# Αναγνώριση Clickbait Βάση Γλωσσικών Χαρακτηριστικών και Random Forests

Νίκος Καπέλλας1, Σαράντος Καπιδάκης1

1 Πανεπιστήμιο Δυτικής Αττικής, Τμήμα Αρχειονομίας, Βιβλιοθηκονομίας και Συστημάτων Πληροφόρησης, Ερευνητικό Εργαστήριο Διαχείρισης της Πληροφορίας, {nkapellas, sarantos}@uniwa.gr

# Clickbait Detection Using Linguistic Features and Random Forests

Nikos Kapellas1, Sarantos Kapidakis1

1 University of West Attica, Department of Archival, Library and Information Studies, Information Management Research Lab, {nkapellas, sarantos}@uniwa.gr

# Περίληψη

Σήμερα, η πλειονότητα των διαδικτυακών εφημερίδων βασίζεται στη δημιουργία εσόδων από τα κλικ των χρηστών και την προβολή διαφημίσεων. Με τα πολυάριθμα ειδησεογραφικά μέσα να ανταγωνίζονται για ένα κομμάτι του μεριδίου της αγοράς, παρατηρείται μια έντονη προσπάθεια για την ‘αρπαγή’ της προσοχής των αναγνωστών. Οι οργανισμοί αυτοί, συχνά δημιουργούν ελκυστικούς τίτλους ώστε να πείσουν τους χρήστες να κάνουν κλικ στα άρθρα που θα τους οδηγήσουν στις ιστοσελίδες τους. Αυτοί οι τίτλοι, που είναι κατασκευασμένοι για να τραβούν την προσοχή, είναι γνωστοί ως clickbait. Και ενώ οι πρακτικές αυτές εκ πρώτης όψεως φαίνονται επικερδής, μακροπρόθεσμα τείνουν να υποβαθμίζουν την ποιότητα των άρθρων και να οδηγούν στην ανεπιτυχή ικανοποίηση των πληροφοριακών αναγκών του κοινού. Από την άλλη, οι επιστήμες της βιβλιοθηκονομίας και της πληροφόρησης, έχουν παράδοση στην πληροφοριακή παιδεία και τη διαφύλαξη της ποιοτικής πληροφορίας. Έτσι, στόχος της παρούσας μελέτης είναι αφενός να εξετάσει και να αξιολογήσει τα γλωσσικά κυρίως χαρακτηριστικά που συνθέτουν τους τίτλους clickbait, και αφετέρου να παρουσιάσει μια πρόταση για την ενίσχυση και τη διεύρυνση του ρόλου των βιβλιοθηκών. Έτσι, κατασκευάζεται ένα σύνολο δεδομένων αποτελούμενο από άρθρα ειδήσεων, και γίνεται χρήση ενός μοντέλου Random Forests, για την αυτόματη αναγνώριση τίτλων clickbait. Το μοντέλο είχε εξαιρετική απόδοση, καθώς είχε αποτέσμα 1.00 στις μετρικές accuracy, precision και recall, κάτι που επιδεικνύει την αποτελεσματικότητα της μεθόδου. Η βαρύτητα των χαρακτηριστικών που χρησιμοποιήθηκαν στο μοντέλο δείχνει ότι μεταβλητές σχετικές με το συναίσθημα δεν ήταν σημαντικές, ενώ μερικές κειμενικές-γλωσσικές μεταβλητές είχαν το μεγαλύτερο αντίκτυπο.

**Λέξεις -κλειδιά:**

Αναγνώριση Clickbait, Random Forests, Επεξεργασία Φυσικής Γλώσσας, Ταξινόμηση, Πληροφοριακή Παιδεία, Επιστήμη της Πληροφορίας

# Abstract

Nowadays, most online newspapers rely on generating revenue from user clicks and displaying ads. With numerous news media competing for a slice of the market share, there is a strong effort to 'grab' the attention of readers. These organizations often create attractive headlines to persuade users to click on articles that will lead them to their websites. These titles, which are constructed to attract attention, are known as clickbait. And while these practices may seem profitable at first glance, in the long run they tend to degrade the quality of articles and lead to meeting unsuccessfully the informational needs of the audience. On the other hand, library and information sciences have a tradition to information literacy and preserving quality information. Thus, the aim of this study is both to examine and evaluate the mainly linguistic features that make up clickbait titles, and to present a proposal on enhancing and expanding the role of libraries. Thus, a dataset consisting of news articles is constructed, and a Random Forests model is used to automatically identify similar titles. The model performed extremely well, achieving a score of 1.00 in three metrics, accuracy, precision, and recall, demonstrating the effectiveness of the methodology. The importance of the features used in the model showed that variables related to sentiment were not important for prediction, while a few textual-linguistic variables were most impactful.

