Detecting Fake News Using Machine Learning

Kelsey Corro
New Mexico State University
Las Cruces, NM
kcorro@nmsu.edu

Shannon Cruse
New Mexico State University
Las Cruces, NM
shanhead@nmsu.edu

Vinny Mikelic
New Mexico State University
Las Cruces, NM
vmikelic@nmsu.edu

MOTIVATION

With the advent of the internet, the barrier of entry to create and spread false or misleading information to a large audience has been reduced significantly [1]. Social media platforms allow for memetic dissemination of information that is stripped of context or is outright fabricated. The lack of regulation and oversight found on these sites can hasten the spread [2] and allow for personal information environments which foster a biased perspective [3].

Such environments are also prone to creating extremist ideas and increasing political polarization [3]. This can lead to widely successful misinformation campaigns that cause substantive and long-lasting negative outcomes for society. Some relatively recent examples are the "Stop the steal" campaign that contributed to the January 6th Capitol attack, and COVID-19 misinformation which resulted in higher numbers of hospitalization and death due to vaccine hesitancy [4].

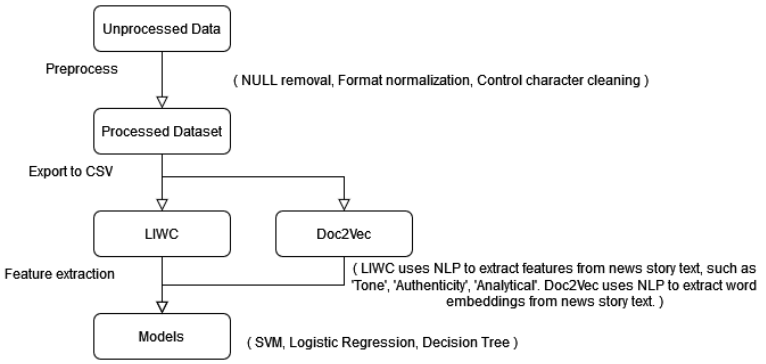
PROBLEM

Although many platforms that contain misinformation are taking action to combat it, there is little incentive to invest into manual moderation of these issues. Misinformation is more complex and harder to validate than other moderation tasks which may be tackled by simple word or phrase searches. However, the largely homogenous and excessive rhetoric used by misinformation campaigns lends itself to being detected automatically using a machine learning approach [5]. Machine learning and natural language processing has been used extensively to categorize a sentences' lexical and syntactic features as bigoted [6] and can naturally extend into the territory of misinformation using different features of speech. This will allow platforms to create better automatic moderation tools to reduce the harm of misinformation. Harm reduction can be accomplished through provenance enhancing, corrective commenting, trustworthy information sourcing, and media literacy resources [7].

Extensive work has been conducted on the topic of fake news, including major social media sites developing their own tools to combat the spread of misinformation. However, with the prevalence of fake news on social media and the real-world harm it can cause in mind, identifying misinformation is crucial. Our intent during this project is to use extensive fake news datasets to train machine learning models to identify misleading content.

SOLUTION

To achieve our goals, fake news datasets from Kaggle (a website that is "home to the world's largest community of data scientists") will be collected and analyzed. Once the data has been collected, it will be processed using natural language processing to extract relevant information from the text data and converted into data that can be classified. Afterward, several machine learning tasks will be implemented such as decision tree classification, logistic regression, naive bayes, and support vector machines to classify news into one of two categories: fake news and real news. There are many aspects to creating a robust model for detecting misinformation. This starts with accurate data for training and testing. Because our model is being trained using human-interpreted facts, we must ensure that our data is accurately tagged as real or fake. This can be done generically through patterns of speech which are recognized to be part of true and false utterances [8]. Implementing generic methods of detecting falsehoods in text, such as utilizing LIWC to analyze psycho-linguistic characteristics, will allow us to better leverage the datasets we have chosen to train our model.



Pipeline of our model construction.

