Naive Bayesian approach to news headline classification

KENNETH ALLEN, Armstrong State University, United States of America
JEFFREY YOUNG, Armstrong State University, United States of America

News article headlines have traditionally been a source of attraction without providing much content. Traditionally, the information value of headlines has been considered miniscule at best. This research effort tests these preconceived notions of headline information quality as well as the usability of minable headline data. The efforts of this study began with attaining a set of approximately one million headlines taken from the Australian Broadcasting Corporation. The data set was processed using a naive Bayes approach to produce both test and training sets. The training set was used to make predictions of weekend, month, season, year, election time, first chronological half of the data set, prime minister's affiliation, and originating publication. The study has concluded that quality information can be extracted and features can be predicted with an accuracy well above that of chance.

Additional Key Words and Phrases: Machine learning, Support Vector Machines, LibSVM, Stock prediction, S&P 500

ACM Reference Format:
Kenneth Allen and Jeffrey Young. 2017. Naive Bayesian approach to news headline classification. 1, 1 (November 2017), **??** pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 BACKGROUND

1.1 Bayes' Theorem

Bayes' Theorem is a simple mathematical formula used for calculating conditional probabilities. It figures prominently in subjectivist or Bayesian approaches to epistemology, statistics, and inductive logic. Subjectivists maintain that rational belief is governed by the laws of probability and lean heavily on conditional probabilities in their theories of evidence and their models of empirical learning. Bayes' Theorem is central to these initiatives both because it simplifies the calculation of conditional probabilities and because it clarifies significant features of subjectivist position. Indeed, the Theorem's central insight that a hypothesis is confirmed by any body of data that its truth renders probable is the cornerstone of all subjectivist methodology.

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

1.2 Naive Bayes Classification

Naive Bayes is a family of algorithms that take advantage of probability theory and Bayes' Theorem to predict the category of a sample (like a piece of news or a customer review). They are probabilistic, which means that they calculate the probability of membership in each category for a given sample, and then output the category with the highest one. The way they get these probabilities is by using Bayes' Theorem, which describes the probability of a feature, based on prior knowledge of conditions that might be related to that feature.

Authors' addresses: Kenneth Allen, Armstrong State University, Department of Computer Science & Information Technology, Savannah, GA, 31419, United States of America, ka3878@stu.armstrong.edu; Jeffrey Young, Armstrong State University, Department of Computer Science & Information Technology, Savannah, GA, 31419, United States of America, jy8672@stu.armstrong.edu.

XXXX-XXXX/2017/11-ART
https://doi.org/10.1145/nnnnnnn.nnnnnnn

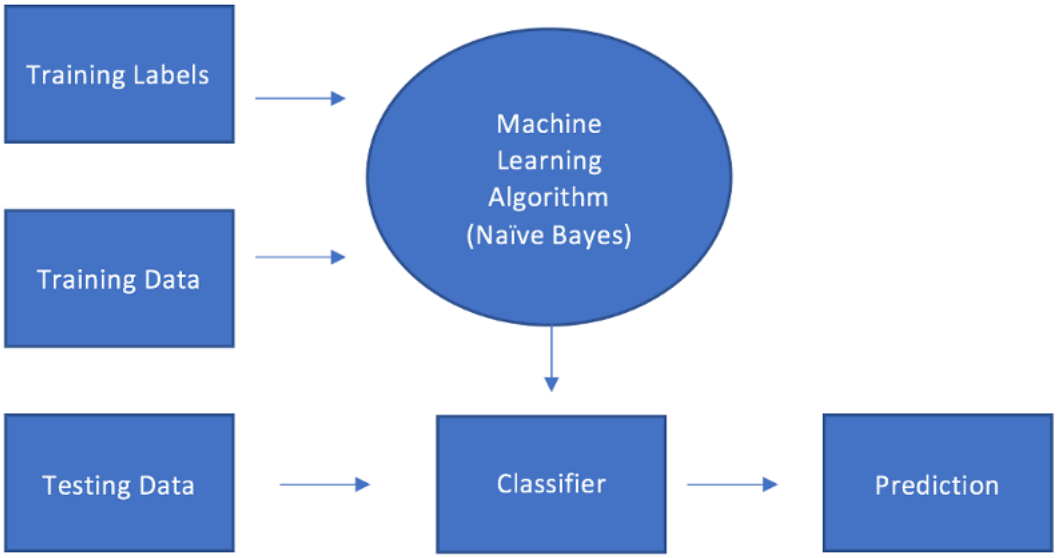

Fig. 1. Block diagram of classifier system.