**Keywords:**

Clickbait Detection, Random Forests, Natural Language Processing, Classification, Media Literacy, Information Science

# Introduction

The digital transformation of news services has changed how news is written, circulated, and consumed. Currently, the online news ecosystem enables the fast dissemination of unverified, propaganda news stories and their mass consumption via social media platforms such as Facebook, messaging applications such as Viber, and Messenger, and multimedia platforms such as YouTube. Even though fake news and propaganda existed long before the use of the Internet, the current ads-based economic model has encouraged the spread of misinformation and disinformation, a phenomenon with significant impact on societies. This economic model dictates that online newspapers or journals do not operate under a subscription fee model for their content to be accessed. Instead, they use advertisements from providers and demand fees based on the visitor traffic on their pages, such as the more people visit a news website, the bigger the revenue. Therefore, advertisers and news organizations have adopted questionable methods for attracting traffic, leading to carelessly produced, inferior articles. Most notably, clickbait is a widely used term that has risen to describe such practices; content that is generated with the purpose of luring audiences into clicking on it, rather than any informative purpose (Molyneux & Coddington, 2020).

Consequently, clickbait news articles tend to blur the lines between fact and fiction. News articles are now perceived not for their credibility or reported facts but rather are perceived based on how their title is phrased, drawing attention to specific facts that captivate readers (Cofaru & Groza, 2022). These articles will be later recalled not because of the details in their content, but because of their sensational headlines. Another potentially hazardous effect is known as “belief perseverance” or “continued influence effect” (Wilkes & Leatherbarrow, 1988). According to it, reasoning that is based on facts that have been shown to be false remains intact until an alternative line of reasoning is offered (Nyhan & Reifler, 2015. This can be explored further, to study the impact of misinformative news in forming an opinion regarding numerous socio-economic news. Also, the authenticity of news seems to play an important role in the news consumption. According to a report of (Marchi, 2012) the value that news authenticity has, explains why for example teenagers prefer blogs, satirical shows, or anything other than traditional media, which they consider identical.

At the same time, it has been noted (Bazaco et al., 2019) that the number of search engine queries for news has been decreasing. The information overload has led to the aggregation and curation of news, such that audiences do not actively seek to read the news but rather follow news stories and their developments uncontextualized via news aggregators, their social media feeds, posts by friends, relatives, or influential figures. As a result, the medium as a textual unit, a paratext according to (Genette, 1997), disappears and each news story competes on its own, against a blend of news articles and other shared content, that is not necessarily journalistic in nature. Thus, the traditional gatekeeping role of the press is weakened, while citizens decide which stories to discuss and share.

Moreover, modern analytics tools allow publishers to know the exact preferences of their audiences. This way, news publishers can adjust their agenda to reflect the interests of citizens (audience agenda). Thus, what users prefer on these social platforms ends up becoming news. This way, important news can be ignored more easily, only to make room for popular news stories of little significance. In other words, the media focus on detecting successful news and then they incorporate these stories into their news feeds. Research has shown (Kapellas & Kapidakis, 2022), that many times groups of news media share the exact same news stories with little or no modifications at all. News media also change their agenda based on popular international stories, as these are more likely to become successful front-page news to small local publishers as well.

Additionally, many times the media deliberately create content that is designed to become viral, even if many times it attracts negative attention. By republishing attractive stories, the news piece loses its fresh status, but it can still be profitable if it redirects traffic to the website. This way, the presumption of commerciality is placed above the informational value of the article and audience satisfaction. Another study (Paor, 2020) reveals that information overload has produced a so-called post-truth society in which citizens tend to appreciate information that confirms their beliefs and ideologies (confirmation bias), rather than attempting to challenge their own views. As uncertainty grows people are more likely to value familiar facts than reliable ones (Gallagher, 2016).

Traditionally, titles serve some basic purposes that clickbait often disregards; communicate the subject of the news to the reader, summarize it, and gain the reader’s attention and trust. According to (Dor & Daniel, 2003) journalists typically use a short, clear, interesting headline that emphasizes new information, popular people, and connecting to prior events or reports. Clickbait titles on the other hand are deliberately ambiguous, bold, or exaggerated, trying to oversell the actual news attached to that title.

The manipulative power of clickbait has been attributed to the curiosity gap theory (Loewenstein, 1994). This theory suggests that a knowledge gap is created between the reader and the news article, by withholding vital information from a teaser title. Psychologically, this gap causes an urge to be filled which can be explained as a form of curiosity. However, even though the information gap theory is widely referenced to explain the effects of clickbait headlines, it has not been proven scientifically, thus it is still debatable how such short, exaggerated messages can exploit human cognition, as well as the scale in which they do so (Potthast et al., 2016).

