Clickbait Classifier - NLP Final Project

Anna Fenske
af2570@nyu.edu

Abstract

This document contains a detailed report of a Naive Bayes classifier which takes article headlines as input and classifies each headline as either "clickbait" or "news". This program is my final project for NYU's undergraduate Natural Language Processing course in the College of Arts and Sciences in the fall semester of 2016.

1 Introduction

This program uses the Naive Bayes classifier contained in the nltk.classify package on a training corpus containing data collected from Buzzfeed, Upworthy, the New York Times, the Wall Street Journal and CNN, and the sets of features containing information about the different types of words and parts of speech (and the quantities of each type) gathered from each headline to classify article titles as either "clickbait" or "news" (non-clickbait).

2 Problem Statement

Clickbait, according to the Oxford English Dictionary, is a term which describes "content, especially that of a sensational or provocative nature, whose main purpose is to attract attention and draw visitors to a particular web page". The presence of this species of journalism can be seen as early as the late 1800s, in a number of political cartoons and opinion pieces condemning the use of "yellow journalism". Yellow Journalism (or "Yellow Press"), coined in the 1890s to describe the tactics used by two competing New York City newspapers (the New York World and the Journal), describes "the use of lurid features and sensationalized news in newspaper publishing to attract readers and increase circulation"[1].

With this definition in mind, I sought to use some of the linguistic properties discussed throughout this course to quantify the "sensationality" or "provocativeness" of a given headline.

Before explaining how I approached this problem, let's explore in more depth the exact definition of clickbait and what forms it might take, and the problems posed by this type of journalism:

2.1 7 Defining Elements of Clickbait (#4 Will Shock You!)

As the use of clickbait by online journals and other digital media organizatoins has become more widespread, we can see a number of distinct elements and strategies which have become common among most clickbait headlines. These elements, listed below, each represent a different common end goal - to pique the interest of a prospective reader.

1. **The Listicle**: Perhaps the most well-known tactics of clickbait, the "listicle" is a headline which describes a numbered or bulleted list of items, ranked or unranked.

   Distinguishing Lists and Listicles: Though headlines describing lists are common enough among non-sensationalized media, a major distinguishing factors between headlines which describe lists and listicle headlines is the promise to the reader that follows. The title of this subsection covers a basic example of this kind of promise: the use of a parenthetical "(Number X will ------)".

   Example: *23 Incredible Photos From 2016 That Prove It Wasn't A Total Dumpster Fire*[2]

[1] Encyclopedia Britannica

[2] http://www.upworthy.com/23-incredible-photos-from-2016-that-prove-it-wasnt-a-total-dumpster-fire?c=tpstream

2. **A Promise to the Reader (cont'd)**: The promise element described above is not unique to listicle-type clickbait headlines. A promise to the reader will often be used to generate intrigue. The promise element is seldom dependent on the content of the article itself, since the promise is itself often a banal (or fairly subjective) claim.

   Common Promises: Listed below are some phrases commonly used in a promise to the reader.

   - *"...Will Make You ____"*
   - *"...Restore [someone's] Faith in Humanity"*
   - *"You Won't Believe _____ ..."*
   - *"If You ____ Then You _____"*

   Examples:

   - *5 Images of Victorian England That Will Make You Rethink LGBTQ History* [3]
   - *If You Love Being Single, You'll Love Emma Morano, The Oldest Person On Earth* [4]
   - *Unimaginable Trauma And Cruelty Didn't Break Them. Their Letters Will Tell You Why* [5]

3. **Reference to "You" or "I"**: Another common device employed in clickbait titles is a direct reference to the reader and/or the author. Titles which make use of this strategy often describe a personal story and are often written in a conversational tone, which exploits the reader's empathy to attract attention more effectively than a headline written in the third-person perspective.

   Stories About Unnamed Subjects: There are many cases in which a clickbait headline written in the third person perspective describes a personal story. In the vast majority of such cases, the subject of the headline will be described exclusively using pronouns or a combination of pronouns and vague nouns ("man", "woman", "guy", nouns describing the subject's occupation, etc.).

   Examples

   - *20 Years Ago, He Helped Two Kids At Disney World. Today, His Story Helped Even More.* [6]
   - *5 Surprising Pieces of Life Advice My Dad Left Me Before He Died* [7]

4. **A Claim. Then a Plot Twist.**: This strategy is more complex then the previous three, involving more than just word choice or sentence structure. Headlines which take this form often begin with the introduction of a fairly typical story, idea, or theme in the first sentence, immediately followed by a much shorter sentence to complicate or undermine the claim in the first. Titles which fall into this category rely entirely on the sense of unresolvedness caused by the "twist" in the second sentence. By introducing a contradiction in the headline and denying any immediate closure, these titles seek to prompt the reader to click on the link in hopes that the article itself will resolve the irritation which the incomplete title has caused.