DATASET

The dataset was collected from Kaggle and is titled "Misinformation & Fake News text dataset 79k." It contains three .csv files, but only two will be used for this project. All information except the text data was stripped from each instance. The instances labeled as 'true' came from "a variety of sources, such as *Reuters, The New York Times, The Washington Post* and more" [9]. While the instances labeled 'fake' came from sources such as "*Redflag Newsdesk, Beitbart, Truth Broadcast Network* which are American right wing extremist websites" [9]. Since the only data available for each instance is the original text data, word embeddings and natural language processing will be used to extract more features from the text data.

The dataset has a sizable number of instances in which we can produce features from the text data and evaluate the association of those features with their given classification: fake or true.

This data can also be used for clustering and topic modeling so that newly written articles can be fed into our model and categorized as fake or true.

STATISTICS / ANALYSIS

The dataset titled "Misinformation & Fake News text dataset 79k" contains 43,642 instances that are classified as 'fake' and 34,975 instances that are classified as 'true', which gives a total of 78,617 instances for the overall dataset.

The first step into preprocessing the data was ensuring the dataset was clean and contained only useful information for LIWC and our model to be trained on. Our data seemingly had little preprocessing done to it, this was assumed because removing null values from our dataset took away 29 of our instances. This should be a simple first step that was not taken by the authors of the data. It's possible our data was programmatically taken from news sites and compiled into one set.

After clearing the data of NULL entries, the entries themselves had to be cleaned and put into a format that is more convenient for us, LIWC, and our model to work with. Initially, our dataset was separated into two files - one containing true stories, one containing false stories. This is obviously a problem for training a model because the data should have a uniform distribution of labels. To solve this issue, the files were concatenated into a single data frame and randomly shuffled.

After these two steps, a final cleaning process was applied to the actual news story text value of each instance. For every news story in the data frame, an algorithm to remove all control characters from each string was applied. This was a relatively fast operation that will allow future natural language procedures to be done without formatting hiccups.

A	B	C
112205	Melania Trump plagiarized part of her Republican National Convention speech. So, Republicans, whose unofficial ther	1
112206	A New York band member called out Austinâ€™s South by Southwest (SXSW) music festival over its contract that den	0
112207	By wmw_admin on November 3, 2016 The Saker â€" The Saker.is Nov 3, 2016 Last May I wrote an article entitled Count	1
112208	Edmondo Burr in Middle East , News , World // 0 Comments For the first time ever China and U.S. ally Saudi Arabia	1
112209	Free Thought Project â€" by Jay Syrmopoulos	1
112210	Leave a reply Charles Hugh Smith â€" The sole output of Americaâ€™s Establishment/Ruling Elite is self-serving hubris	1
112211	WASHINGTON (Reuters) - New U.S. Commerce Secretary Wilbur Ross said President Donald Trump did not endorse a p	0
112212	AL RAWDAH, Egypt (Reuters) - People wounded in an attack on a mosque in Egypt s North Sinai region that killed more	0
112213	WASHINGTON (Reuters) - A battle over implementation of the Iran nuclear deal erupted on the U.S. Senate floor on W	0
112214	BERLIN (Reuters) - U.S. President Donald Trumpâ€™s move to revise a travel ban on citizens of certain Muslim-majorit	0
112215	The Muslim couple who stormed an office holiday party Wednesday in Southern California, mowing down 14 people	0
112216	BEIJING (Reuters) - China has set a clear direction on reform and opening up to the world and will not deviate from th	0
112217	The words Extraordinary Claims needs to be banished when talking Extraterrestrials page: 1 link I was reading an articl	1
112218	A struggling mom sent an email to all 122 of Mississippi s lawmakers asking for help getting insulin and an insulin pum	1
112219	WASHINGTON (Reuters) - The U.S. Senate on Tuesday overwhelmingly backed the expansion of NATO to allow Monte	0
112220	ERBIL/SULAIMANIYA, Iraq (Reuters) - Kurds voted in large numbers in an independence referendum in northern Iraq o	0
112221	BRUSSELS (Reuters) - NATO urged all countries to step up efforts to enforce sanctions on North Korea and stop its wea	0
112222	Despite decades of fervent student protests that reached a peak last fall, the president of Yale announced on Wednes	0
112223	Edmondo Burr in Sci/Environment // 0 Comments Yesterday the sun erupted with a huge solar flare sending streams	1
112224	It appears that the transcript of the NFLâ€™s Opening Night media event for the upcoming Super Bowl omitted most c	0
112225	HANOVER, Germany (Reuters) - Members of the anti-immigrant Alternative for Germany (AfD) party elected a right-w	0
112226	Jill Stein agreed to do an interview with The View hacks to discuss her recount effort. The interview becomes very aw	1
112227	Hate Speech as a Weapon: Reporters Are Charged for Covering Disturbances 26, 2016	1
112228	GOP establishment consultant Rick Wilson, one of the most vile â€œNever Trumpersâ€ who once said the donor class	0
112229	Julian Zelizer is a professor of history and public affairs at Princeton University and a New America fellow. He is the	0
112230	Will the new 3-story Islamic Museum include pictures of Americans jumping out of their flame engulfed offices in the	1
112231	Sberbank plans to introduce Samsung pay technology in near future October 28, 2016 TASS banks , apps Sberbank is	1
112232	21st Century Wire says In the wake of the Friday release of the ODNI s 25 page report from US intelligence agencies (	1
112233	Globalization and technology are routinely cited as drivers of inequality over the last four decades. While the	1
112234	We reported earlier about Dallas cop killer, Micah X. Johnson and his affiliation with the New Black Panther group, as	1
112235	Rep. Barbara Lee ( ) said on Tuesday that she does not agree with statesâ€™ rights because President Donald Trump b	0
112236	BELIZE CITY, Belize â€" One oâ€™clock arrived. Relatives gathered at a hotel bar to watch Olympic gymnastics on tele	0
112237	So Trump opposes "free trade" and Hillary is all for it. This is not denied, yet T/O spins a BS story to make Trump the vi	1
112238	When Fox News host Stuart Varney bragged about Donald Trump wanting to imprison and strip citizenship from Amer	1
112239	BRUSSELS (Reuters) - The European Union agreed on Friday to move Brexit talks onto trade and a transition pact but sc	0