Nonetheless, curiosity creates a rather uncomfortable sensation that can be met by exploratory behavior. Thus, the basic purpose of satisfying a provoked curiosity is to induce pleasure. This anticipation of pleasure, the promise of future rewards, could be the driving force behind the success of clickbait (Venneti & Alam, 2018) (Potthast et. al, 2018). Another study (Beleslin et al., 2017) discusses that behind the clickbait style of writing, there is a pleasure delayer (so-called reverse pyramid) mechanism, that tempts the audience to read the whole page of news, down to its last line, to provide the desired information. On the other hand (Knobloch et. al, 2004) found that the linear type of narrative evokes more suspense than the inverted type, and reading enjoyment is at a maximum when narration is prepared in the linear form.

In (Munger et. al, 2020) authors investigated and dismissed, based on their findings, the hypothesis that clickbait headlines have a negative effect on polarization, trust in the media, and information retention. They conducted an experiment with two different participant pools, to evaluate the individual-level preference for clickbait, and randomly assigned sets of participants to read either clickbait or traditional headlines. Another research that questions the effectiveness of clickbait headlines is (Molina, 2021). In it, the author presents research that moves in the opposite direction. They argue that clickbait in general is operationalized in different ways in each research, thus results are contradictory and inconclusive. Most of the relevant studies utilize different characteristics of clickbait. For example, (Molyneux & Coddington, 2020) (Scacco & Muddiman, 2020) compared user engagement and perceptions between non-clickbait headlines and clickbait characteristics such as the use of interrogative pronouns (who, what, when, where, why) and the question-as-title headlines (i.e. What are the benefits of early running?).

Other characteristics that have been studied (Chakraborty et al., 2016) (Rony, Hassan, & Yousuf, 2017) (Venneti & Alam, 2018) related to clickbait headlines, are the use of demonstrative adjectives (this, that, these, those), hyperboles, exclamation and question marks, length, and structure of the headline and important or popular events and figures. Similar characteristics are often used in relevant tasks as well, such as event detection in news articles (Kapellas & Kapidakis, 2023). Even though in general clickbait receives more engagement than non-clickbait, in some cases users engage more with non-clickbait content. Other forms of clickbait include articles that supposedly contain lists of things, so-called “listicles” (Vijgen, 2014). These present a common structure as the titles usually start or contain a cardinal number (“3 phrases that a couple’s therapist will never say to her partner”), and bold adjectives or nouns that convey authority. Another researched feature (Potthast et al., 2016) is phoricity in headlines. More specifically two forms of forward-references that can be identified at the discourse level and phrase level are deixis (“This news will blow your mind.”) and cataphora (“This name is hilarious.”), which are used to provoke the reader’s curiosity. Or for example the headline (“He loves Beatles, menthol cigs... and longs for muscles like Van Damme [sic]”) is cataphoric in the sense that “He” refers to a name that is located within the text (Blom & Hansen, 2015).

The research of (Chen et al., 2015) proposed that automatic identification of clickbait could utilize three distinct types of features: a) lexico-semantic and pragmatic linguistic patterns (e.g. unresolved pronouns, affective and suspenseful language, action words, overuse of numerals, and reverse narratives), b) not fitting image placement with a possible emotional load, and c) user reading and commenting behavior. Another similar study (Chen & Rubin, 2017) suggests that the methods for detecting clickbait can include lexical features (word level), complex language, and grammatical features, to the news genre classification. The vocabulary used (lexics) in the title can influence the perception of the headline. (Lex, Juffinger, & Granitzer, 2010) found a list of stylometric features of text (i.e., parts of speech, word length, and subjective terms) that enable the distinction between “yellow” and high-quality journalistic formats of news articles. In (Chakraborty et al., 2016) authors used the following list of features to develop a browser extension that alerts users about possible clickbait: the length of headlines, length of words the deadline consists of, the length of the syntactic dependencies, extremely positive words (hyperboles), internet slangs, punctuation patterns, common bait phrases, sentence subjects, determiners, possessive cases, word n-grams, repetition, part of speech categorization, and syntactic n-grams.

In this dynamic environment, library professionals can bridge the gap between technological innovation and information literacy, reinforcing the importance of responsible digital citizenship in the age of information overload. It is apparent that evaluating resources is becoming increasingly challenging, especially for younger audiences. Unarguably a shift is needed in the way resource searching and evaluation is taught to students in information literacy programs. The information-literate individual must be able to distinguish between different types of news such as fake news, biased or propaganda articles, clickbait, misinformation, rumors, satire, advertisements, or accurate and verified news stories, among others.

The American Library Association (ALA) created an “Information Literacy Competency Standards” for higher education to assess the skills of individuals at the college and university levels. According to these standards, the vast amount of information can be used effectively only with a complementary set of necessary abilities, as it will not self-create a more informed citizenry (Neely-Sardon & Tignor, 2018). Information literacy has been identified as one of the most effective methods to combat fake news (Batchelor, 2017). Critical thinking skills are of major importance on that path, and librarians are recognized as experts and specialists in teaching this set of skills (Eva & Shea, 2018). On the same note, newly coined literacies, such as media literacy, algorithmic literacy (Dogruel, Masur, & Joeckel, 2022), and meta literacy, provide a framework to specifically address the skills to actively participate and manage information in complex multimedia environments. Meta literacy (Mackey & Jacobson, 2010) can be defined as the combination of media literacy, digital literacy, visual literacy, cyberliteracy, transliteracy, and information fluency. On the other hand, algorithmic literacy can be defined as being aware of the use of algorithms in online applications, platforms, and services, knowing how algorithms work, being able to critically evaluate algorithmic decision-making, as well as having the skills to cope with or even influence algorithmic operations.