   Examples:

   - *It's The Hubble Telescope's Most Famous Image. Here's How It Almost Didn't Happen.* [8]
   - *If You Could Press A Button And Murder Every Mosquito, Would You? Because That's Kinda Possible.* [9]
   - *This Story Starts Out So American, Normal, And Everyday. Until The End.* [10]

5. **Reference to Visual Media**: Many clickbait headlines describe videos, comic strips, or some other form of visual media. Non-text-based articles as a rule are more likely to grab

[3] http://www.upworthy.com/5-images-of-victorian-england-that-will-make-you-rethink-lgbtq-history?c=tpstream

[4] http://www.upworthy.com/if-you-love-being-single-youll-love-emma-morano-the-oldest-person-on-earth?c=tpstream

[5] http://www.upworthy.com/unimaginable-trauma-and-cruelty-didnt-break-them-their-letters-will-tell-you-why?c=tpstream

[6] http://www.upworthy.com/20-years-ago-he-helped-two-kids-at-disney-world-today-his-story-helped-even-more?c=tpstream

[7] http://www.upworthy.com/5-surprising-pieces-of-life-advice-my-dad-left-me-before-he-died?c=tpstream

[8] http://www.upworthy.com/its-the-hubble-telescopes-most-famous-image-heres-how-it-almost-didnt-happen?c=tpstream

[9] http://www.upworthy.com/if-you-could-press-a-button-and-murder-every-mosquito-would-you-because-thats-kinda-possible

[10] https://www.upworthy.com/this-story-starts-out-so-american-normal-and-everyday-until-the-end

attention since they do not require as much effort as text-based articles do. Comics appear to be particularly popular subjects for clickbait articles which reference visual media. However, since a description of the content alone does not a clickbait title make, referencing visual media is often used in tandem with one or more other strategies in this list.

Example: *A Short Comic Gives The Simplest, Most Perfect Explanation of Privilege I've Ever Seen.*[11]

6. **From the Perspective of** ____: Branching off of the "personal story" style discussed earlier, some clickbait titles which use this strategy advertise articles which provide a specific perspective on an often common experience.

   Example: *6 Muslim American Women Share Their Thoughts On The Election*[12]

7. **Descriptive Words and Sweeping Generalizations**: Perhaps the most easily identifiable tic associated with clickbait titles is the consistent use of adverbs, adjectives (especially superlatives) to "spice up" titles which would sound more mundane otherwise. The use of sweeping generalizations about groups of people (specifically the opinions of groups of people), indicated by the use of words like "everyone" and "no one", operate as hyperbole in the same way that superlatives do. Superlatives and adverbial phrases rarely appear in news (non-clickbait) headlines since the purpose of such titles is to give a succinct and informative outline of the content of the article; these phrases imply the presence of opinion which will underlie any factual information the article communicates. And, in the case of clickbait articles, this opinion often is the main focus of the piece.

   Example: *These Weird Nanorobots Could Make Chemotherapy Treatment Easier* [13]

Each of the 7 strategies outlined in the above list confirm that the main distinction between a clickbait and a non-clickbait headline is its core purpose. While non-clickbait titles aim to summarize the content of the pieces they describe, the task of clickbait is simply to hook potential readers. The structure of a clickbait headline is seldom dependent on the content of the article.

It is also important to note that these 7 tactics are often used in combination; clickbait headlines will usually employ two or more of these strategies at the same time.

For Example: *12 Funny Comics That Might Help You Feel A Bit Less Anxious Today* [14] employs tactics 1, 2, 3, and 5.

2.2 The Uses and Dangers of Yellow Journalism Today

In recent years, the use of clickbait-style headlines has become a more and more popular tactic for increasing the reader count for many online news sources. Some digital media organizations (*Buzzfeed*[15] and *Upworthy*[16], for example) have become notorious for their use of exaggerated and gratuitously informal clickbait headlines - giving rise to parody websites like *Clickhole*[17] which mimic the style of these headlines.

