BT5153 Final Project Report
Team Area51

Name	Student ID	Email
Koey Huixin	A0045254L	koey.huixin@u.nus.edu
Li Wei Cheng	A0186371Y	e0320609@u.nus.edu
Lin Congren	A0056412N	lin.congren@u.nus.edu
Ng Pui Yee	A0206669U	puiyee.ng@u.nus.edu
Tay Choon Wee	A0056188W	tay.choon.wee@u.nus.edu

1. OVERVIEW

In the Digital Age, news and stories get circulated at tremendous speed with the help of social media platforms such as Facebook and WhatsApp. While technological aids have helped to make news and information more accessible, the rise of fake news in the world has become a growing concern in recent years.

The intent of fabricating and spreading fake news is not always malicious. Some people may fabricate fake news simply "for fun", while others simply forward questionable news articles without checking the source. Regardless of the intent, the circulation of fake news makes it difficult for society to identify which news articles are real, and which are not, which may result in public confusion and unnecessary panic in some cases.

Some measures have been put in place to combat the circulation of fake news in Singapore. For instance, the Singapore Government passed the Protection from Online Falsehoods and Manipulation Act in 2019 to help combat the circulation of fake news in Singapore. Under this act, individuals who knowingly disseminate online false statements that threaten the security of Singapore may face a fine or even imprisonment. While this will certainly serve as a deterrence to the circulation of fake news, the public may not always be able to differentiate between the truth and lies.

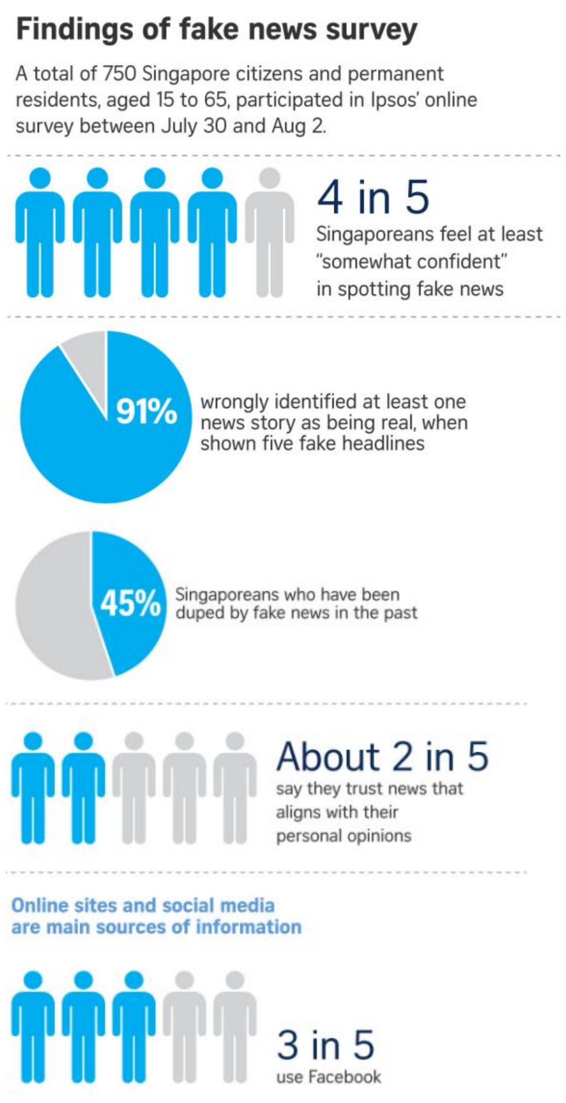


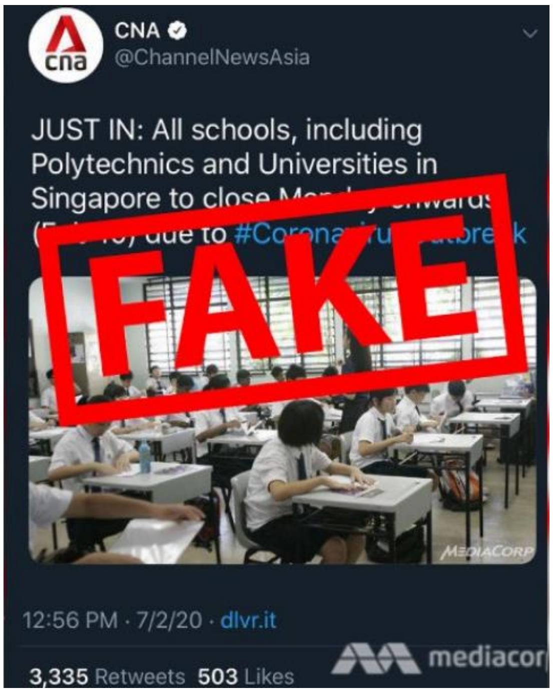
Figure 1 - Infographics on fake news survey in Singapore