Involved parties must gain an understanding as to in which ways the information flow is changed by the algorithm-driven platforms students and teenagers are using and the technologies and social incentives that shape the circulation of news and information in societies today. Algorithms are not inherently good or bad. Instead, their impact depends on what they are programmed to do, why they operate this way or another, how users interact with them on social platforms, how they can be manipulated to their benefit, and lastly, to acknowledge the way algorithms feed on personal data to operate.

In other terms, information literacy, not to be confused with library instruction (Head et. al, 2020), can be described as an effort of librarians, information scientists, and educators, to provide the practices necessary to prepare students to discover and interpret information ethically, to experience how automated decision systems operate to recommend news, information, or other services (Sahoo et al., 2021). In (Lim, 2020) the author examined the content of 21 academic institutional library guides to explore how fake news is perceived by librarians and which strategies are used for detecting them. Findings suggested that library professionals need to pay attention to psychological factors when evaluating facts in their strategies relevant to news sources and fake news. In (Faix & Fyn, 2020) the authors read the “Framework for Information Literacy for Higher Education” of the Association of College and Research Libraries (ACRL) and proposed that librarians should take an all-inclusive approach to the misinformation problem and promote critical thinking by incorporating concepts from that framework.

In the following section the methodology followed is described. Results are presented and findings, limitations and future work are discussed.

# 1 Related Work

In (Geckil, 2018), data were gathered from new websites and Twitter. The data was tagged as clickbait or non-clickbait and compared with the TF-IDF method. Based on that comparison an incoming headline was classified as clickbait or non-clickbait, using a reliability index. Similarly, in (Potthast et al., 2018) authors constructed a news corpus of annotated Twitter tweets. To avoid biases, tweets were sampled from the top 27 most retweeted news publishers, covering a 5-month period. This corpus has been used to evaluate 12 clickbait detectors submitted to the Clickbait Challenge 2017. The research of (Zhou, 2017) proposed a self-attentive network for clickbait detection. They treat the regression task as a multi-classification problem, where a token-level self-attentive mechanism is applied to the hidden states of bi-direction recurrent units. This mechanism, which ranked 1st in the final evaluation of the Clickbait Challenge 2017, enables the generation of task-specific vector representation for tweets by attending important tokens. Another Twitter-related research (Potthast et. al, 2016), in which authors constructed the first clickbait corpus of nearly 3.000 Twitter tweets and they developed a 215-feature clickbait model that enables a Random Forest classifier with satisfactory results.

In (Rubin et al., 2019) authors developed a browser system that offers automatic detection and highlighting of clickbait, satirical fake news, and fabricated news articles. The proposed algorithm is based on finding patterns of subtle lexico-syntactic features in text and the reported accuracy of these detectors varies considerably given the variety of styles, formats, and content of online news. In (Cofaru & Groza, 2022) researchers experiment and analyze three feature sets to address automatic clickbait classification. They make use of word embeddings to observe relationships among words within the headline, investigate the language that is used to tempt users based on information extracted from the headline, and analyze how a combination of the previous features affects the performance of the model.

On the other hand (Molyneux & Coddington, 2020) focuses on understanding what audiences think of news aggregation and clickbait. Their study uses published original and aggregated news articles as stimuli in two online experiments to test readers' perceptions. Regarding clickbait, their results suggest that clickbait headlines may lower perceptions of the credibility and quality of that news and the news media. Similarly, (Molina, 2021) argues that users do not reliably click more often on headlines classified as clickbait by automated classifiers and they conduct three studies; a quasi-experiment using headlines classified as clickbait by three machine-learning models, a controlled experiment varying the headline of an identical news story to contain a single clickbait feature, and a computation analysis of four clickbait classifiers using real-world data. Their results point out that clickbait did not generate more curiosity than non-clickbait headlines and that while some headlines generate more engagement, these headlines cannot be safely classified as clickbait due to the low agreement between classifiers.

