Fake News Detection

Abstract

Fake News Detection: This project employs Natural Language Processing and Machine Learning techniques to develop a robust system for automatically identifying and classifying fake news articles from genuine ones, contributing to the fight against misinformation and enhancing information integrity in the digital age.

Problem Definition

- The problem is to develop a fake news detection model using a Kaggle dataset.
- The goal is to distinguish between genuine and fake news articles based on their titles and text.
- This project involves using natural language processing (NLP) techniques to preprocess the text data, building a machine learning model for classification, and evaluating the model's performance.

Design Thinking

i) Data Source

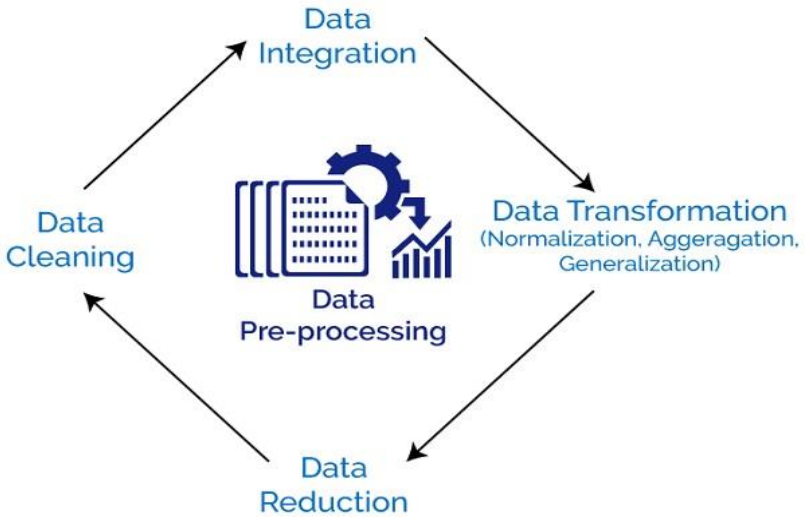
title	text	subject	date
Donald Trump Sends Out Embarrassing New Yearâ€™s Eve Message; This is t	Donald Trump just couldn t wish all Americans a Happy New Year and	News	December 31, 2017
Drunk Bragging Trump Staffer Started Russian Collusion Investigation	House Intelligence Committee Chairman Devin Nunes is going to have	News	December 31, 2017
Sheriff David Clarke Becomes An Internet Joke For Threatening To Poke Peop	On Friday, it was revealed that former Milwaukee Sheriff David Clarke	News	December 30, 2017
Trump Is So Obsessed He Even Has Obamaâ€™s Name Coded Into His Websi	On Christmas day, Donald Trump announced that he would be back t	News	December 29, 2017
Pope Francis Just Called Out Donald Trump During His Christmas Speech	Pope Francis used his annual Christmas Day message to rebuke Donal	News	December 25, 2017
Racist Alabama Cops Brutalize Black Boy While He Is In Handcuffs (GRAPHIC	The number of cases of cops brutalizing and killing people of color se	News	December 25, 2017
Fresh Off The Golf Course, Trump Lashes Out At FBI Deputy Director And Jan	Donald Trump spent a good portion of his day at his golf club, markin	News	December 23, 2017
Trump Said Some INSANELY Racist Stuff Inside The Oval Office, And Witness	In the wake of yet another court decision that derailed Donald Trump	News	December 23, 2017
Former CIA Director Slams Trump Over UN Bullying, Openly Suggests Heâ€™s	Many people have raised the alarm regarding the fact that Donald Tru	News	December 22, 2017
WATCH: Brand-New Pro-Trump Ad Features So Much A** Kissing It Will Mak	Just when you might have thought we d get a break from watching pe	News	December 21, 2017
Papa Johnâ€™s Founder Retires, Figures Out Racism Is Bad For Business	A centerpiece of Donald Trump s campaign, and now his presidency, h	News	December 21, 2017
WATCH: Paul Ryan Just Told Us He Doesnâ€™t Care About Struggling Familie	Republicans are working overtime trying to sell their scam of a tax bill	News	December 21, 2017
Bad News For Trump â€" Mitch McConnell Says No To Repealing Obamacar	Republicans have had seven years to come up with a viable replacem	News	December 21, 2017
WATCH: Lindsey Graham Trashes Media For Portraying Trump As â€˜Kooky,â	The media has been talking all day about Trump and the Republican P	News	December 20, 2017
Heiress To Disney Empire Knows GOP Scammed Us â€" SHREDS Them For Ta	Abigail Disney is an heiress with brass ovaries who will profit from the	News	December 20, 2017