In 2018, a survey was conducted in Singapore by global independent market research agency, Ipsos *(refer to Figure 1)*. The result shows four in five Singaporeans say that they can confidently spot fake news, but when put to the test, about 90 per cent mistakenly identified at least one out of five fake headlines as being real *(Ng, H., 2018)*. Hence, we believe that more can be done to help the public distinguish between real and fake news, and this became the genesis of this project.

Through this project, we aim to create a machine learning model with the task (T) of classifying online news articles based on their reliability. The performance (P) will be measured by the accuracy of the fake news articles that are detected, and the experience (E) will be historical news articles that have been labeled fake or reliable.

The business application of this project is twofold. At a consumer level, it can be used by the public to help identify fake news and scams, reducing their likelihood of falling for them. For instance, in the recent and ongoing case of the COVID-19 outbreak, there has been many cases of fake news such as school closure and fake DORSCON escalation report from MOH circulating around social media and online resources. By identifying such fake news early, we can help to control spreading the wrong information to the people around us and create unnecessary public panic.

At a business level, news outlets can use machine-learning models like this to distinguish between fake and genuine reports of new events, allowing them to better deliver breaking news with a low risk of public embarrassment. Government agencies such as the Singapore Police Force can also avoid unnecessary investigations arising from such fake reports.

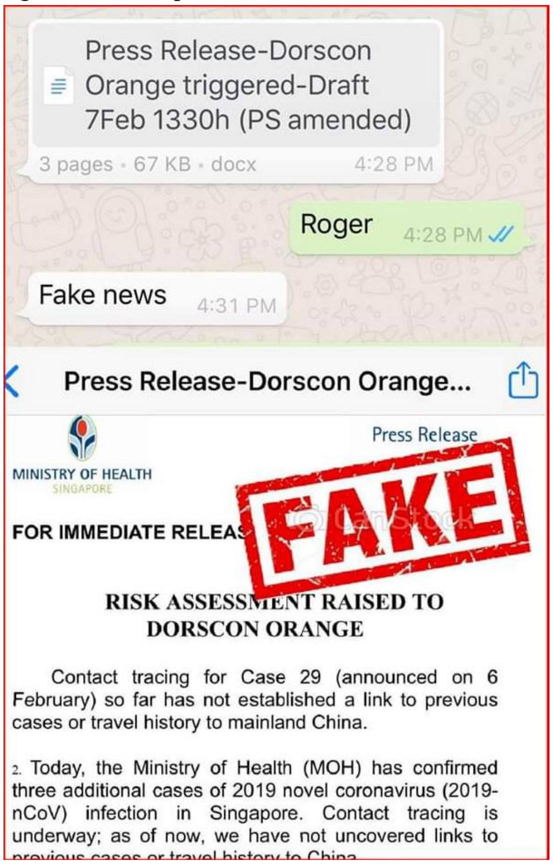

CNA
@ChannelNewsAsia

JUST IN: All schools, including Polytechnics and Universities in Singapore to close M

12:56 PM · 7/2/20 · dlvr.it

3,335 Retweets 503 Likes

Figure 2 - Example of fake news debunked


Press Release-Dorscon Orange triggered-Draft 7Feb 1330h (PS amended)
3 pages · 67 KB · docx 4:28 PM

Roger 4:28 PM

Fake news 4:31 PM

Press Release-Dorscon Orange...

MINISTRY OF HEALTH
SINGAPORE

Press Release

FOR IMMEDIATE RELEAS

RISK ASSESSMENT RAISED TO DORSCON ORANGE

Contact tracing for Case 29 (announced on 6 February) so far has not established a link to previous cases or travel history to mainland China.

2. Today, the Ministry of Health (MOH) has confirmed three additional cases of 2019 novel coronavirus (2019-nCoV) infection in Singapore. Contact tracing is underway; as of now, we have not uncovered links to previous cases or travel history to China.

