



UNIVERSITY OF TORONTO  
DEPARTMENT OF MECHANICAL & INDUSTRIAL ENGINEERING

# **Fact or Fake News: Developing a Fact Checking Algorithm**

## **MIE1624H1: Introduction to Data Science & Analytics Consulting Report**

Prepared by: Group #5

Hijun Seo

Jeong Cheol Seok

Meji Lee

Mickole Mulano

Negar Balaghi

Submitted to:

Oleksandr Romanko

Department of Mechanical & Industrial Engineering

University of Toronto

Email: romanko@romanko.ca

November 28<sup>th</sup>, 2019

## EXECUTIVE SUMMARY

The prominence of fake news has caused great concerns on its negative impact towards individuals and society. As such, the development of automatic fake news detection systems is of critical importance to maintain an unbiased news ecosystem. We approached the problem from the fields of Natural Language Processing and Machine Learning, and developed a model that combines Term Frequency - Inverse Document Frequency feature extraction process and a Bernoulli Naïve Bayes classifier. With tuning simplicity and high testing efficiency, our model obtained a validation accuracy of 47.4% and testing accuracy of 46.9% as part of the 2019 Schulich Leaders Prize Competition. Our model also provides insights on the importance of taking claimants into account when classifying the veracity of a statement.

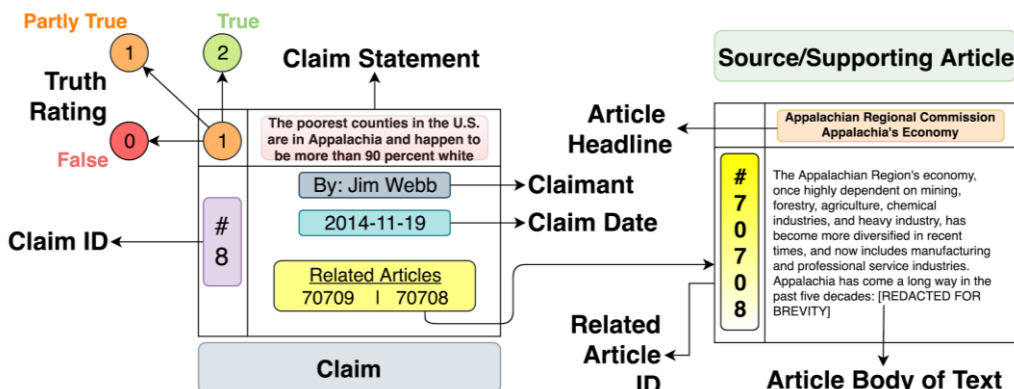
## 1. INTRODUCTION

At the peak of the 2016 United States presidential election, the 20 most widely discussed false election narratives posted on hoax news sites and hyperpartisan blogs had generated over 8.7 million shares, comments, and reactions on Facebook alone, compared to the 7.6 million engagements for the 20 frequently-discussed election stories from 19 prominent online news publishers [1]. Similarly, \$130 billion USD in stock value was momentarily eradicated following the spread of fake news claiming former US President Barack Obama was injured [1]. As such, the prominence of disinformation in online content has garnered significant attention to its detrimental effects towards the public and society for its ability to potentially influence and erode political, economic, and social realities [2-3].

While there have been human interventions to verify information veracity (such as the International Fact-Checking Network) at a high accuracy, manual identification of fake news is deemed impractical due to the substantial volume of digital news publications and its rapid diffusion worldwide, and the limitations of human operators to have a vast knowledge of the covered topics [2][4-5]. Recent advancements in Machine Learning (ML) and Natural Language Processing (NLP) have led to an interest in developing automatic, computational fake news detection systems [2]. NLP is an emerging subfield in artificial intelligence and linguistics that analyzes word patterns and could provide statistical correlations between news publications. With the sheer volume of digital content, ML classification algorithms are able to automatically learn and improve its own performance with more training data.

Traditional methods include a content-based approach where algorithms analyze the online news content from its headline, body of text, and any accessible metadata, and compare this to verified factual information [2][4]. While several endeavours have been made, they still lack the required robustness to reliably predict the veracity of digital content due to the complexity of the human language [2-5]. Systems must also remain politically and socially unbiased as fake and reliable news exists in balance at both ends of the spectrum [2][6].