Dataset Link: https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset

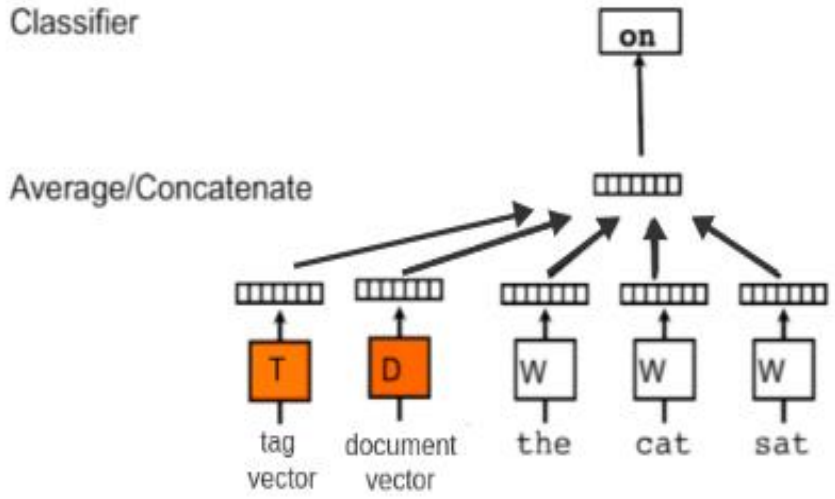
- Dataset source - Kaggle
- ID, Title , Text , Subject , Date
- Label 1 - Unreliable
- Label 0 – Reliable

ii)Data Preprocessing

- Perform various text cleaning steps (remove all non-alphanumeric characters, delete stopwords, delete missing rows, etc.)
- For Doc2Vec, convert to LabeledSentences(), comma separated word format



Doc2Vec Model



⧫ Based on Word2Vec model

⧫ Preserves word order information

⧫ Extracts Word2Vec features and adds an additional “document vector” with information about the entire document

iii) Feature Extraction

- Feature Extraction and Pre-Processing The embeddings used for the majority of our modelling are generated using the Doc2Vec model.
- The goal is to produce a vector representation of each article. Before applying Doc2Vec, we perform some basic pre-processing of the data.
- This includes removing stopwords, deleting special characters and punctuation, and converting all text to lowercase.
- This produces a comma-separated list of words, which can be input into the Doc2Vec algorithm to produce an 300-length embedding vector for each article.
- Doc2Vec is a model developed in 2014 based on the existing Word2Vec model, which generates vector representations for words [5].
- Word2Vec represents documents by combining the vectors of the individual words, but in doing so it loses all word order information. Doc2Vec expands on Word2Vec by adding a ”document vector” to the output representation, which contains some information about the document as a whole, and allows the model to learn some information about word order.
- Preservation of word order information makes Doc2Vec useful for our application, as we are aiming to detect subtle differences between text documents.

iv)Model Selection

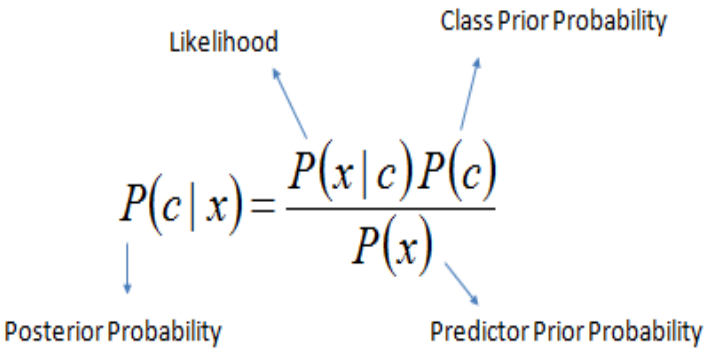
1) Naive Bayes

♦ Classification technique based on Bayes' theorem with an assumption of independence among predictors