Figure 3 - Example of fake news debunked

2. DATA EXPLORATION

The dataset chosen for this project is the FakeNewsCorpus obtained from Github (*Szpakowski M., 2018*), which is an open source dataset composed of millions of news articles collected from a curated list of 1001 domains. It contains 9 news csv files format with 27GB of data and represents 15 features with 12 categories of news type. Out of the 12 news types *(refer to Table 1)*, **fake** and **reliable** are selected to identify fake news from our problem statement while other news types were not used in this study. This is because the other news types appear to consist mainly of opinion pieces which are not objectively true or false. Hence, they do not fit the business question for the fake news classifier.

Table 1 - News types from the dataset.

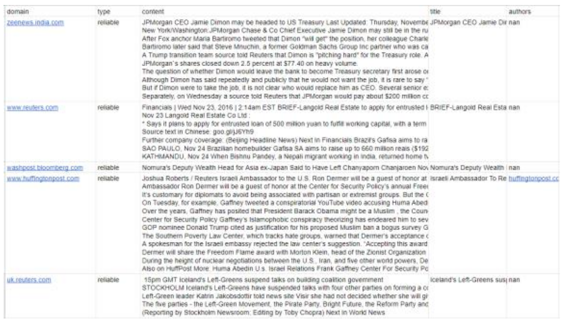
News Type	**Description**
fake	Sources that entirely fabricate information, disseminate deceptive content, or grossly distort actual news reports
satire	Sources that use humor, irony, exaggeration, ridicule, and false information to comment on current events.
bias	Sources that come from a particular point of view and may rely on propaganda, decontextualized information, and opinions distorted as facts.
conspiracy	Sources that are well-known promoters of kooky conspiracy theories.
state	Sources in repressive states operating under government sanction.
junksci	Sources that promote pseudoscience, metaphysics, naturalistic fallacies, and other scientifically dubious claims.
hate	Sources that actively promote racism, misogyny, homophobia, and other forms of discrimination.
clickbait	Sources that provide generally credible content, but use exaggerated, misleading, or questionable headlines, social media descriptions, and/or images.
unreliable	Sources that may be reliable but whose contents require further verification.
political	Sources that provide generally verifiable information in support of certain points of view or political orientations.
reliable	Sources that circulate news and information in a manner consistent with traditional and ethical practices in journalism.

Out of the 15 features, the following features were selected for the machine learning solution (*refer to Table 2*).

Table 2 - Selected features from the dataset.

Feature Name	**Description**
Domain	URL Source of the news article
Type	Label for the news (fake vs reliable)
Content	The content body for the news
Author	Author of the news article
Title	The title for the news

Data cleaning process involves removal of duplicates and null contents from Fake and Reliable news categories. Prior to data exploration and model learning, data was chunked into appropriate size during reading of the raw data into our local computer to extract 100,000 rows from each of the fake and reliable news datasets to build our machine learning models.