In this report, we propose a fake news detection model that innovatively combines and adopts an NLP based-approach and ML classification techniques. Our model was developed in accordance to the 2019 Schulich Leaders Prize Competition guidelines. Several NLP based-approaches and ML classification techniques were studied and compared with a dataset composed of real and fake news online publications. We present these experimental evaluations which yielded promising results when compared to current industry solutions, and discuss the implications of our model against the current state of fake news distribution.



**Figure 1:** Sample representation of provided dataset consisting of real-life claim statements and associated metadata. Each claim consists of a statement, the author/claimant, a unique numeric identifier, and date of publication. Additionally, each statement has been professionally fact checked, and its veracity is numerically encoded and mapped to its truth rating category.

## 2. PROBLEM IDENTIFICATION

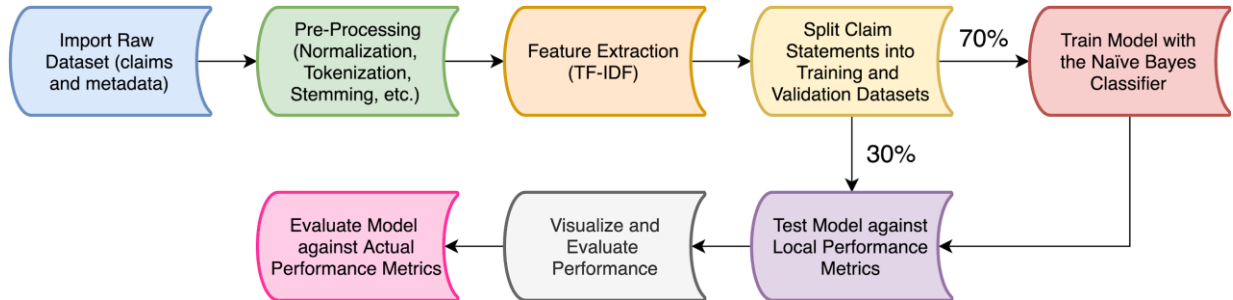
Developed by the Schulich Foundation and the University of Waterloo, Canada, the 2019 Schulich Leaders Prize Competition invites teams to employ artificial intelligence to solve the global issue of fake news and the spread of

disinformation. The objective of the detection model is to verify a set of claims and flag its veracity based on three categories: True, Partly True, and False.

The provided training dataset consists of 15555 claims (see Figure 1). Each claim is linked to at least two related articles (or metadata) which can be either a source or supporting article<sup>1</sup>. All claims and associated metadata were compiled across nine different fact-checking websites<sup>2</sup>. Using the provided dataset, our detection model developed in a Python 3.7 environment, complies with the following criteria:

- **Effectiveness:** The model must achieve a robust and high accuracy as measured by the weighted average of the precision and sensitivity (or probability of detection) of the predictions of truth ratings for each claim.
- **Efficiency:** The model must provide a robust output of the predictions of truth ratings for each claim within a specified time limit<sup>3</sup>.

### 3. DATA ANALYSIS



**Figure 2:** Generalized classification process used to create the ML model for fake news detection.

#### 3.1 DATA CLEANING

The raw format of each claim statement is unsuitable for analysis using an ML classification technique, which necessitates processing of the text into a machine-readable format. Preprocessing of text data includes converting all letters to lowercase and removing all numbers, punctuation, symbols, and stop words<sup>4</sup>. To reduce inflected words to their root form, stemming is applied to text<sup>5</sup>. To represent the text features numerically, important words in the dataset were extracted using the Term Frequency - Inverse Document Frequency (TF-IDF) method<sup>6</sup>.

#### 3.2 DATA VISUALIZATION

The current industry solutions for predicting the polarity (how positive or negative a statement is) or subjectivity (how objective a statement is) of a claim can be implemented using publicly available Python libraries such as TextBlob. However, the application of such methods on the claims data results in poor correlation with the claim labels (see Figures 3 and 4). Furthermore, the most frequently used words within each of the three truth rating categories are largely similar (see Figure 5). This overlap suggests that use of only word frequencies to train an ML classifier will result in poor classification of the claims. In contrast, correlation of the claimants data and the claim truth labels displays clear differences between the three types of claims, to the point of achieving near deterministic categorization for some of the claimants (see Figure 6). For example, all the claims made by the claimant 'Various websites' are false. Hence, the claimants data seem to provide more informative features for training of a predictive ML classifier.

<sup>1</sup> Source articles fundamentally contains the claim, which may be phrased differently, whereas the supporting article provides evidence to support the claim statement.