Take a set of classes C, a domain S, an element $s \in S$, and a set of s's features $a_1, ..., a_n$. We use simple Laplace smoothing to compensate for some low-probability features occuring zero times in a limited training set. The classifier is a function $f : S \mapsto C$ defined as such:

$$f(s) = \underset{c \in C}{\operatorname{argmax}} P(c|s)$$

$$P(c|s) = \frac{P(c)}{P(s)} P(s|c)$$

$$P(s|c) = \prod_{i=1}^{n} P(a_i|c)$$

$$P(a_i|c) = \frac{\text{training samples in class with feature} + 1}{\text{total training samples in class} + n}$$

2 DATA PROCESSING

2.1 Training Labels

Labeled corpora may be used to train and evaluate a wide range of learning algorithms. Assigning a label is considered a judgment task performed by a human (worker, judge, expert, annotator, etc.). Labels for this undertaking were determined exclusively by date and originating publication.

2.2 Data

Data used in test and training sets was sourced from two sources. The first dataset was sourced from the Australian Broadcasting Corporation (http://www.abc.net.au/news/) and consisted of approximately one million

news headlines with corresponding dates. The second dataset was sourced from Examiner.com (previously at http://www.examiner.com/) and consisted of approximately three million headlines with corresponding dates.

The first dataset was trained and used to extrapolate weekend, month, season, year, election time, first chronological half of the data set, and current prime minister's party affiliation. The second dataset was used to train the classifier to distinguish between headlines from two different publications: a longstanding, public, professional broadcaster and a crowdsourced publishing platform with loose standards. Essentially, it was being challenged to distingish between news for a regional or global audience, as well as recognizing so-called 'clickbait' designed exclusively to be psychologically enticing.

For each test, 10% of the data was randomly removed and used as a test set while the remaining 90% was used as a training set.

2.3 Naive Bayes Classifier

The formulas described in the Background section were implemented in Java. Since the product of so many miniscule fractions can easily underflow an IEEE 64-bit double-precision floating-point number's exponent, the calculations are done using Java's built-in `BigDecimal` data type, with a precision setting of 32 bits for each of the exponent and significand.

2.4 Prediction

Prediction quality was measured with a bevy of statistics. Accuracy is simple to calculate ($\frac{\text{correct classifications}}{\text{tested elements}}$), and can be easily compared to the accuracy of a feature-ignorant guess.

For binary criteria, more informative values can be generated. Use TP, FP, FN, and TN, to represent true positives, false positives, false negatives, and true negatives, respectively. Especially important are true positive rate ($TPR = \frac{TP}{TP+FN}$), true negative rate ($TNR = \frac{TN}{FP+TN}$), positive predictive value ($PPV = \frac{TP}{TP+FP}$), and negative predictive value ($NPV = \frac{TN}{FN+TN}$).

The Naive Bayesian Classifier does not have a sensitivity parameter, so it only creates one point on a Receiver Operating Characteristic (ROC) graph. We can trivially interpolate between that point and the points $(0, 0)$ and $(1, 1)$ by imagining a series of classifiers that randomly choose with a certain probability whether to refer to our classifier or to automatically reject or accept, respectively (as in Fig. 2). (This is equivalent to 'convex hull' techniques, but for a single data point.) In this way, we get a curve and can calculate the area under it simply as $AUC = \frac{TPR+TNR}{2}$. The F_1 score is calculated as the harmonic mean of TPR and PPV:

$$F_1 = \frac{2}{\frac{1}{TPR} + \frac{1}{PPV}}$$

3 RESULTS AND PERFORMANCE

Table 1 gives a summary of the performance of the classifier given different labeling criteria, on a partial and full headline dataset. The actual Accuracy can be compared favorably to a Random Guess, which is made only with the knowledge of the relative class frequencies (for instance, guessing an article was published on a weekend with a probability of $\frac{2}{7}$ or on a weekday with a probability of $\frac{5}{7}$). The ROC-AUC (Area Under the Curve of the Interpolated Receiver Operating Characteristic) and F_1 score, as described above, also give meaningful insight. (The F_1 scores presented are for the 'positive' conditions; the value for the 'negative' condition would be different.)

All of our classifier configurations outperformed random guessing by a significant margin. Perhaps our most significant result comes when we try to use machine learning to sort the full set of four million headlines between their originating publications. We can delve into the internals of the collated data for more insight: Table 2