domain	type	content	title	authors
zeenews.india.com	reliable	JPMorgan CEO Jamie Dimon may be headed to US Treasury Last Updated: Thursday, Novembe New York/Washington: JPMorgan Chase & Co Chief Executive Jamie Dimon may still be in the ru After Fox anchor Maria Bartiromo tweeted that Dimon "will get" the position, her colleague Charle Bartiromo later said that Steve Mnuchin, a former Goldman Sachs Group Inc partner who was ca A Trump transition team source told Reuters that Dimon is "pitching hard" for the Treasury role. A JPMorgan's shares closed down 2.5 percent at $77.40 on heavy volume. The question of whether Dimon would leave the bank to become Treasury secretary first arose o Although Dimon has said repeatedly and publicly that he would not want the job, it is rare to say " But if Dimon were to take the job, it is not clear who would replace him as CEO. Several senior e Separately, on Wednesday a source told Reuters that JPMorgan would pay about $200 million cc	JPMorgan CEO Jamie Dir	nan
www.reuters.com	reliable	Financials | Wed Nov 23, 2016 | 2:14am EST BRIEF-Langold Real Estate to apply for entrusted l Nov 23 Langold Real Estate Co Ltd : * Says it plans to apply for entrusted loan of 500 million yuan to fulfil working capital, with a term Source text in Chinese: goo.gl/jJ6Yh9 Further company coverage: (Beijing Headline News) Next In Financials Brazil's Gafisa aims to ra SAO PAULO, Nov 24 Brazilian homebuilder Gafisa SA aims to raise up to 660 million reais ($192 KATHMANDU, Nov 24 When Bishnu Pandey, a Nepali migrant working in India, returned home tv	BRIEF-Langold Real Esta	nan
washpost.bloomberg.com	reliable	Nomura's Deputy Wealth Head for Asia ex-Japan Said to Have Left Chanyaporn Chanjaroen Nov	Nomura's Deputy Wealth	nan
www.huffingtonpost.com	reliable	Joshua Roberts / Reuters Israeli Ambassador to the U.S. Ron Dermer will be a guest of honor at Ambassador Ron Dermer will be a guest of honor at the Center for Security Policy's annual Free It's customary for diplomats to avoid being associated with partisan or extremist groups. But the C On Tuesday, for example, Gaffney tweeted a conspiratorial YouTube video accusing Huma Abed Over the years, Gaffney has posited that President Barack Obama might be a Muslim , the Coun Center for Security Policy Gaffney's Islamophobic conspiracy theorizing has endeared him to sev GOP nominee Donald Trump cited as justification for his proposed Muslim ban a bogus survey G The Southern Poverty Law Center, which tracks hate groups, warned that Dermer's acceptance c A spokesman for the Israeli embassy rejected the law center's suggestion. "Accepting this award Dermer will share the Freedom Flame award with Morton Klein, head of the Zionist Organization During the height of nuclear negotiations between the U.S., Iran, and five other world powers, De Also on HuffPost More: Huma Abedin U.s. Israel Relations Frank Gaffney Center For Security Po	Israeli Ambassador To Re	huffingtonpost.cc
uk.reuters.com	reliable	.15pm GMT Iceland's Left-Greens suspend talks on building coalition government STOCKHOLM Iceland's Left-Greens have suspended talks with four other parties on forming a co Left-Green leader Katrin Jakobsdottir told news site Visir she had not decided whether she will gi The five parties - the Left-Green Movement, the Pirate Party, Bright Future, the Reform Party and (Reporting by Stockholm Newsroom; Editing by Toby Chopra) Next In World News	Iceland's Left-Greens sus	nan

Figure 4 - Sample screenshot of dataset with selected features

Next, n-grams word clouds are processed through the cleaned dataset of fake and reliable news types.

From the Uni-gram word cloud (*refer to Figure 5*), fake and reliable news share similar terms such as "news" and "people". However, terms such as "source", "http", and "fact" appear unexpectedly in the Fake news category. They may be used to increase reliability and gain trust from the readers.

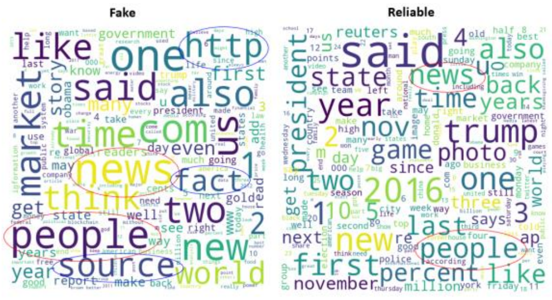


Figure 5 - Uni-gram word cloud

From the Bi-gram word cloud (*refer to Figure 6*), reliable news can be seen as association with "new york", "united states", "donald trump" compared to "story_fact", "readers_think", "http_www" which frequently appears on Fake news.

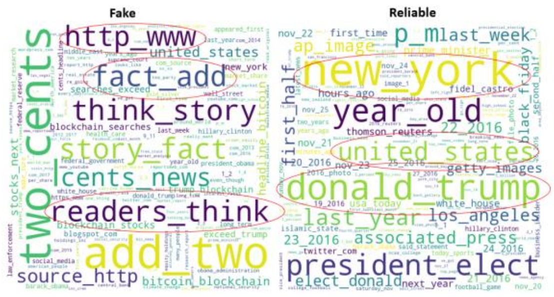


Figure 6 - Bi-gram word cloud

Lastly, in the Tri-gram word cloud (*refer to Figure 7*), reliable news capture readers' attention through important terms such as "president", "usa", "donald_trump" and "barack_obama". Date is also a significant feature in reliable news. Apart from "fact", "story" terms found continuously on fake news, "blockchain" is another word to catch audience attention.

