News Categorization Case Study Summary:

Here given set of dataset contains headlines, URLs, and categories for 422,937 news stories collected by a web aggregator between March 10th, 2014 and August 10th, 2014. News categories in this dataset are labelled:

b: business;

t: science and technology;

e: entertainment; and

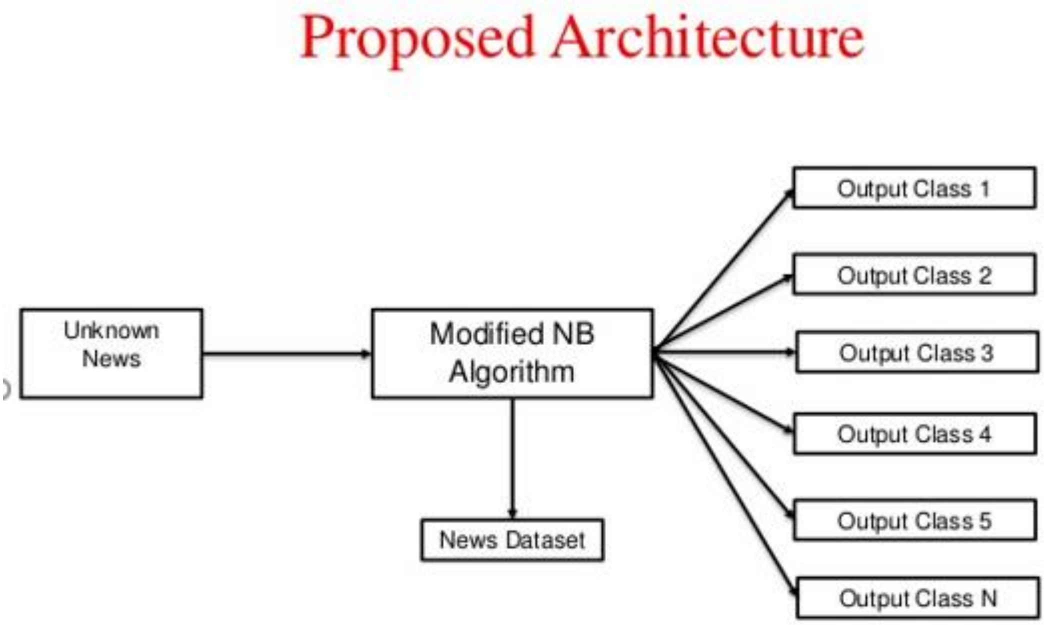
m: health.

It is Classification problem since here we need to classify or predict the category (business, entertainment, etc.) of a news article

Existing news portals on the WWW aim to provide users with numerous articles that are categorized into specific topics. Such a **categorization procedure improves presentation of the information** to the end-user. We further improve usability of these systems by presenting the architecture of a personalized news classification system that **exploits user's awareness of a topic in order to classify the articles in a 'peruser' manner.** The system's classification procedure bases upon a new text analysis and classification technique that represents documents using the vector space representation of their sentences.

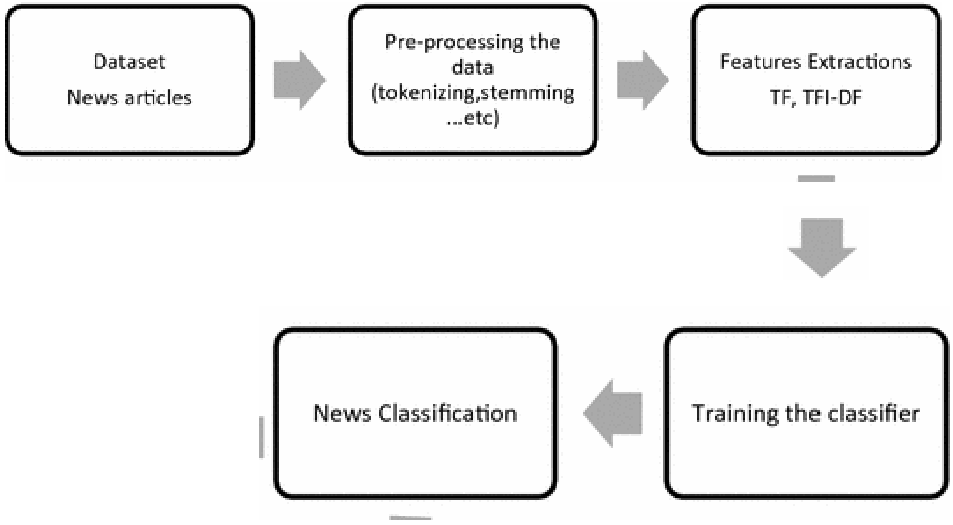
GENERAL ARCHITECTURE OF THE SYSTEM

The system1 consists of distributed sub-systems that cooperate in order to provide end-user with categorized news articles from the web that meet his personal needs.



Proposed Architecture

Steps involved:



1)Getting the data and load the data & do Cleaning of Data, Exploration of Data to make it consistent for Analysis.

2) Explore the data, and understand what the data is all about and feature selection.

Performance of the model depends.

Choice of Algorithm
Feature Selection
Feature Creation
Model selection

Feature selection is also known as variable selection.

Here in this case study we have variables ID, TITLE, URL, PUBLISHER, CATEGORY, STORY, HOSTNAME, TIMESTAMP where **by information value category of news purely depends on Head line of article i.e. Title, so selected this variable for model building.**

3)Choosing of Algorithm

what features do you have? What are their nature, what values can they have, what are their distributions?

Our predictor variable is a corpus and for text classification, Naïve Bayes is the best algorithm. Naive Bayes is a family of algorithms based on applying Bayes theorem with a strong(naive) assumption, that every feature is independent of the others, in order to predict the category of a given sample**. They are probabilistic classifiers, therefore will calculate the probability of each category using Bayes theorem, and the category with the highest** probability will be output. Naive Bayes classifiers have been successfully applied to many domains, particularly Natural Language Processing(NLP).

But, why Naive Bayes classifiers?

We do have other alternatives when coping with NLP problems, such as Support Vector Machine (SVM) and neural networks. However, the simple design of Naive Bayes classifiers make them very attractive for such classifiers. Moreover, they have been demonstrated to be fast, reliable and accurate in a number of applications of NLP.

4) Several algorithms were used to find the best fit. Here are the best fits:

Naive Bayes Classifier: 92.3% accuracy

So finally concluded the project by taking **Naive Bayes Classifier: 92.3% accuracy**