Although this style of journalism - playing entirely on the emotions and curiosity of the reader - has existed in the literary world since before the birth of the internet, this trend of sensationalized headlines and personal confessions seems to have coincided with the rise of social media. The sheer amount of information accessible by one person has grown exponentially in the past decade as social networking sites like Facebook, Tumblr and Twitter (among others) have gained widespread popularity as vehicles for sharing thoughts, photos, and links. With the pervasiveness and rapid growth of this online information-sharing network, digital media organizations have aimed to make their content the *most* eye-catching in order to stand out among the thousands of other links available to a prospective reader on his or her Facebook news feed. The goal of these articles is no longer to communicate meaningful information, but to appeal to the reader in order to go "viral".

Says PBS' Jeffrey Dvorkin on clickbait: "It's

[11] http://www.upworthy.com/a-short-comic-gives-the-simplest-most-perfect-explanation-of-privilege-ive-ever-seen

[12] http://www.upworthy.com/6-muslim-american-women-share-their-thoughts-on-the-election?c=tpstream

[13] http://www.upworthy.com/these-weird-nanorobots-could-make-chemotherapy-treatment-easier?c=tpstream

[14] http://www.upworthy.com/12-funny-comics-that-might-help-you-feel-a-bit-less-anxious-today?c=tpstream

[15] http://www.buzzfeed.com

[16] http://www.upworthy.com

[17] http://www.clickhole.com

rarely newsworthy, but it does attract eyeballs. The assumption seems to be that audiences might stay for the 'serious' content after gorging on the fluff."

The overwhelming popularity of digital media has caused its fair share of problems. Though we have not yet seen any drastic concrete issues as a result of clickbait news, there has been a major shift in the emphasis of the content of many journals from the information presented to the method of presentation. Today, many major news journals (CNN, the New York Times, etc.) have begun to experience this same push to produce viral content. Digital media organizations like Buzzfeed and Upworthy have begun to reap larger and larger profits as they continue to pump out exaggerated and opinionated content presented by titles which manipulate the reader into clicking in, links which do not provide as intriguing a rundown of the substance of their articles have lost readers. Consequently, many major news sources have begun to publish articles with sensationalized packaging, simply to keep up in the now-dominant world of digital media.

We begin to see some of the more concrete risks of embellished and exaggerated news when we consider its role in the context of the most recent general election (2016). The spread of "fake news", specifically that which surrounded the election, has been facilitated by articles of this sort: playing on the emotions - predominantly feelings of fear and distrust - of voters to both draw in readers and promote the sharing of such articles. The circulation of these "fake news" headlines, paired with the hesitation by reputable news sources to "play dirty" (their reluctance to participate in the publication of dramatized news) kindled much of the conflict and hatred which characterized this election.

3 Approach

Since the classification algorithm used was provided by the Naive Baye's Classifier in the nltk.classify.NaiveBayesClassifier package, the majority of my work was in gathering training data and extracting features from each headline to generate a feature set which would

3.1 Training Data

The training corpus used in this project comprised around 16,000 manually annotated headlines from Buzzfeed (3711 titles), Upworthy (5055 titles), CNN (730 titles), the Wall Street Journal (6442 titles) and the New York Times (1186 titles). Titles from Buzzfeed and Upworthy were labelled as "clickbait", and titles from CNN, the New York Times and the Wall Street Journal were labelled as "news"[18].

3.2 Forms

As explained in the previous section, there are many multi-token phrases which appear commonly among clickbait titles and can indicate the usage of one or more of the elements of clickbait headlines outlined in section 1.1. While writing this program, I was tasked with identifying occurrences of these phrases in each headline. There were two major considerations which needed to be addressed in my approach to this task:

1. Variation in both the exact word choice and the length of the phrase itself

2. Variation in the placement of the phrase in the title

My original plan for this task involved building a Context Free Grammar and using some built-in functions provided by nltk to recognize a number of distinct phrase structures. However, building a CFG requires a number of terminals, which I could not provide because (1) I could not accurately predict the exact word choice of any given clickbait headline and (2) using terminals, however many, could ultimately allow for a headline following the pattern one of these phrase structures to avoid being identified as having one of these structures because some unique words in the title were not included in the CFG.

Instead, I implemented a simple function, `has_form`, which takes as input a tokenized title and a **form**, a phrase structure represented as a list of non-empty and empty strings and some nested lists of non-empty strings.

Representing Phrases as Lists: The list representation I implemented allows for a list to be made up of elements falling into one of the three categories:

1. `" "` = *Blank Space*: Empty strings represent elements in the phrase structure that can be filled by any word. This takes the place of a

[18]"News" annotation refers to non-clickbait titles

terminal for phrase elements which the programmer (I) could not predict, but were not essential in identifying the overall pattern.