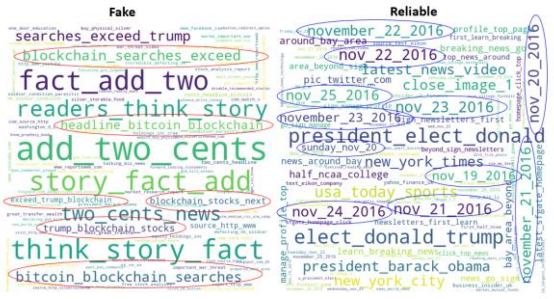


Figure 7 - Tri-gram word cloud

3. SYSTEM ARCHITECTURE

Considering the size of the dataset and complexity of text transformation, PySpark program was adopted to facilitate the study. The platform being used in this study is Google Cloud Platform - a Google bucket was created as the file system to store the dataset and notebooks, another function being used is Google VM Cluster. The cluster was created containing 5 VMs, each providing 4 CPUs and 25 GB RAM. The VM environment was configured with Debian 9, Hadoop 2.9 and Spark 2.4. While creating the cluster, the bucket was associated with the cluster and enabled anaconda service and jupyter notebook, so that the cluster can automatically load the notebook into Jupyter.

Figure 8 illustrates the system architecture setup explaining how all the resources were linked and worked together.

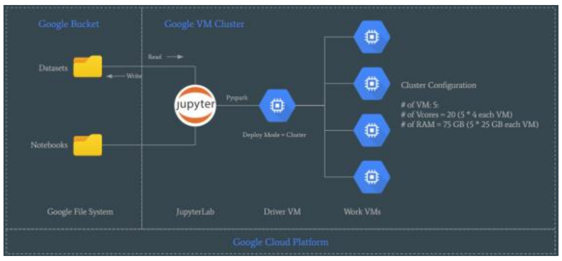
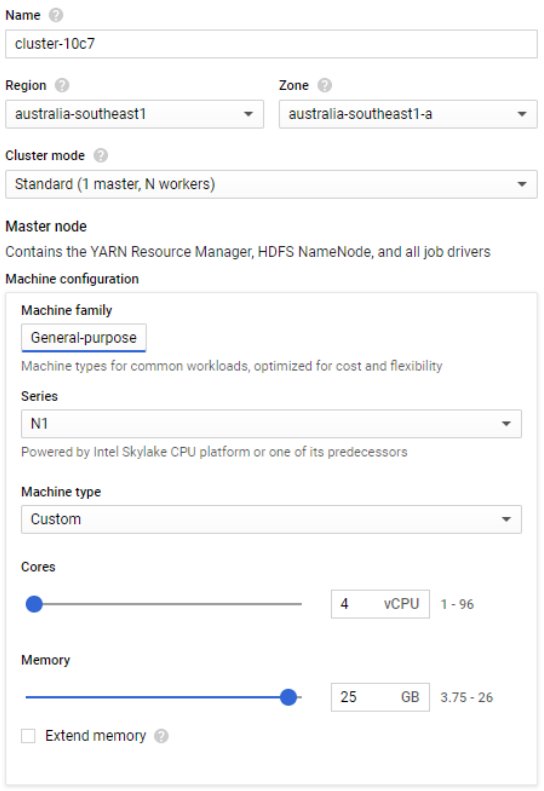


Figure 8 - System Architecture Diagram

On the side note, in order to maximize the cluster computing power, the "spark deploy mode" was set as "cluster", so that the cluster can choose any free VM as the driver node and rest of the VMs as the work nodes (*refer to Figures 9 to 11*). With that setup, multiple jobs can be performed concurrently to allow the development process to be more efficient.


Name
cluster-10c7

Region: australia-southeast1
Zone: australia-southeast1-a

Cluster mode
Standard (1 master, N workers)

Master node
Contains the YARN Resource Manager, HDFS NameNode, and all job drivers
Machine configuration

Machine family
General-purpose
Machine types for common workloads, optimized for cost and flexibility