In (Bourgonje et al., 2017) authors present a system to detect the stance of the headlines with respect to their corresponding article bodies. This system is part of a larger platform for the curation of digital content, and it can aid the community as a tool that detects news articles that need a more thoughtful approach, where human intervention might be needed to decide the news's legitimacy or credibility. In (Beleslin et al., 2017) authors present the results of quantitative research, in which results confirm the hypothesis that there is a negative attitude towards clickbait titles. This finding is linked to the time spent on news consumption as well as the frequency of reading news on the internet overall. Findings showed that clickbait titles are most frequent in entertainment and lifestyle news sections, while they are less frequent in politics. In (Bazaco et al., 2019) researchers conduct an empirical content analysis, to check the presence of clickbait articles posted on Facebook and Twitter in two French newspapers. They detected a high average of clickbait content as well as the prevalence of the clickbait techniques used such as incomplete information, the predominance of soft news, repetition, serialization, and use of hyperboles in those headlines. In (Chen & Rubin, 2017) which is a study in progress, authors use a Q-methodology approach to perceive the subjectivity around clickbait. Study participants were asked to evaluate clickbait headlines, taken from BuzzFeed in 2016, on a 3-point scale and then rank them into two categories, clickbait, or not-clickbait, according to their perception of the term. Participants were also interviewed and allowed to comment on their decisions and these results showed that they indicated several clickbait features such as profanity, forward referencing, and colloquial phrasing. In terms of content, the participants were more likely to classify soft news headlines, like sports or entertainment, as clickbait. In (Chen et al., 2015) authors examine potential methods for automatically detecting clickbait as a form of deception. They survey methods for textual and non-textual features and suggest that hybrid approaches can achieve superior results than following only specific features.

Research coming from the LIS field, in (Chakraborty et al., 2016) professional and cultural issues surrounding online news are reviewed. It is argued that to address such problems, firstly there is a need for proactive public engagement by educators, librarians, and information specialists to promote digital literacy practices and secondly, to develop automated tools to enable the work of journalists in curating, verifying, and fact-checking news, and to assist news readers to identify and filter suspicious information. (Neely-Sardon & Tignor, 2018) discusses and places fake news, in the context of information literacy instruction for college students. In (Paor, 2020) a literature review was conducted to identify the current themes and gaps within the literature on fake news, information literacy, and librarianship. The research findings center around fake news definitions, discussing information literacy frameworks, and outlining efforts made by libraries to refute the spread of misinformation and educate their users. In (Sahoo et al., 2021) authors assessed both qualitatively and quantitatively the published research regarding fake news. Results showed that fake news is a growing research area in the discipline that enabled the production of a good number of studies within the last four years and received numerous citations.

# 2 Data Collection and Annotation

The first into the methodology followed was the collection of data from the web. This was accomplished by using an automated web scraping technique, based on several Python libraries such as Requests, BeautifulSoup, and Newspaper3k[[1]](#footnote-1). The web scraping process began by carefully auditing numerous Greek online news media to find representative samples for the clickbait experiment. This selection ensures that the dataset will contain (in equal parts) both clickbait and non-clickbait articles. As the number is not relevant to the context of this research, two news media were selected to proceed and gather our data. The first covers a wide range of topics such as politics, economy, sports, while the second mostly published articles concerning wellness. It is important that the research is conducted under a controlled environment so that it can be easily understood and generalized, if needed. The landing page of each website was given as input to the web scraper, that visited internal links and extracted their text. Based on the above, a total of 1.565 text documents were collected. Links, multimedia such as image, audio or video, and other non-textual elements are not stored along with the article’s text and therefore not considered.

These text documents were annotated by a single reviewer, placing them into the following two categories, clickbait, and non-clickbait. Based on their title and other elements, 816 documents were classified as non-clickbait, while the rest 749 were classified as clickbait. Normally this task is performed by a group of people to eliminate any conceptual or other bias. By splitting the dataset into two equal halves, by including the same amount of positive and negative samples, it is ensured that the model trained and used to identify clickbait articles, will also remain unbiased. As an extra step to eliminate spam, incomplete, or other documents from the dataset, articles with less than 600 characters were also not considered, decreasing the overall “noise” in our data. Consequently, a new total was formed counting 1.113 articles, from which 438 are classified as clickbait and 675 as non-clickbait. In the following chapter the training and clickbait detection process are described, along with the initial results.

# 3 Training

For this step of the process, the dataset described previously is transformed into a data frame by processing and extracting specific information from each article. This way, training and clickbait detection stages are prepared by consolidating the information contained in the data. Thus, 15 features have been devised for analysis, to help in understanding the nature of clickbait articles. In essence, the model used is trained based on these features

|  |  |  |
| --- | --- | --- |
| **Variable name** | **Meaning** | **Notes** |
| Title | The title of the article | Type: String |
| Title Length | The length of the title | Measured in characters |
| Title Sentiment | Sentiment of the title | Values: [-1, 1] |
| Date | Publication date | Format: 2023-06-19 |
| Time | Publication time | Format: 07:00:34+00:00 |
| Article Text | Text of the article | Type: String |
| Article Text Length | Length of the text | Measured in characters |
| Article Text Sentiment | Sentiment of the article’s text | Values: [-1, 1] |
| Title-Text Sentiment Similarity | Similarity between title and text sentiment | Values: [0,1] |
| Title-Text Similarity | Similarity between title and text | Values: [0,1] |
| Curiosity Gap | Curiosity gap related terms (list) | Values: True/False |
| Numbered List | Listicle type of article | Values: True/False |
| Adjectives | Excessive use of adjectives | Values: True/False |
| Punctuation | Excessive use of punctuation | Values: True/False |
| Clickbait Score | Article classification label | Values: 0, 1 |

and their values, aiming to classify “new” articles, that are not part of the training or testing sets, as clickbait or non-clickbait. These 15 features are discussed in the following table:

Table 1: Dataset variables

# Even though some of the above variables are self-explanatory, some require further explanation as to what is their meaning and their values. Variables related to sentiment (title sentiment, article text sentiment) represent the overall sentiment of the words that compose the title and the article’s text. Their values range from -1 to 1, with -1 meaning negative sentiment and 1 meaning positive sentiment. Similarly, variables related to similarity take values from 0 to 1, where 0 is not similar and 1 is similar. For True/False binary values, some ground rules were set to for the variable to take either value. For example, the Curiosity Gap variable examines if the title of the article contains terms associated with clickbait terminology. The constructed list of terms (both uppercase and lowercase instances) is as follows (translated from Greek): discover, learn, secrets, ways, how, when, which, why, some, someone, some, who, what, find, choose, how much, who, where.

# Likewise, the numbered list examines if the title starts with any number, while adjectives and punctuation examine whether the title makes excessive usage of adjectives (POS tags) and punctuation, such as exclamation marks or question marks. Lastly, Clickbait score that was provided by the human annotator and is considered the ground truth, takes values 0 for non-clickbait articles or 1 for clickbait articles.

# To train the Random Forests model, we separated the dependent variable, which is the Clickbait Score, from the 14 rest, which are the independent variables. The objective here is for the model to train on the independent variables and try to predict the dependent variable. Furthermore, the dataset was partitioned into two smaller subsets, the training set, and the testing set, maintaining a ratio of 80-20. This technique (Hold-out cross-validation) allocates 80% of the total samples for training while keeping 20% of the total samples for testing the model performance. Random Forests classifier creates many trees using a random subset of the training data and a random subset of the dataset features.

# 4 Results

Using the testing subset, the algorithm can evaluate the model’s performance. Results show that the model performed flawlessly in all three metrics, precision, recall and accuracy, achieving a score of 1.00.

Table 2: Testing Results – Precision, Recall, F1-score

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1-score** | **Support** |
| 0 | 1.00 | 1.00 | 1.00 | 132 |
| 1 | 1.00 | 1.00 | 1.00 | 90 |

Table 3: Testing Results – Accuracy, Macro avg., Weighted avg.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1-score** | **Support** |
| **Accuracy** |  |  | 1.00 | 222 |
| **Macro avg.** | 1.00 | 1.00 | 1.00 | 222 |
| **Weighted avg.** | 1.00 | 1.00 | 1.00 | 222 |

The Support column shows the number of actual occurrences of each class in the dataset. For class 0 (non-clickbait), there are 132 instances, and for class 1 (clickbait), there are 90 instances, while Macro avg. and Weighted avg. provide average values for the three metrics.

Additionally, the model evaluates the importance of each feature in the process. These are displayed below, in order of importance. Apparently, only 5 of the 14 features contributed to the model’s predictive performance.

Table 4: Feature Importance

|  |  |
| --- | --- |
| **Feature** | **Importance** |
| Curiosity Gap | 0.639175 |
| Numbered List | 0.128413 |
| Article Text Length | 0.124740 |
| Title length | 0.064068 |
| Title-Text Similarity | 0.043604 |

The achieved results show the effectiveness of the method, even though it is possible that due to the relatively small number of testing samples the model overfitted, learning the data too well. This subset might not fully represent the diversity of real-world data, thus performing poorly on unseen data, and not being able to generalize easily. It is important to note that variables related to sentiment Title Sentiment, Article Text Sentiment, and Title-Text Sentiment Similarity were expected not to be helpful as for every article they had a score of 0.5 (neutral) for the first two, and 1 (similar) for the third. This potentially shows that sentiment measures are not reliable or appropriate for such kind of analysis. Moreover, regarding the Title-Text Similarity variable, out of the total articles, only 50 of them had a similarity score lower than 0.5. Most similar article title-text have a score of 0.81, while less similar have a score of 0.3, with an average of 0.63.

Furthermore, the average length of Article Text Length is 3.707 characters, with 601 characters minimum and 26.569 maximum characters. Similarly, Title Length variable averages 64 characters, with 9 characters minimum and 144 characters minimum. Following, we will discuss the overall progress of the research, limitations, and future endeavours.

