SUBJECT: ISM 6359
DATA MINING

TOPIC: FAKE NEWS DETECTION

SHEETAL MURALI

BUSINESS REASON:
FAKE NEWS DETECTION
TOOL: RAPID MINER

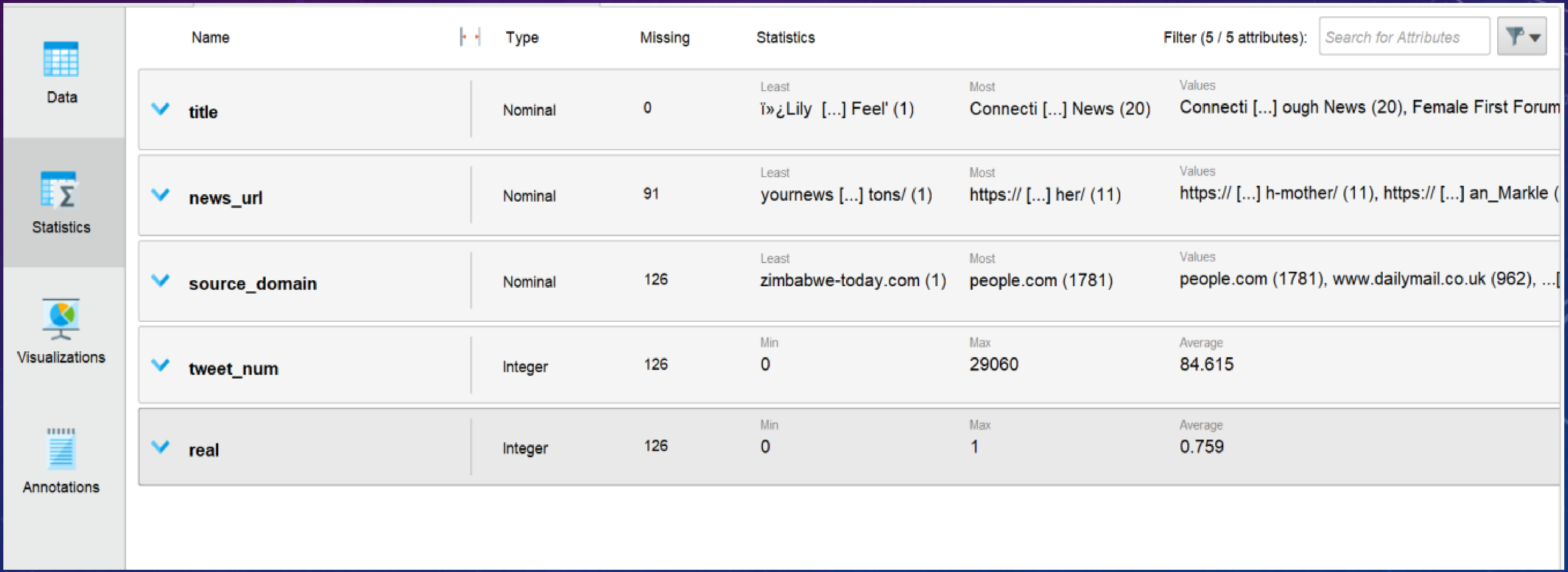
- In today’s world of Social Media, one of the major problem we face is misinformation.
- Millions of articles are being published every minute-the scope of human manual detection of a real or fake news is not feasible.
- The model created will test the unseen data, and accordingly detect fake articles and can be used and integrated with any system for future use.

DATA STRUCTURE

- The data I have collected is from Kaggle.com. This data consists of the following information:
- title: title of the article – 21724 Unique Values (Sampled it down to 5000)
- news_url: URL of the article
- source domain: web domain where article was posted.
- tweet_num: number of retweets for this article.
- real: label column, where 1 is 'real' and 0 is 'fake'.