Series
N1
Powered by Intel Skylake CPU platform or one of its predecessors

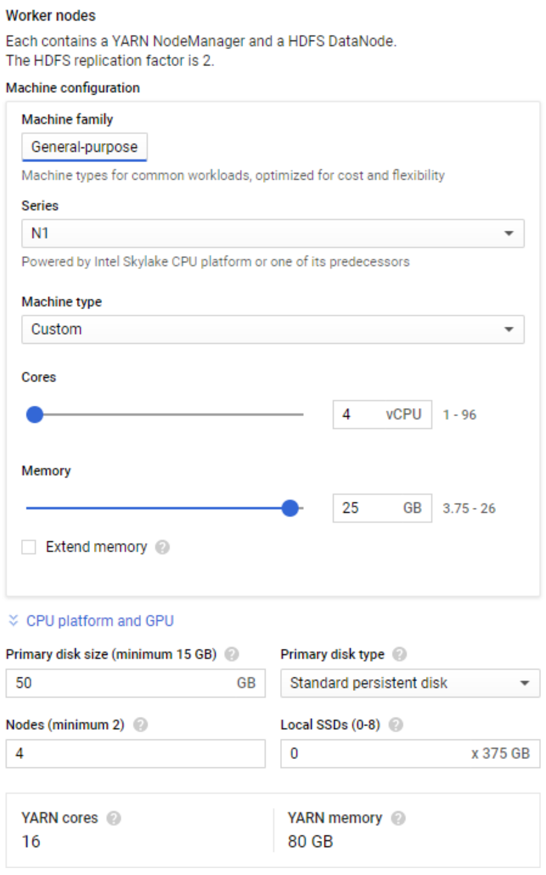
Machine type
Custom

Cores
4 vCPU 1 - 96

Memory
25 GB 3.75 - 26

Extend memory

Figure 9 - Master Node Configuration


Worker nodes
Each contains a YARN NodeManager and a HDFS DataNode.
The HDFS replication factor is 2.
Machine configuration

Machine family
General-purpose
Machine types for common workloads, optimized for cost and flexibility

Series
N1
Powered by Intel Skylake CPU platform or one of its predecessors

Machine type
Custom

Cores
4 vCPU 1 - 96

Memory
25 GB 3.75 - 26

Extend memory

CPU platform and GPU

Primary disk size (minimum 15 GB): 50 GB
Primary disk type: Standard persistent disk

Nodes (minimum 2): 4
Local SSDs (0-8): 0 x 375 GB

YARN cores: 16
YARN memory: 80 GB

Figure 10 - Work Nodes Configuration

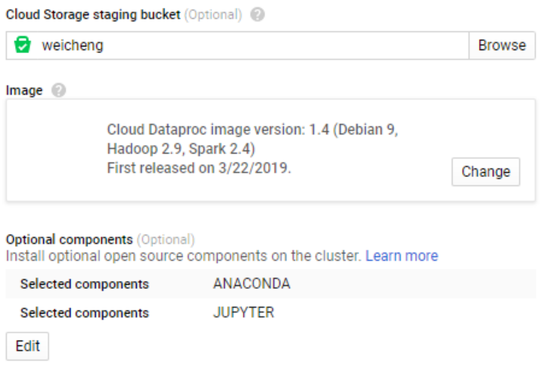

Cloud Storage staging bucket (Optional)
weicheng
Browse

Image
Cloud Dataproc image version: 1.4 (Debian 9, Hadoop 2.9, Spark 2.4)
First released on 3/22/2019.
Change

Optional components (Optional)
Install optional open source components on the cluster. Learn more

Selected components	ANACONDA
Selected components	JUPYTER

Edit

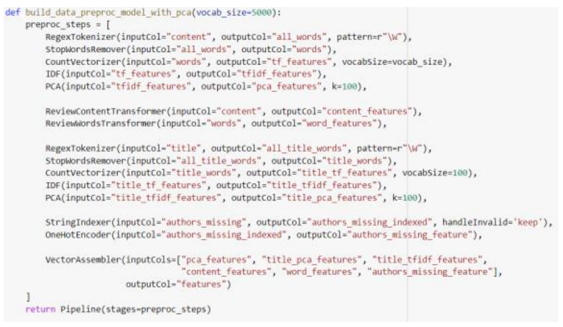
Figure 11 - Cluster Services Configuration

4. DATA PRE-PROCESSING

Most of our pre-processing steps focused on tokenizing existing unstructured text fields such as “Content” and “Title”. Only 1 field was manually created i.e. Missing_Authors.