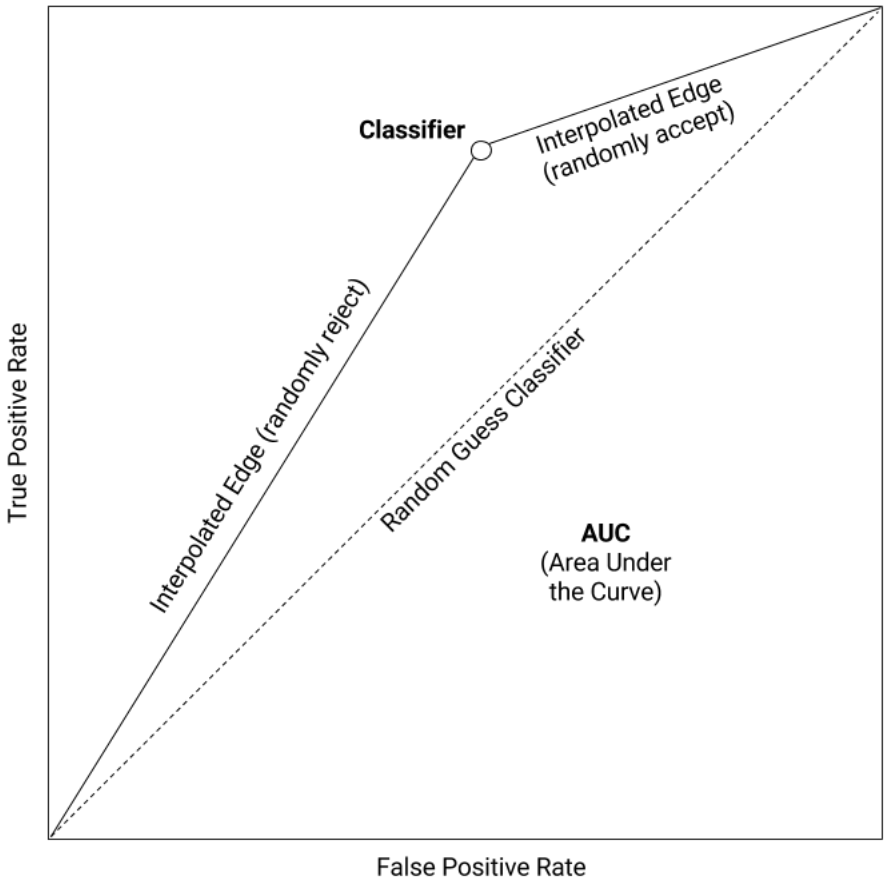

Fig. 2. Interpolated Receiver Operating Characteristic.

Table 1. Classifier Results

Criterion	Dataset	Accuracy	Blind Guess	ROC-AUC	F_1 score
publication month	ABC	0.169	0.083	-	-
publication season	ABC	0.346	0.250	-	-
publication year	ABC	0.225	0.067	-	-
first chronological half	ABC	0.710	0.500	0.710	0.717
published on weekend	ABC	0.816	0.592	0.619	0.356
published in 90 days before natl. election	ABC	0.916	0.861	0.508	0.047
published under Labor Party Prime Minister	ABC	0.648	0.508	0.634	0.563
original publication	ABC & Examiner.com	0.931	0.617	0.912	0.953

Table 2. Most Indicative Words

Rank	'ABC' Word	Explanation	'Examiner.com' Word	Explanation
1	qld	Abbreviation for Queensland	year-old	Human-interest stories
2	nsw	Abbreviation for New South Wales	ncaa	College basketball
3	nrn	Rural Australia news section	mom	Stories for & about mothers
4	australias	Originating country	spoilers	Story details of movies/TV
5	bendigo	Australian city and bank	rumors	Gossip stories
6	townsville	Australian city	dwts	'Dancing with the Stars', a TV show
7	gippsland	Australian region	honors	Headlines about awards
8	bikie	Australian slang for "biker"	dlc	'Downloadable content', a videogame term
9	aust	Abbreviation for Australia	gluten-free	Health/diet craze
10	bushfire	Australian term for "wildfire"	moms	Stories for & about mothers

presents the words most strongly identified with each of the ABC and Examiner.com. They scored highest in a formula that compares the effect in our Laplace-smoothed Bayesian equations:

$$\text{score}(\text{word}, \text{class}) = \frac{1 + \text{instances of word in class}}{(\text{number of classes} - 1) + \text{instances of words in all other classes}}$$

4 DISCUSSION

From the test results, it can be concluded that news headline data has a greater amount of informational value than previously assumed. This study has concluded that it is possible to extrapolate usable information in regard to a multitude of inquiries. Also, the approach outlined in this report can be used to predict the informational quality of the underlying article of which the headline represents.

5 WORKS CITED

- A practical explanation of a Naive Bayes classifier. (2017, October 03). Retrieved October 09, 2017, from https://monkeylearn.com/blog/practical-explanation-naive-bayes-classifier/
- Joyce, J. (2003, June 28). Bayes' Theorem. Retrieved October 09, 2017, from https://plato.stanford.edu/entries/bayes-theorem/
- A simple explanation of Naive Bayes Classification. (n.d.). Retrieved October 10, 2017, from https://stackoverflow.com/questions/1
simple-explanation-of-naive-bayes-classification#20556654
- Mitchell, T. M. (2015). Machine learning. Johanneshov: MTM.