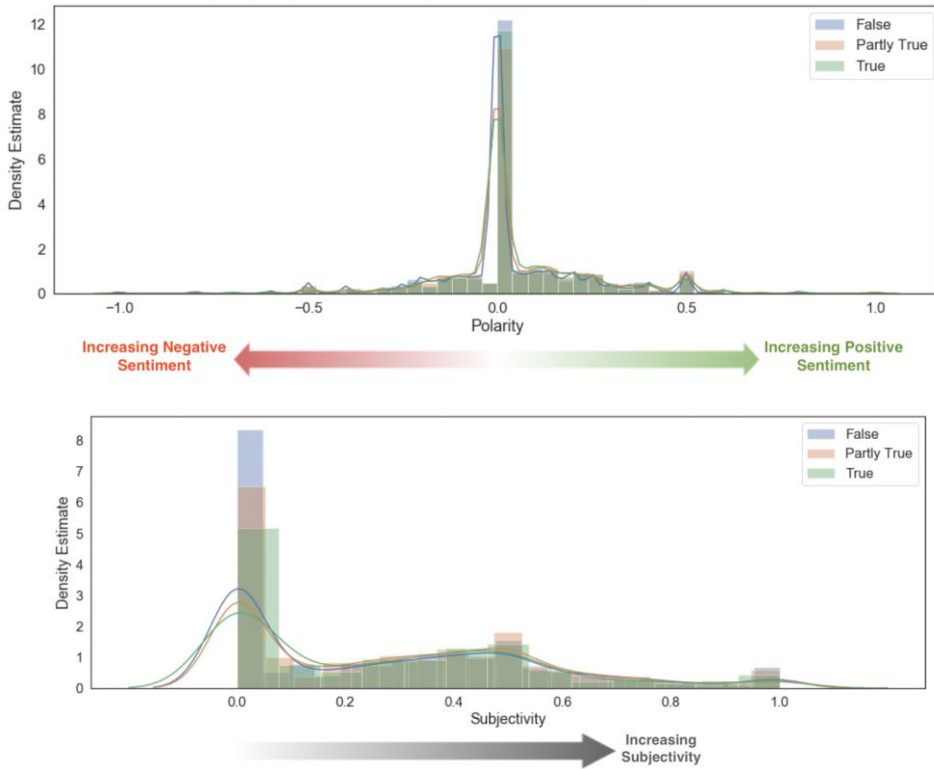
<sup>2</sup> Articles were compiled from: politifact.com, snopes.com, washingtonpost.com, weekllystandard.com, africacheck.org, factscan.ca, factcheck.afp.com, polygraph.info, factcheck.org

<sup>3</sup> For compliance with the 2019 Schulich Leaders Prize Competition, the maximum run time was set to 30 minutes.

<sup>4</sup> Stop words are the most commonly used words in a language that often do not carry important information, including most pronouns (e.g. "my", "you", "this", "where") and indicative verbs (e.g. "is", "are", "were").

<sup>5</sup> Stemming often attempts to find the root word by eliminating the characters at the end of a word, including suffix characters (e.g. "-ing", "-es", ...).

<sup>6</sup> TF-IDF calculates the frequency of specific words in each claim, and assigns a relative importance weight to them based on the frequency of that word's appearance in the entire document.

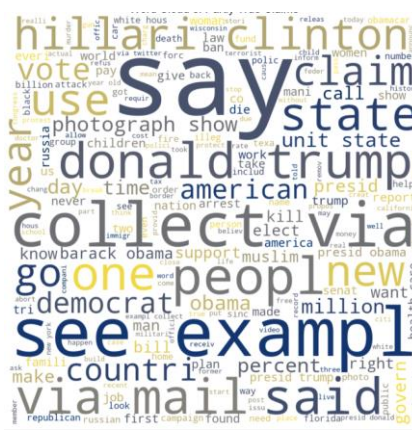


**Figure 3:** Distribution of true, partially true, or fake claims per polarity sentiment rating indicating the similarities in the general sentiment associated with the claims from each category.