A pipeline was built consisting of the following steps in data pre-processing:

- RegexTokenizer to take in only alphanumeric characters
- StopWordsRemover to remove common words such as “I”, “and”, “the” etc
- CountVectorizer to convert a collection of text documents to a matrix of token counts
- IDF to measure the importance of every term. Specifically, it is a mathematical calculation IDF(t) = log_e(Total number of documents / Number of documents with term t in it).
- Principal Component Analysis (PCA) on tf-idf features to reduce n-dimensional data to k-dimensional data to speed things up in machine learning. In our case, we set k = 100.



```python
def build_data_preproc_model_with_pca(vocab_size=5000):
    preproc_steps = [
        RegexTokenizer(inputCol="content", outputCol="all_words", pattern=r"\W"),
        StopWordsRemover(inputCol="all_words", outputCol="words"),
        CountVectorizer(inputCol="words", outputCol="tf_features", vocabSize=vocab_size),
        IDF(inputCol="tf_features", outputCol="tfidf_features"),
        PCA(inputCol="tfidf_features", outputCol="pca_features", k=100),

        ReviewContentTransformer(inputCol="content", outputCol="content_features"),
        ReviewWordsTransformer(inputCol="words", outputCol="word_features"),

        RegexTokenizer(inputCol="title", outputCol="all_title_words", pattern=r"\W"),
        StopWordsRemover(inputCol="all_title_words", outputCol="title_words"),
        CountVectorizer(inputCol="title_words", outputCol="title_tf_features", vocabSize=100),
        IDF(inputCol="title_tf_features", outputCol="title_tfidf_features"),
        PCA(inputCol="title_tfidf_features", outputCol="title_pca_features", k=100),

        StringIndexer(inputCol="authors_missing", outputCol="authors_missing_indexed", handleInvalid='keep'),
        OneHotEncoder(inputCol="authors_missing_indexed", outputCol="authors_missing_feature"),

        VectorAssembler(inputCols=["pca_features", "title_pca_features", "title_tfidf_features",
                                   "content_features", "word_features", "authors_missing_feature"],
                        outputCol="features")
    ]
    return Pipeline(stages=preproc_steps)
```

Figure 12 - Code snippets for data pre-processing including RegexTokenizer, StopWordsRemover, CountVectorizer, IDF and PCA transformation.

Finally, the data was split into 80% training and 20% testing. The training dataset was used to feed into building models, whereas the testing dataset was used to evaluate the performance of the model on unseen data.

From the data preparations, four different datasets were created. The first dataset contains only the contents from the news corpus and the second dataset contains the title and the missing author indicator on top of the contents of the news corpus. Then PCA transformation was applied to the two datasets to generate the third and forth datasets (*refer to Table 3*).

Table 3 - Multiple datasets created for machine learning model

S/No	Dataset Description
1	Content only (without PCA)
2	Content + Title + Missing Authors (without PCA)
3	Content only (with PCA)
4	Content + Title + Missing Authors (with PCA)

5. MACHINE LEARNING MODELS

The problem was defined to be a binary classification problem, hence, the following five supervised machine learning techniques were applied:

- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier
- Gradient Boosted Tree Classifier
- Naive Bayes Classifier

However, Naive Bayes Classifier does not allow negative values, hence the Naive Bayes will not be used for the datasets with PCA transformation

The results for each of the models are being tabulated based on their Accuracy score and ROC-AUC score that were chosen to measure the performance of the models (*refer to Tables 4 to 8*).

Table 4 - Results for Logistic Regression: best performance for logistic regression yielded an accuracy of 95.3% and ROC-AUC score of 98.8%.

Dataset No.	Accuracy Score	ROC-AUC Score
1	94.6%	98.5%
2	95.3%	98.8%
3	88.9%	94.9%
4	89.8%	95.3%

Table 5 - Results for Decision Tree Classifier: best performance for decision tree classifier yielded an accuracy of 93.0% and ROC-AUC score of 97.2%.

Dataset No.	Accuracy Score	ROC-AUC Score
1	93.0%	97.1%
2	93.0%	97.2%
3	83.1%	90.2%
4	82.8%	89.9%

Table 5 - Results for Random Forest Classifier: best performance for random forest classifier yielded an accuracy of 93.3% and ROC-AUC score of 98.5%.

Dataset No.	Accuracy Score	ROC-AUC Score
1	93.3%	98.5%
2	93.1%	98.5%
3	87.2%	94.6%
4	86.8%	94.3%

Table 7 - Results for Gradient Boosted Tree Classifier: best performance for gradient boosted tree classifier yielded an accuracy of 95.9% and ROC-AUC score of 99.4%. Both datasets without PCA transformation gave the same results. Dataset 1 is preferred over dataset 2 since it has less features included (a simpler model).