1. Convert data set into a frequency table

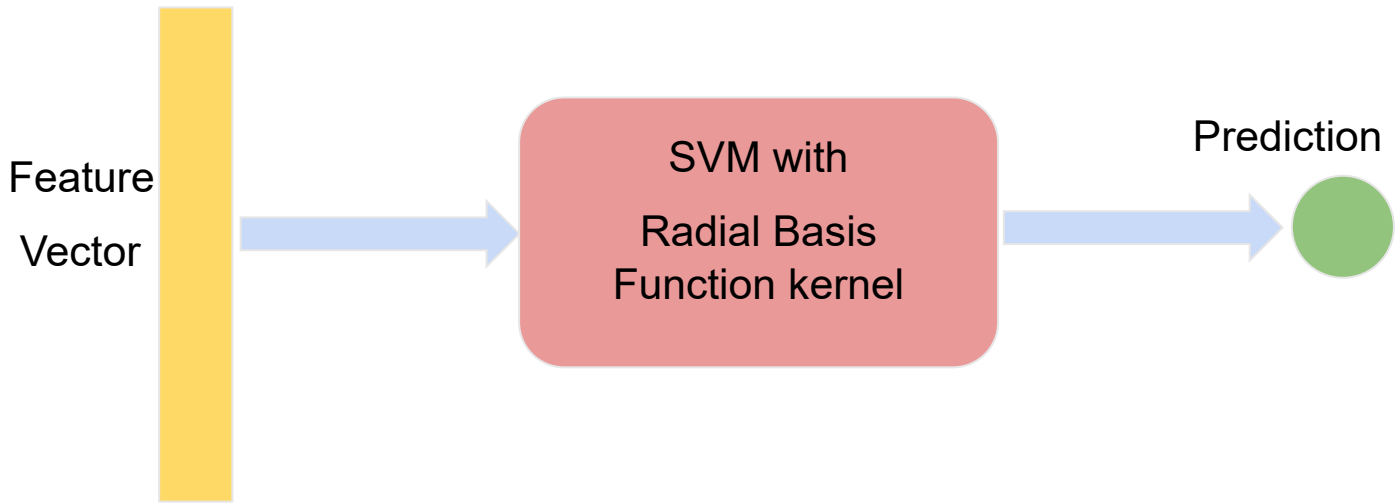
2. Create likelihood table by finding probabilities