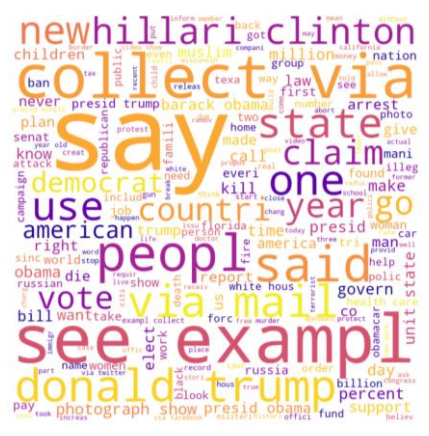
**Figure 4:** Distribution of truth ratings per subjectivity rating indicating similar distribution of subjective expression of feeling, views, or beliefs present in claims from each category.



(a) True Claims

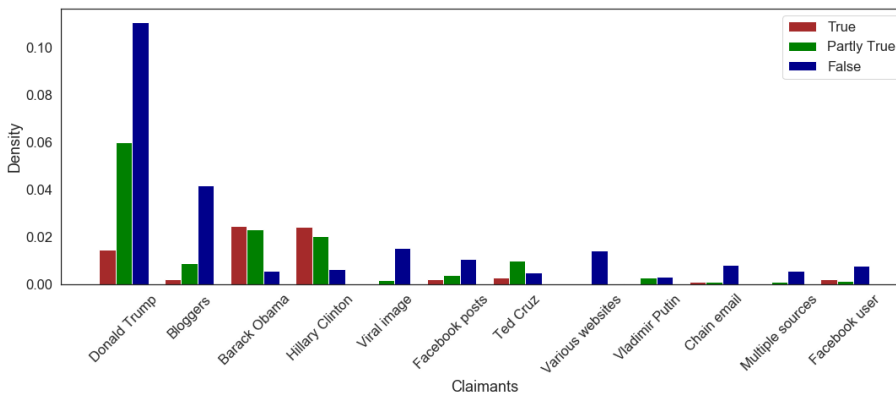


(b) Partly True Claims



(c) False Claims

**Figure 5:** Visual representation of frequently used words in each of the three categories.



**Figure 6.** Distribution of claims truth rating for commonly referenced claimants.

### 3.3 ALGORITHM APPLICATION

Based on our exploratory data analysis, the addition of the claimants data to the word frequency data improves the accuracy of classification. Therefore, these two features were used to build the classification model. The cleaned and vectorized data was split into training and validation sets (see Figure 2), and tested with eight different ML classifiers. The performance of each algorithm was evaluated by calculating the F1-score<sup>7</sup> (see Table 1). Bernoulli Naïve Bayes classifier was selected as the final model, as it yielded the best performance amongst those tested.

Naïve Bayes is a commonly used classifier in NLP applications due to its performance and simplicity. Naïve Bayes classifier learns the pattern of the text data based on the occurrence of each term and compares to the contents in all categories. Afterwards, it calculates the probability of each label for a given text, and outputs the label with the highest probability.

**Table 1:** Comparison of model performance against local performance metrics

ML Classifier	Training F1-Score	Validation F1-Score
Logistic Regression	60.66%	45.82%
Linear Discriminant Analysis	61.89%	46.63%
Multinomial Naïve Bayes	55.03%	46.06%
Bernoulli Naïve Bayes	54.29%	47.36%
Support Vector Machine (linear kernel)	59.42%	45.35%
XGBoost	44.34%	43.98%
Random Forest	97.17%	43.96%
Decision Trees	99.88%	41.99%

### 3.4 ALGORITHM EVALUATION

Our best model was submitted to the 2019 Schulich Leaders Prize Competition and a test F1-score of 46.9% was received, placing our team 22nd amongst 74 contestants<sup>8</sup>. In comparison to the other model accuracies listed in the competition guidelines, our model performed better than their Naïve Bayes classifier, but worse than the other complex models. However, our model has the advantage of being simple to tune and efficient in testing. This allows potential future re-training of the model, especially with a much larger dataset, to be possible at an efficient rate.

**Table 2:** Comparison of model performance against competition guidelines

ML Classifier	Test F1-Score
Our Detection Model with Bernoulli Naïve Bayes	46.9%
Naïve Bayes	38.2%
Bi-Directional Long Short-Term Memory	49.4%
Bi-Directional Encoder Representations from Transformers	51.1%

## 4. DISCUSSION

Early detection of fake news is of critical importance to redesign the information ecosystem of the digital age and develop “a news ecosystem and culture that values and promotes truth” [7]. Inaccurate information, especially information that is spread with malicious intent, can have detrimental effects on organizations from defamation of individuals or small entities to misdirection of mass public opinion on election campaigns that can affect the selected leader of a country.