# 5 Discussion

Clickbait detection is a feasible task with potential value for relevant institutions. The online news industry, like many other sectors, is greatly influence by the market, a fact that can sometimes have a negative impact on the quality of information-news. Most organizations rely on ad revenue as their primary monetization strategy, with only a few opting for alternative models such as paywalls with additional features. Consequently, marketing techniques like clickbait, which may be considered a legit marketing technique or fall into a 'grey' area, find widespread adoption across various media outlets.

Clickbait is characterized by titles designed to attract attention, potentially overshadowing the quality of the article content. Determining whether clickbait is related to forms of spam articles or low-quality journalism can be challenging, and this contributes to the difficulty readers face in identifying such instances, in the middle of the information overload of mass platforms.

Linguistic features, among others, play a role in empowering intelligent systems to detect clickbait. Our approach highlights the significance of considering title and text length as well as their similarity. While sentiment analysis is theoretically applicable, it proves to be unreliable in our experiment. However, the danger is real; the potential misuse of sentiment to target a specific audience, such as inducing caution and fear, can prove to be a manipulative force in attracting audience attention.

Future plans involve refining the curiosity gap glossary and exploring alternative algorithms to compare their performance. From a library-information science standpoint, we believe there is room for expansion in utilizing news data and implementing media literacy programs to enhance awareness in this domain.

**Acknowledgements**

Funding for this research was provided by the Special Account for Research Grants of the University of West Attica.

# Bibliography

Batchelor, O. (2017). Getting out the truth: The role of libraries in the fight against fake news. Reference Services Review, 45(2), 143–148. https://doi.org/10.1108/RSR-03-2017-0006.

Bazaco, Á., Redondo, M., Sánchez-García, P., & de Valladolid. (2019). Clickbait as a strategy of viral journalism: Conceptualization and methods.

Beleslin, I., Njegovan, B. R., & Vukadinović, M. S. (2017). Clickbait titles: Risky formula for attracting readers and advertisers.

Blom, J. N., & Hansen, K. R. (2015). Click bait: Forward reference as lure in online news headlines. Journal of Pragmatics, 76, 87-100.

Bourgonje, P., Schneider, J. M., & Rehm, G. (2017). From Clickbait to Fake News Detection: An Approach based on Detecting the Stance of Headlines to Articles.

Chakraborty, A., Paranjape, B., Kakarla, S., & Ganguly, N. (2016). Stop Clickbait: Detecting and Preventing Clickbaits in Online News Media.

Chen, Y., Conroy, N. J., & Rubin, V. L. (2015). Misleading Online Content: Recognizing Clickbait as “False News.”.

Chen, Y., Conroy, N., & Rubin, V. (2015). News in an online world: The need for an "Automatic Crap Detector." The Proceedings of the Association for Information Science and Technology Annual Meeting (ASIST2015), Nov. 6–10.

Chen, Y., & Rubin, V. L. (2017). Perceptions of Clickbait: A Q-Methodology Approach. Q.

Cofaru, V., & Groza, A. (2022). Detecting clickbaits using Deep Forest.

Dogruel, L., Masur, P., & Joeckel, S. (2022). Development and Validation of an Algorithm Literacy Scale for Internet Users. Communication Methods and Measures, 16(2), 115-133. https://doi.org/10.1080/19312458.2021.1968361.

Dor, Daniel. (2003). On Newspaper Headlines as Relevance Optimizers. Journal of Pragmatics - J PRAGMATICS, 35.

Eva, N., & Shea, E. (2018). Marketing libraries in an era of "fake news". Reference & User Services Quarterly, 57(3), 168–171.

Faix, A., & Fyn, A. (2020). Framing fake news: Misinformation and the ACRL framework. Portal: Libraries and the Academy, 20(3), 495–508.

Gallagher, B. (2016). What impact does fake news have on the real world? In Mary Jane. December 5th.

Geçkil, A., Müngen, A. A., Gündogan, E., & Kaya, M. (2018). A clickbait detection method on news sites. In 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). Barcelona, Spain. https://doi.org/10.1109/ASONAM.2018.8508452.

Gerard Genette, G. (1997). Paratexts: Thresholds of Interpretation. Cambridge University Press, Mar 13, 1997 - Literary Criticism.

Head, A. J., Fister, B., & MacMillan, M. (2020). Information literacy in the age of algorithms.

Kapellas, N. and Kapidakis, S. (2022). A Text Similarity Study: Understanding How Differently Greek News Media Describe News Events. In Proceedings of the 14th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2022) - KEOD; ISBN 978-989-758-614-9; ISSN 2184-3228, SciTePress, pages 245-252. DOI: 10.5220/0011589700003335.

Kapellas, N., & Kapidakis, S. (2023). Event Detection in News Articles: A Hybrid Approach Combining Topic Modeling, Clustering, and Named Entity Recognition. In Proceedings of the 15th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - Volume 2: KEOD (pp. 272-279). ISBN 978-989-758-671-2. ISSN 2184-3228. DOI: 10.5220/0012234300003598.