Dataset No.	Accuracy Score	ROC-AUC Score
1	95.9%	99.4%
2	95.9%	99.4%
3	90.4%	96.8%
4	90.5%	96.8%

Table 8 - Results for Naive Bayes Classifier: best performance for naive bayes classifier yielded an accuracy of 87.3% and ROC-AUC score of 91.5%.

Dataset No.	Accuracy Score	ROC-AUC Score
1	87.3%	91.5%
2	86.9%	90.9%
3	-	-
4	-	-

As mentioned previously, the Naive Bayes does not take in negative values, hence, there were no results for the two datasets that underwent PCA transformation in the Naive Bayes classifier (*refer to Table 7*).

6. INSIGHTS

Insights can then be drawn based on the results from the performance of the models.

Firstly, there were only slight differences in the results between the datasets using the content only and datasets that included the title and missing authors. This shows that the title and missing authors indicator did not have significant contribution towards the prediction of the labels for the news corpus.

Secondly, it can be observed that the results using the datasets without applying PCA transformation were significantly better than the models using the PCA features. Hence, we can infer that the original datasets without PCA transformation will perform better for the prediction.

To summarise the results (*refer to Table 9*), the gradient boosted tree classifier gave the best results with an accuracy score of 95.9% and ROC-AUC score of more than 99% with both Datasets 1 and 2 (*refer to Table 3*). However, the logistic regression model which is the next best model, is also an honorable mention with an accuracy score of 95.3% and AUC score of 98.8% with Dataset 2 (*refer to Table 3*).

Table 9 - Best performance results from each machine learning model.

ML Model	Dataset No.	Accuracy Score	ROC-AUC Score
Logistic Regression	2	95.3%	98.8%
Decision Tree	2	93.0%	97.2%
Random Forest	1	93.3%	98.5%
Gradient Boosted Tree	1	95.9%	99.4%
Naive Bayes	1	87.3%	91.5%

Despite the slight differences in the results of less than 1% between the logistic regression model against the gradient boosted tree classifier, the runtime for the logistic regression, which is around 4 minutes, was significantly faster than the gradient boosted tree classifier which took more than 3 hours to train and predict.

7. CONCLUSION

The gradient boosted tree classifier gave the best results with an accuracy score of around 96% and AUC score of 99%. Which means that out of 100 news, 96 will be classified correctly. And out of 100 news that is classified to be fake, we have a 99% confidence that the prediction is correct. While the runner up, logistic regression, is not far behind with an accuracy score of around 95%, has a significantly faster runtime as compared to that of the gradient boosted tree

However, there were some limitations to the machine learning models.

Unfortunately, we were not able to obtain enough local news corpus data that is applicable to the Singapore context. The data that we were able to get are mainly US-related data which was apparent in the data exploration. Some of the words that appeared many times were phrases such as “Donald Trump”, “New York” and “United States”. Also, most of the data were dated in late 2016 when the US presidential election happened. Hence, the trained models may not be effective for the more recent events or news.

In addition, the word semantics were not taken into consideration in the machine learning model. As a result, the interpretability of the models will be lacking in terms of the capability of using the model to generate or understand key words in a "fake" news corpus. Instead, we have to rely on the model to predict whether the news report is reliable or fake.

Nevertheless, despite the limitations, with the high accuracy that the models were able to achieve, our team believes that the models can be retrained with the relevant datasets, to give the classification predictions for the business applications.

REFERENCES

Ng H., 2018, *4 in 5 Singaporeans confident in spotting fake news but 90 per cent wrong when put to the test: Survey, Straits Times*: https://www.straitstimes.com/singapore/4-in-5-singaporeans-confident-in-spotting-fake-news-but-90-per-cent-wrong-when-put-to-the

Lorent S., 2018-2019, *Fake News Detection Using Machine Learning*: https://matheo.uliege.be/bitstream/2268.2/8416/1/s134450_fake_news_detection_using_machine_learning.pdf

Szpakowski M., 2018, *Fake News Corpus, Github Repository*: https://github.com/several27/FakeNewsCorpus

CODE SOURCE (GITHUB)

https://github.com/liweicheng178/fakenews_classification