Unfortunately, once a claim has been published on the internet, there is usually no way for an organization to force its author to revoke it. Larger social media organizations, such as Facebook, Twitter, and Instagram, are actively working on methods to combat misinformation. Most current methods focus on identification of false user accounts, occasional deletion of false statements, and insertion of true information to neutralize false claims [8]. However, the danger with these interventions lies in its threat to freedom of speech and free press. A recent report on global press freedoms, showed that media freedom was at its lowest point in 13 years [9]. An alternative solution that is not limited to social media platforms is pre-bunking, where true information is preemptively injected online before false information is

<sup>7</sup> The F1-score is a measure of the model’s accuracy, which takes into account the model’s precision and recall. Precision is a measure of model’s exactness in making predictions; a low precision value indicates frequent false positive predictions. Recall is a measure of the model’s sensitivity; a low recall value indicates frequent false negative predictions.

<sup>8</sup> Ranking obtained as of November 26<sup>th</sup>, 2019, 10:00 pm EST



even created, which saves the effort of fact-checking and debunking it before a false claim has a chance to affect the masses [10]. However, the question remains of what information and how it should be communicated must be decided without the initial basis of a false claim.

The analysis presented in this report identifies important features for detection and classification of misinformation. During the exploratory analysis phase, we highlighted the importance of taking claimants into account when classifying the truthfulness of a statement. Our presented classifier uses this feature to make significant improvements to a basic Naïve Bayes classifier. Using TF-IDF, an optimized set of features was selected such that the most informative aspects of the dataset are used for classification. However, it must be recognized that some of the decisions that improve model efficiency may result in reduced model accuracy.

As part of the data clean-up, a stemming function was utilized to reduce word inflection. While stemming is effective in reducing diversity of words in a specific document, it does not take the meaning of the words into account, which may have eliminated important features from the dataset. An alternative approach for inflection reduction is lemmatization, which accounts for the meaning and implication of each word during clean-up.

The decision to forgo utilizing the related articles textfiles is largely related to the reduction of classifier efficiency, which resulted in considerable increased runtime but insignificant increases in classifier accuracy. It's important to note that related articles are not crucial features only within the scope of this specific project, as the training dataset contains claims that are similar to the claims in the final performance evaluation. The use of related articles would be critical in the cases where the training and validation datasets belong to different topics. For example, a classifier that was trained on claims related to politics would be expected to perform well on predicting truthfulness in other datasets that contain political claims. However, the same classifier would not be expected to perform as well on dataset containing medical claims. The use of related articles and engineering of features that determine claim validity based on the correlation between claim statements and the related documents would be robust enough to handle claims from any general topic.

## 5. CONCLUSION

The first step to correcting misinformation is to identify and assess the degree of inaccuracy. The model presented in this report provides an approach to classify information presented in a claim based on its veracity. Using a data set of claims already categorized by professional fact checkers, a machine learning algorithm was trained to identify features associated with true, partly true, or false claims, and use these features to predict veracity of future claims presented to it. The model was optimized for effective, yet efficient classification of claims. However, there is still room for improvement.

Future work can be done to improve model accuracy. The current dataset used to train our algorithm on had uneven class distribution, with much fewer true claims than false and partly true claims. Training on a more even dataset could result in better accuracy since the current algorithm's error matrix (see Supplementary Figure A4) suggests most of the error comes from incorrect prediction of true claims.

Furthermore, we envision our solution to be refined for ease of use through the development of a user interface (UI). One possible application is to determine defamation. The UI could be designed to collect relevant statements about the company through specified keywords from popular social media platforms such as Facebook, Twitter, or Reddit. With this data, fake news can be monitored, especially those with negative sentiment. Thus, the model will allow the company to preemptively prove the rumour as false or divert the attention with positive news.

With the current boom in technology and use of social media, information, especially fake news, spreads quickly [11]. What might start out as one defaming tweet can lead to a company's downfall [12]. Our solution can prevent such consequences of defamation through assessment and monitoring of fake news to inform companies when to take action.