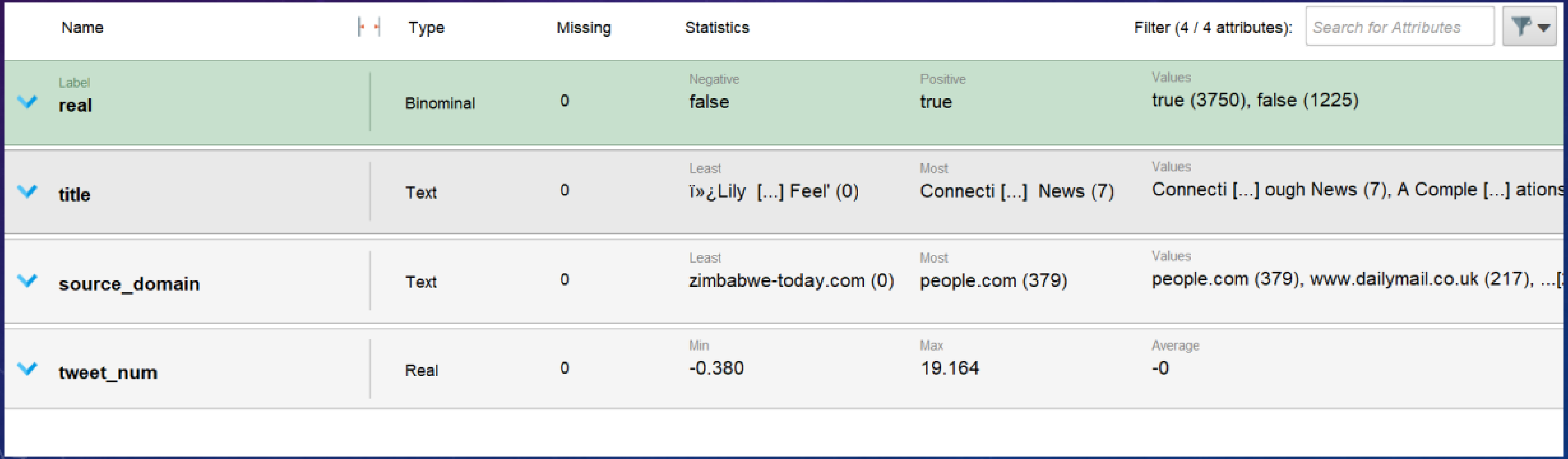
TEXT PREPARATION

- Loading the data



Name	Type	Missing	Statistics		
title	Nominal	0	Least ï»¿Lily [...] Feel' (1)	Most Connecti [...] News (20)	Values Connecti [...] ough News (20), Female First Forum
news_url	Nominal	91	Least yournews [...] tons/ (1)	Most https:// [...] her/ (11)	Values https:// [...] h-mother/ (11), https:// [...] an_Markle (
source_domain	Nominal	126	Least zimbabwe-today.com (1)	Most people.com (1781)	Values people.com (1781), www.dailymail.co.uk (962), ...[
tweet_num	Integer	126	Min 0	Max 29060	Average 84.615
real	Integer	126	Min 0	Max 1	Average 0.759

- Set Role – Label “Real”
- Select Attributes- Remove irrelevant attribute “URL”
- Numerical to Binomial – Label
- Filter Examples – Missing values
- Sample – 5000 random values
- Normalize – number of tweets ranging from (from ‘0’ to ‘29060’)
- Nominal to Text