2. `"nonempty"` = *Distinct Token*: Nonempty strings represent terminal phrase elements.

3. `["one", "two"]` = *Space with Multiple Options*: Nested lists of nonempty strings represent phrase elements for which there are multiple terminal options.

Order of Operations in `has_form` (and `helper`):
`has_form(form, sentence)`:

1. If the number of elements in the form is greater than the number of tokens in the sentence, return `False`.

2. Otherwise:

   - If the first token of the form is an empty string, call `helper( form[1:], sentence[i+1:] )` for all indices i such that 0 <i ≤ `(len(sentence) - len(form))`.
   - If the first token of the form is a nonempty string, call `helper( form[1:], sentence[i+1:] )` for all indices i such that 0 <i ≤ `(len(sentence) - len(form))` and the element at index i in the sentence is equal to the first token of the form.
   - If the first token of the form is a list of nonempty strings, call `helper( form[1:], sentence[i+1:] )` for all indices i such that 0 <i ≤ `(len(sentence) - len(form))` and the element at index i in the sentence is in the list at the first index of the form.

3. Return `True` if and when `helper` returns `True` while iterating through the indices in the tokenized sentence (as outlined above). If we reach the end of the sentence without returning any value, return `False` as we have not found any occurrence of the given form in the given sentence.

`helper(form, sentence)`:

1. If the length of the form equals 0, return `True`, as we have reached the end of the form without encountering a mismatched element in the sentence.

2. If the length of the form is greater than the length of the sentence, return `False`.

3. Otherwise:

   - If the first element of the form is a list of nonempty strings and the first element of the given sentence is contained in that list, return the value given by a recursive call `helper( form[1:], sentence[1:] )`.
   - If the first element of the form is an empty string, return the value given by a recursive call `helper( form[1:], sentence[1:] )`, since we do not need to check for a match in the sentence for a blank space.
   - If the first element of the form is a nonempty string and the first element of the sentence is equal to that string, return the value given by a recursive call `helper( form[1:], sentence[1:] )`.
   - Otherwise, return `False`.

The operation performed by `has_form` does not account for variation in the number of tokens in the applied phrase structure (That is, if a blank token were to be occupied by more than one token in the sentence, this function would not recognize the structure). However, this can be easily added, and will be added in the future.

3.3 Features

multiple_sentences: True if the headline contains multiple sentences, false otherwise.

has_past_tense: True if the headline contains one or more occurrences of past tense verbs, false otherwise.

has_ing: True if the headline contains one or more occurrences of verbs ending in -ing, false otherwise.

start_tag: String containing the Penn Treebank POS tag (gathered from nltk.pos_tag()) of the first token in the headline.

start_tag_bigram: String containing the Penn Treebank POS tags of the first two tokens in the headline.

Pronoun Type	Word List
Personal	*I, me, you, she, her, he, him, we, us, they, them*
Relative	*that, which, who, whom, whose, whichever, whoever, whomever*
Demonstrative	*this, these that those*
Indefinite	*anybody, anyone, anything, each, either, everybody, everyone, everything, neither, nobody, nothing, somebody, someone, something, both, all, any, most, some*
Interrogative	*what, who, which, whom, whose*
Possessive	*my, your, his, her, mine, yours, his, hers, our, their, ours, theirs*
Subject	*I, you, she, he, we, they*
Object	*me, you, her, it, him, us, you, them*

Table 1: Table 1: Pronoun Types

conjunction: True if the headline contains one or more occurrences of "and" or "but", false otherwise. [19]

multiple_you: True if the headline contains multiple occurrences of "you", "your" or "yours", false otherwise.

has_pronoun: True if the headline contains one or more occurrences of any pronoun, false otherwise. [20]

multiple_pronouns: True if the headline contains multiple occurrences of any pronoun, false otherwise.

has_adverb: True if the headline contains one or more adverb, false otherwise.

has_adjective: True if the headline contains one or more adjective, false otherwise.

may_have_list: True if the headline contains one or more occurrences of a cardinal number ("CD" in POS tag sequence given by nltk.pos_tag()) and one or more occurrences of a plural noun or plural proper noun ("NNS" or "NNPS"), false otherwise.

indicative_remark: True if the headline meets one or more of the following requirements, false otherwise.