## BIBLIOGRAPHY

- [1] X. Zhou, R. Zafarani, K. Shu and H. Liu, "Fake News", Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining - WSDM '19, 2019. Available: 10.1145/3289600.3291382 [Accessed 27 November 2019].
- [2] M. Della Vedova, E. Tacchini, S. Moret, G. Ballarin, M. DiPierro and L. de Alfaro, "Automatic Online Fake News Detection Combining Content and Social Signals", 2018 22nd Conference of Open Innovations Association (FRUCT), 2018. Available: 10.23919/fruct.2018.8468301 [Accessed 27 November 2019].
- [3] S. Gilda, "Evaluating machine learning algorithms for fake news detection", 2017 IEEE 15th Student Conference on Research and Development (SCOREd), 2017. Available: 10.1109/scored.2017.8305411 [Accessed 27 November 2019].
- [4] Á. Figueira and L. Oliveira, "The current state of fake news: challenges and opportunities", Procedia Computer Science, vol. 121, pp. 817-825, 2017. Available: 10.1016/j.procs.2017.11.106 [Accessed 27 November 2019].
- [5] H. Ahmed, I. Traore and S. Saad, "Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques", Lecture Notes in Computer Science, pp. 127-138, 2017. Available: 10.1007/978-3-319-69155-8\_9 [Accessed 27 November 2019].
- [6] K. Shu, S. Wang and H. Liu, "Beyond News Contents", Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining - WSDM '19, 2019. Available: 10.1145/3289600.3290994 [Accessed 27 November 2019].
- [7] D. Lazer et al., "The science of fake news", Science, vol. 359, no. 6380, pp. 1094-1096, 2018. Available: 10.1126/science.aao2998 [Accessed 27 November 2019].
- [8] West, D. M. (2017). *How to combat fake news and disinformation*. Washington: Brookings. Available: <https://www.brookings.edu/research/how-to-combat-fake-news-and-disinformation/> [Accessed 27 November 2019]
- [9] Freedom Press. (2017). *New Report: Freedom of the Press 2017 - Press Freedom's Dark Horizon*. Washington. Available: <https://freedomhouse.org/article/new-report-freedom-press-2017-press-freedom-s-dark-horizon> [Accessed 27 November 2019]
- [10] Cook J, Lewandowsky S, Ecker UKH "Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence." PLOS ONE, 2017 Available: 12(5): e0175799 [Accessed 27 November 2019]
- [11] Langin, K. (2018). *Fake news spreads faster than true news on Twitter—thanks to people, not bots*. Washington: AAAS. Available: 10.1126/science.aat5350 [Accessed 27 November 2019]
- [12] Atkinson, C. (2019). Fake news can cause 'irreversible damage' to companies — and sink their stock price. New York: NBC News. Available: <https://www.nbcnews.com/business/business-news/fake-news-can-cause-irreversible-damage-companies-sink-their-stock-n995436> [Accessed 27 November 2019]

## APPENDICES

## MIE1624 Group 5


View team profile


**Members**

- Mickole Mulano (team lead)
- Hijun Seo
- Jessie Lee
- JeongCheol Seok
- Negar Balaghi

**Requests**

Username	Name
----------	------

**Tagline** 

**Description** 


**Tags** 

Figure A1: Proof of attendance at the 2019 Schulich Leaders Prize Competition

**Previous submissions**  
Total number of submissions : 11

#	Date	Status	Score
1	11/19/19, 20:54	Calculation in progress	0
2	11/20/19, 23:05	Calculation in progress	0
3	11/21/19, 10:08	Calculation in progress	0
4	11/21/19, 11:37	Calculation in progress	0
5	11/21/19, 17:20	Calculation in progress	0
6	11/21/19, 18:05	Calculation in progress	0
7	11/21/19, 19:11	Final	0.302275
8	11/22/19, 19:28	Final	0.01631
9	11/25/19, 15:01	Final	0.014359
10	11/25/19, 17:33	Final	0.410642
11	11/25/19, 20:55	Final	0.469259