3. Use Naive Bayesian equation to calculate posterior probability for each class



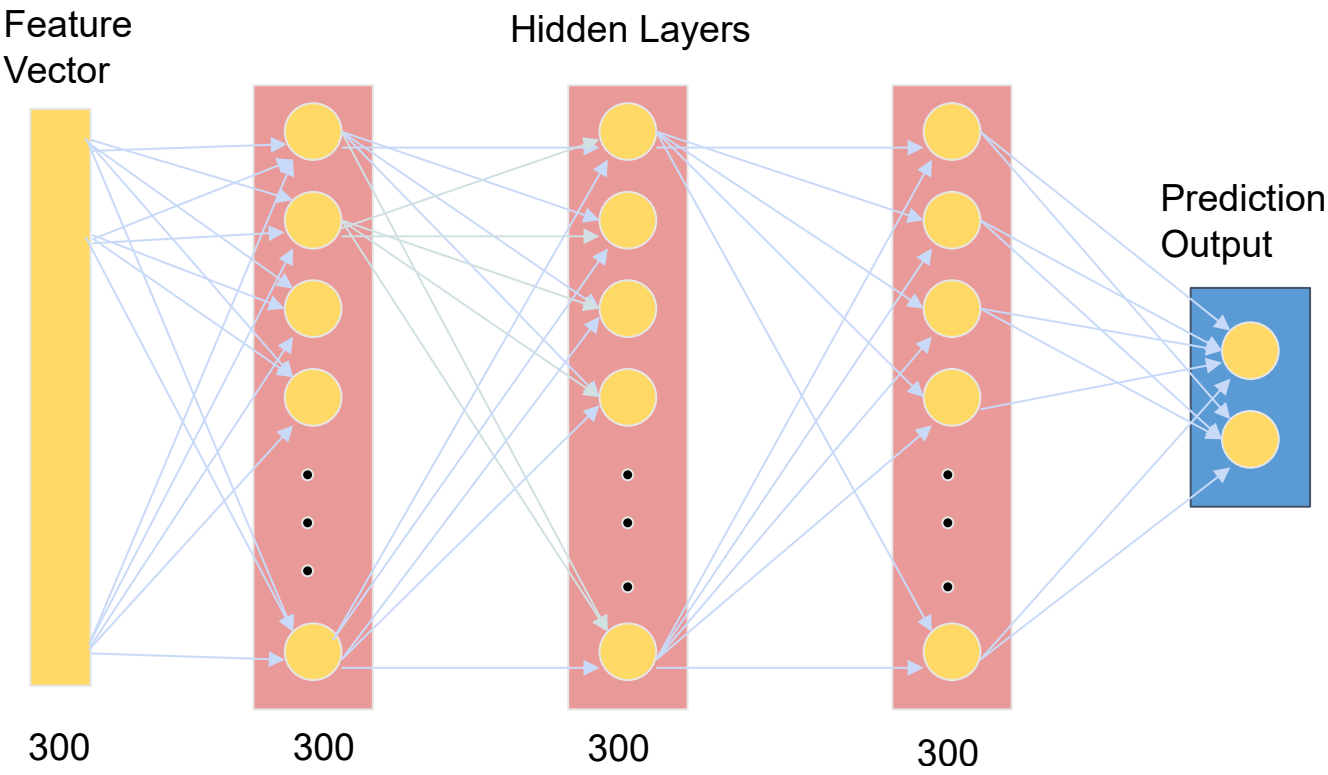
$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

2) Support Vector Machine (SVM)



A Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification and regression tasks. It's particularly popular for classification problems. SVMs are powerful and versatile because they can handle both linear and non-linear data by finding an optimal hyperplane that best separates different classes or predicts numerical values.

