2011.
- [8] A. Gianotto, "Downworthy: A browser plugin to turn hyperbolic viral headlines into what they really mean," [downworthy.snipe.net/](https://downworthy.snipe.net/).
- [9] M. Potthast, S. Kopsel, B. Stein, and M. Hagen, "Clickbait detection," in Advances in Information Retrieval. Springer, 2016.
- [10] J. Dvorkin, "Column: Why click-bait will be the death of journalism," [pbs.org/newshour/making-sense/what-you-dont-know-aboutclick-bait-journalism-could-kill-you/](https://pbs.org/newshour/making-sense/what-you-dont-know-aboutclick-bait-journalism-could-kill-you/).