Figure A2: Proof of submission to the 2019 Schulich Leaders Prize Competition





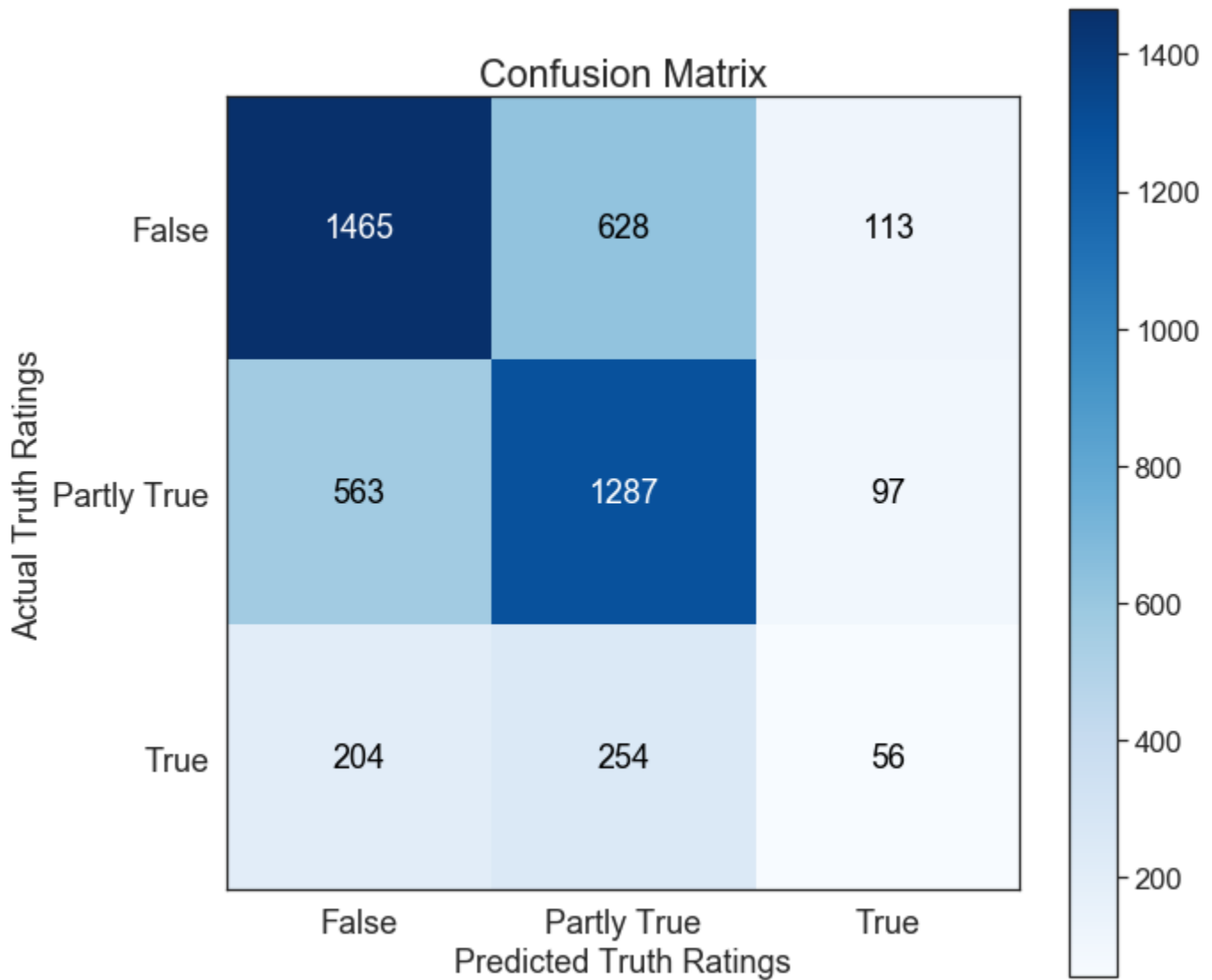
<b>DataCup</b>  Mickole Mulano  Dashboard Competitions  Log out EN 	14	McGill_NLP	0.508509
	15	glia-team	0.502138
	16	FoF	0.499765
	17	hello-world	0.495999
	18	ditto	0.495908
	19	lululu	0.494268
	20	Verse	0.490922
	21	Master of Science	0.488436
	22	MIE1624 Group 5	0.469259
	23	MLAIR	0.461804
	24	Akina	0.459159
Challenge Prizes Timeline Rules Data Resources Evaluation <b>Leaderboard</b> Teams Submission My Team	25	MIE1624 Group7	0.453867
	26	Veritas	0.43149

Figure A3: Current ranking at the 2019 Schulich Leaders Prize Competition as of November 26<sup>th</sup> 2019, 10:00 pm EST





**Figure A4:** The error/confusion matrix shows how the predicted label compares to the true label. The figure suggests that the model is not best at predicting 'true' claims since there were more claims that were predicted incorrectly as 'false' or 'partly true' than correctly as 'true'. This may be because of the uneven class distribution in the data.