- The headline contains at least one question mark and at least one occurrence of the word "you", "you", or "yours".
- The headline contains at least one occurrence of an interrogative pronoun (Table 1 for list of interrogative pronouns), but NO occurrences of a question mark.
- The headline contains one or more occurrences of an exclamation mark(s) ("!", "?!", "!!", "!!!").

indicative_words: True if the headline contains one or more occurrences of words which fall into one of the following categories:

- *Curse*: Curse words
- *Vague*: Words which provide little descriptive information or extremely general references to an individual or a group of people
- *Medium*: Words which describe a type of visual media or artistic medium
- *Social Media*: Words which reference the internet, internet trends, or social media
- *Slang*: Acronyms (text abbreviations such as LOL, WTF, etc.) or other slang terms
- *Emotion*: Words which indicate some sort of emotion
- *Command*: Words which command the reader to do something (e.g. *Listen, Watch, See*, etc.)
- *Curiosity*: Words which do not fall into any of the above categories but are still commonly used by clickbait titles to attract the attention of the reader

multiple_indicative_words: True if the headline contains multiple occurrences of words in the above categories.

indicative_start: True if the headline contains a sentence whose first token is a pronoun in Table 1

[19] I chose to reduce the list of conjunctions to these two since other conjunctions were not common enough among both clickbait and news titles to provide any meaningful data for classification.

[20] I did not include reflexive pronouns since they did not occur commonly enough to provide any meaningful data: in my training corpus, fewer than 1% of clickbait and news titles contained any reflexive pronouns.

or whose first token is a common clickbait starting tokens (*Meet, Why, How, When, If, [number], For, Now, Here*), false otherwise.

the_start: True if the headline contains a sentence beginning with a determiner (DT) followed by either a cardinal number (CD) or an adjective (JJ, JJR, or JJS), false otherwise.

indicative_word_start: True if the headline contains a sentence beginning with a word falling into one of the "indicative_word" categories described above, false otherwise.

has_form: True if the headline contains one of the following forms, false otherwise. Listed below are the forms considered in this program:

- ["that", "will", " "]
- [["make", "give"], "you", " "]
- ["faith", "in", "humanity"]
- ["here", "'s", ["how", "what", "why", "the"]]
- ["here", ["is", "'s", "are"]]
- [POSS_PRONOUNS[21], "story"]
- ["people", "are"]
- ["let", "'s"]
- ["DT", ["DC", "JJ"]][22]
- ["CD", ["NNS", "NNPS"]][22]
- ["CD", " ", ["NNS", "NNPS"]][22]

indicative_form: True if the headline contains one of the forms in Table 2 or if the headline contains an indicative remark (described above), false otherwise.

indicative_form_start: True if the headline contains an indicative form or if the headline contains an indicative starting token, false otherwise.

4 Results

4.1 Test Data

The test corpus used comprised 5,286 manually annotated headlines: 2,182 clickbait headlines scraped from Buzzfeed, and 3,104 non-clickbait headlines scraped from the New York Times. Since both Buzzfeed and the New York Times were sources of data in the training corpus, any duplicates found in the test corpus were removed prior to testing, so test results would not be skewed.

[21] refers to the list of possessive pronouns given in Table 1

[22] These forms were included in a search performed by `has_form` on the POS tag sequence of the given headline.

Name	Value	Likelihood Ratio
start_tag_bigram	PRP VBD	C : N = 54.7 : 1.0
start_tag_bigram	WDT NNP	C : N = 54.4 : 1.0
start_tag_bigram	NN TO	N : C = 33.9 : 1.0
multi_ind_words	True	C : N = 31.6 : 1.0
start_tag	WDT	C : N = 29.0 : 1.0
start_tag_bigram	MD PRP	C : N = 25.5 : 1.0
multi_you	True	C : N = 25.2 : 1.0
multi_pronouns	True	C : N = 20.7 : 1.0
start_tag_bigram	WP NN	C : N = 19.4 : 1.0
start_tag_bigram	NNS :	N : C = 17.8 : 1.0

Table 2: Most Informative Features

	Clickbait	News
Precision	0.914187	0.979103
Recall	0.971586	0.935889
F-Measure	0.942013	0.957009
Accuracy	0.9506242906	

Table 3: Evaluation Metrics[23]

4.2 Classification Results

After running the classifier on the test set of feature vectors generated from the test corpus described above, the system identified 2,319 clickbait headlines (2,120 of which were correctly classified), and 2,967 news headlines (2,905 of which were correctly classified). Table 2 lists the most informative features as given by the `show_most_informative_features()` function provided by `nltk.classify`.

4.3 Evaluation Metrics

Table 3 lists the evaluation metrics calculated from the results of the program on this test corpus, and the overall accuracy of the program.

References