Filter (4 / 4 attributes): Search for Attributes

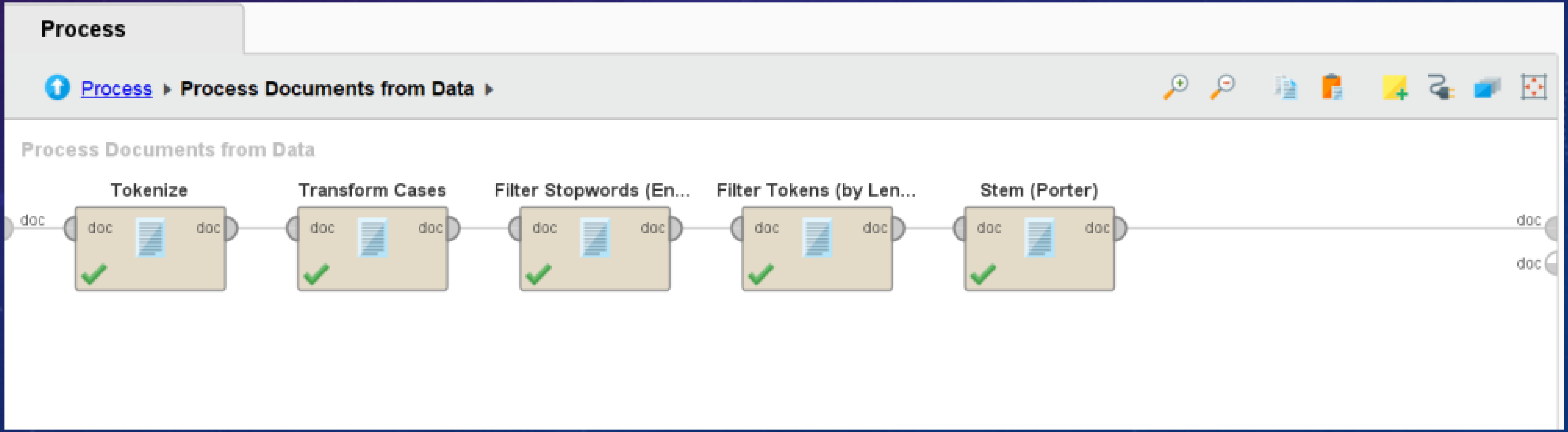
Name	Type	Missing	Statistics		
Label real	Binominal	0	Negative false	Positive true	Values true (3750), false (1225)
title	Text	0	Least ï»¿Lily [...] Feel' (0)	Most Connecti [...] News (7)	Values Connecti [...] ough News (7), A Comple [...] ations
source_domain	Text	0	Least zimbabwe-today.com (0)	Most people.com (379)	Values people.com (379), www.dailymail.co.uk (217), ...[
tweet_num	Real	0	Min -0.380	Max 19.164	Average -0

PROCESS DOCUMENT FROM DATA

- Tokenization
- Transform Cases
- Filter Stopwords
- Filter Token by length
- Stemming-Porter