3) Neural Network

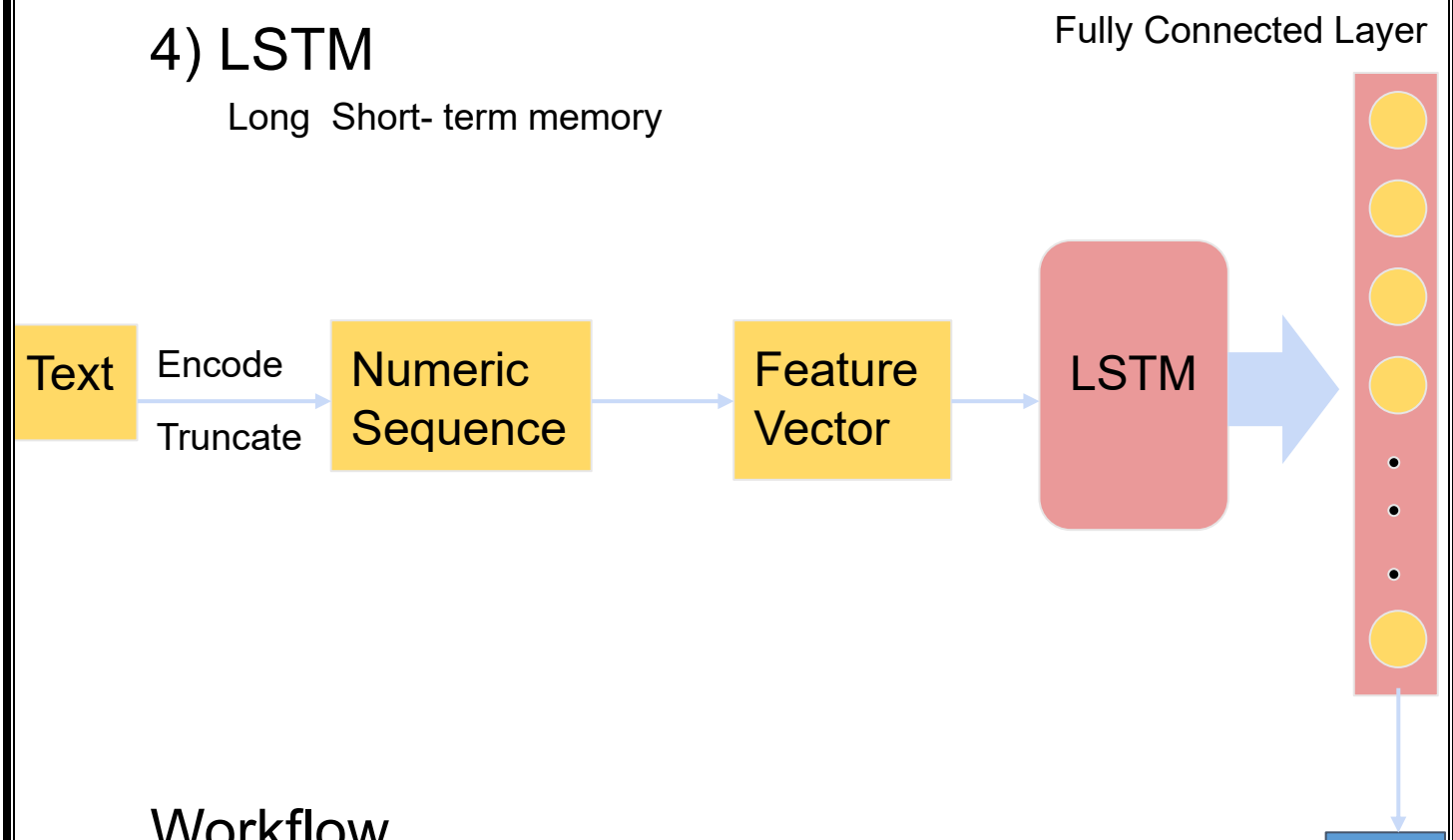


Feed-forward Neural Network

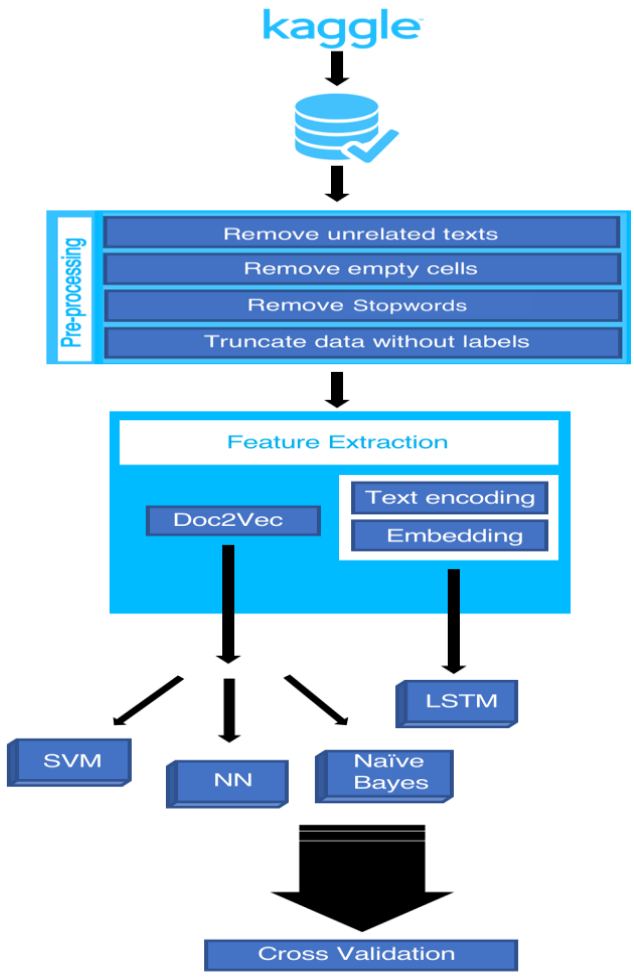
- Feed-forward Neural Network We implemented two feed-forward neural network models, one using Tensorflow and one using Keras.
- Neural networks are commonly used in modern NLP applications [7], in contrast to older approaches which primarily focused on linear models such as SVM's and logistic regression.
- Our neural network implementations use three hidden layers. In the Tensorflow implementation, all layers had 300 neurons each, and in the Keras implementation we used layers of size 256, 256, and 80, interspersed with dropout layers to avoid overfitting.