Knobloch, Silvia & Patzig, Grit & Mende, Anna-Maria & Hastall, Matthias. (2004). Affective News: Effects of Discourse Structure in Narratives on Suspense, Curiosity, and Enjoyment While Reading News and Novels. Communication Research, 31, 259-287. https://doi.org/10.1177/0093650203261517.

Lasya Venneti and Aniket Alam. Clickbaits: Curious hypertexts for news narratives in the digital medium. In Proceedings of Workshop on Narrative and Hypertext ’17 (NHT17), July 2017, Prague, Czech Republic.

Lex, E., Juffinger, A., & Granitzer, M. (2010). A Comparison of Stylometric and Lexical Features for Web Genre Classification and Emotion Classification in Blogs. In TIR 2010 - 7th International Workshop on Text-Based Information Retrieval, in Conjunction with DEXA 2010 (pp. 10-14). https://doi.org/10.1109/DEXA.2010.24.

Lim, S. (2020). Academic library guides for tackling fake news: A content analysis. Journal of Academic Librarianship, 46(5), 102195. https://doi.org/10.1016/j.acalib.2020.102195.

Loewenstein, G. (1994). The psychology of curiosity: A review and reinterpretation. Psychological Bulletin, 116(1), 75–98. https://doi.org/10.1037/0033-2909.116.1.75

Mackey, T., & Jacobson, T. (2010). Reframing Information Literacy as a Metaliteracy. College & Research Libraries, 72, 62-78. https://doi.org/10.5860/crl-76r1

Marchi, R. (2012). With Facebook, Blogs, and Fake News, Teens Reject Journalistic “Objectivity.” Journal of Communication Inquiry, 36(3), 246–262. https://doi.org/10.1177/0196859912458700.

Molina, M. D. (2021). Does Clickbait Actually Attract More Clicks? Three Clickbait Studies You Must Read.

Molyneux, L., & Coddington, M. (2020). Aggregation, Clickbait and Their Effect on Perceptions of Journalistic Credibility and Quality.

Munger, Kevin & Luca, Mario & Nagler, Jonathan & Tucker, Joshua. (2020). The (Null) Effects of Clickbait Headlines on Polarization, Trust, and Learning. Public Opinion Quarterly, 84, 49-73. https://doi.org/10.1093/poq/nfaa008.

Neely-Sardon, A., & Tignor, M. (2018). Focus on the facts: A news and information literacy instructional program. The Reference Librarian.

Nyhan, B., & Reifler, J. (2015). The effect of fact-checking on elites: A field experiment on U.S. state legislators. American Journal of Political Science, 59, 628-640. https://doi.org/10.1111/ajps.12162

Oliva, C., Marín, I. P., Lago-Fernández, L. F., & Arroyo, D. (2022). Rumor and clickbait detection by combining information divergence measures and deep learning techniques.

Paor, S. D. (2020). Information literacy and fake news: How the field of librarianship can help combat the epidemic of fake news. The Journal of Academic Librarianship.

Potthast, M., Kopsel, S., Stein, B., & Hagen, M. (2016). Clickbait detection. In European Conference on Information Retrieval (pp. 810–817). Springer.

Potthast, M., Gollub, T., Komlossy, K., Schuster, S., Wiegmann, M., Fernandez, E. P. G., Hagen, M., & Stein, B. (2018). Crowdsourcing a Large Corpus of Clickbait on Twitter.

Rony, M. M. U., Hassan, N., & Yousuf. (2017). Diving deep into clickbaits: Who use them to what extents in which topics with what effects? arXiv:1703.09400v1. Retrieved from https://arxiv.org/abs/1703.09400.

Rubin, V., Brogly, C., Conroy, N., Chen, Y., & Cornwell, S. E. (2019). A News Verification Browser for the Detection of Clickbait, Satire, and Falsified News.

Sahoo, J., Sahu, S. C., & Mohanty, B. (2021). Research on Fake News: An Empirical Analysis of Selected Library and Information Science Journals. 41(4).

Scacco, J. M., & Muddiman, A. (2020). The curiosity effect: Information seeking in the contemporary news environment. New Media & Society, 22(3), 429–448. https://doi.org/10.1177/1461444819863408.

Vijgen, B. (2014). The Listicle: An Exploring Research on an Interesting Shareable New Media Phenomenon. Studia Universitatis Babes-Bolyai-Ephemerides, (1), 103–122.

Wilkes, A. L., & Leatherbarrow, M. (1988). Editing Episodic Memory following the Identification of Error. The Quarterly Journal of Experimental Psychology Section A, 40(2), 361–387. https://doi.org/10.1080/02724988843000168.

Zhou, Y. (2017). Clickbait Detection in Tweets Using Self-Attentive Network (arXiv:1710.05364). arXiv. http://arxiv.org/abs/1710.05364.

1. https://github.com/kapelnick/clickbait\_detection [↑](#footnote-ref-1)