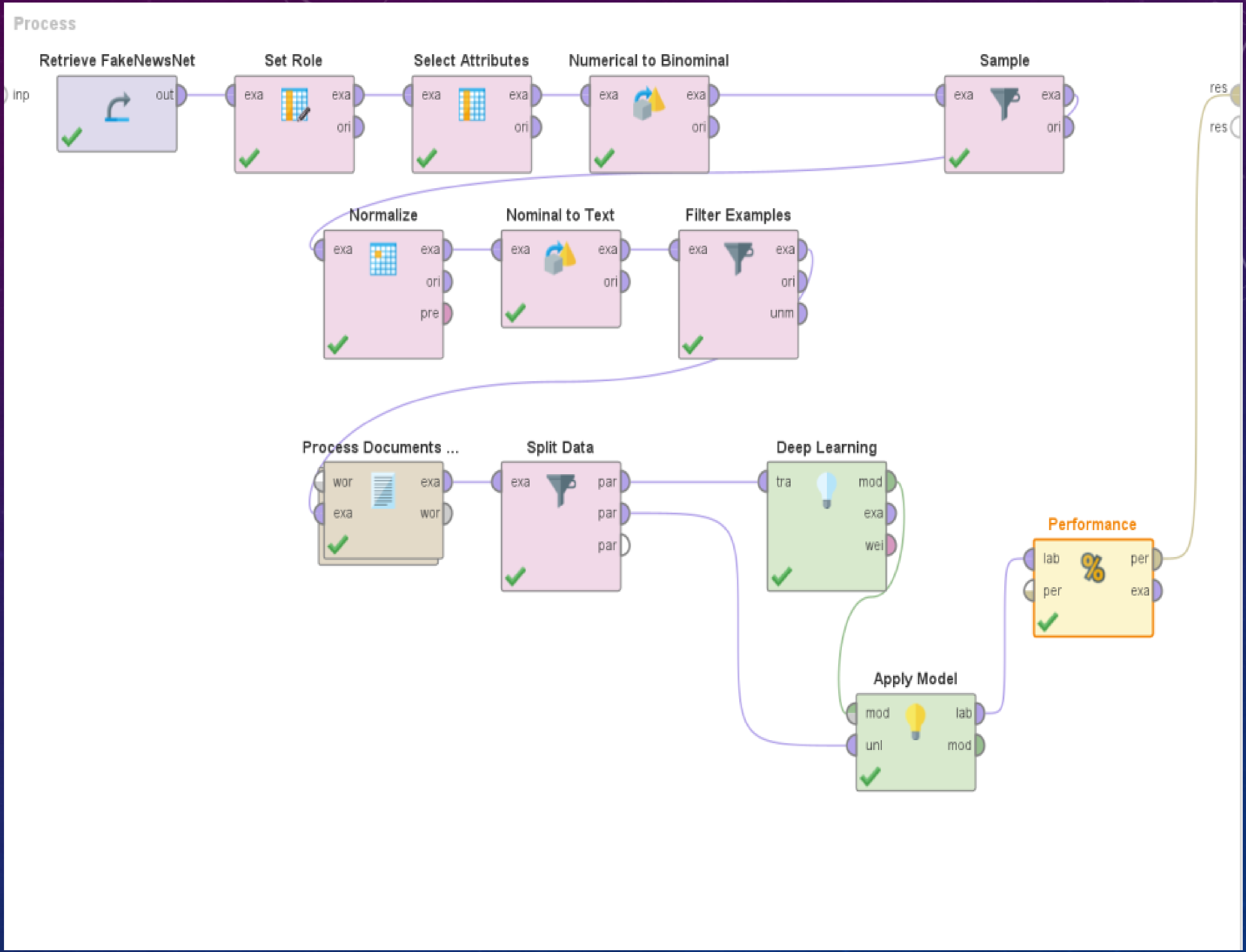
SPLIT DATA: 70% Training set

30% Test set



ALGORITHMS USED

- Naïve Bayes
- Decision Tree
- Random forest
- Deep Learning
- SVM


Process
Retrieve FakeNewsNet
inp
out
Set Role
exa
exa
ori
Select Attributes
exa
exa
ori
Numerical to Binominal
exa
exa
ori
Sample
exa
exa
ori
res
res
Normalize
exa
exa
ori
pre
Nominal to Text
exa
exa
ori
Filter Examples
exa
exa
ori
unm
Process Documents ...
wor
exa
exa
wor
Split Data
exa
par
par
par
Deep Learning
tra
mod
exa
wei
Performance
lab
per
per
exa
Apply Model
mod
unl
lab
mod

ANALYSIS OF THE DATA

Operators	Attribute count
Tokenization	12282
Transform cases	10079
Filter Stopwords	9804
Filter tokens by length	9171
Stem	7538

Accuracy for different Algorithms

Model	Split Data 70%-30%
Decision Tree	78.19%
Random Forest	75.38%
Sequential Minimal Optimization (SVM)	81.51%
Naïve Bayes	72.28%
Deep Learning	81.57%

PERFORMANCE BASED ON WEIGHTS

Binary Term Occurrences	77.39%
Term Occurrences	79.2%
Term-Frequency	75.88%
TF-IDF	81.57%

DATA SOURCE:

I found the dataset used to build this model on Kaggle.com

https://www.kaggle.com/datasets/algord/fake-news?select=FakeNewsNet.csv

3 W'S

1) What went well?

Topics taught in class could be easily related to the business problem I was solving. Video lectures helped me simultaneously work on RapidMiner as we kept learning new features.

2) What did not go well?

Initially I started the project with 20000+ documents, but text mining operator did run because of space constraints on the laptop due to which I had to reduce the data to 5000 examples by using random sampling

3) What would you use Differently Next Time?

Since the local host does not have enough memory to run Text Mining on a large data set, I would explore cloud services like AWS to scale my algorithms on a better computive and memory intensive host.