4) LSTM

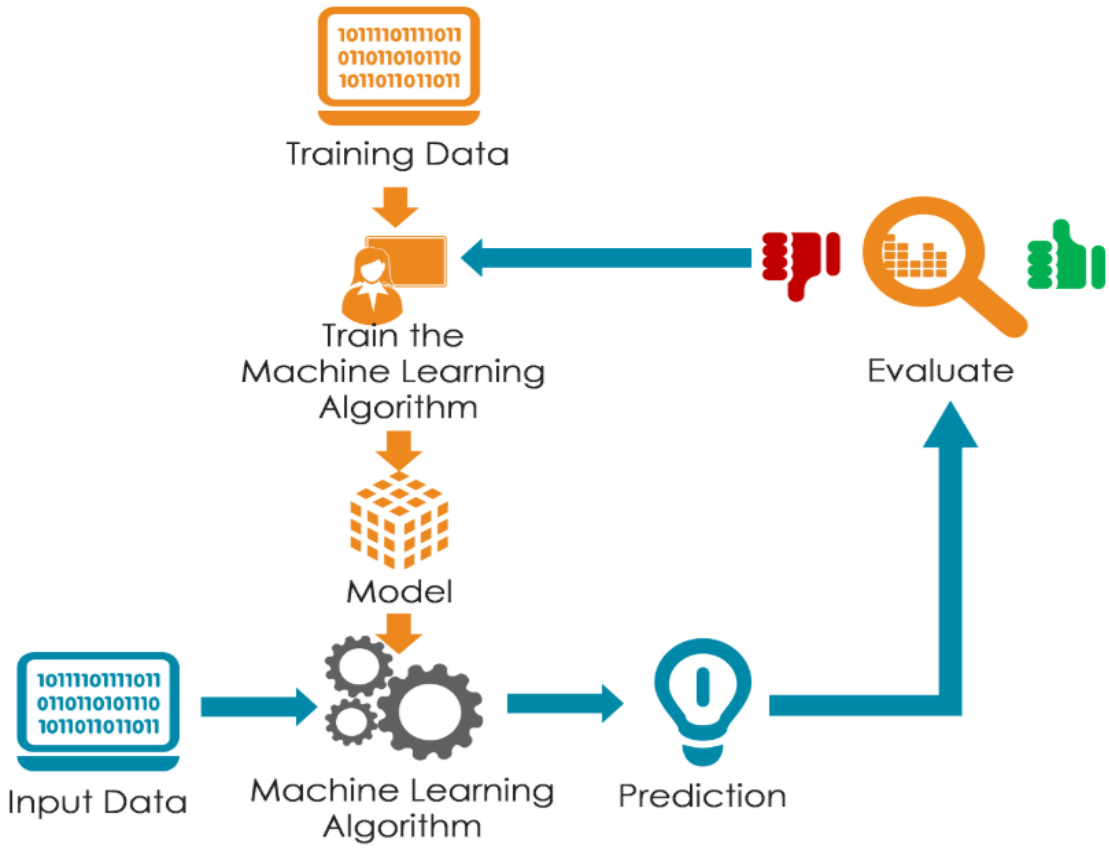
Long Short- term memory



Workflow



v)Model Training



♦ Models used-

- Naive Bayes
- Support Vector Machine (SVM)
- Neural Network
- Long Short-Term Memory (LSTM)

vi)Evaluation

Model	Accuracy
Naive Bayes	72.94%
SVM	88.42%
Neural Network using TensorFlow	81.42%
Neural Network using Keras	92.62%
LSTM	94.53%