Usable data, composing approximately 95% of instances.

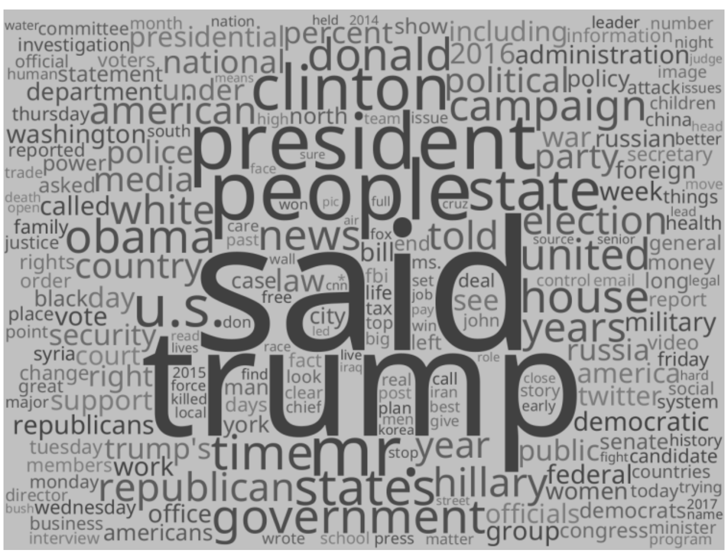
This is the result of our dataset after preprocessing. Each instance contains a (Text, Label) feature pairing. The "Text" feature corresponds to the news article text,

and the "Label" feature corresponds to whether the story is real (value of 1) or fake (value of 0). The strange characters that are seen within the strings are encoding errors from excel and do not appear in UTF view modes.

Unusable data (non-alphanumeric text)

This is an example of non-alphanumeric text that will be cleaned by LIWC, as it will only accept characters from its dictionary language (English).

LIWC



Word cloud of our dataset, constructed by LIWC

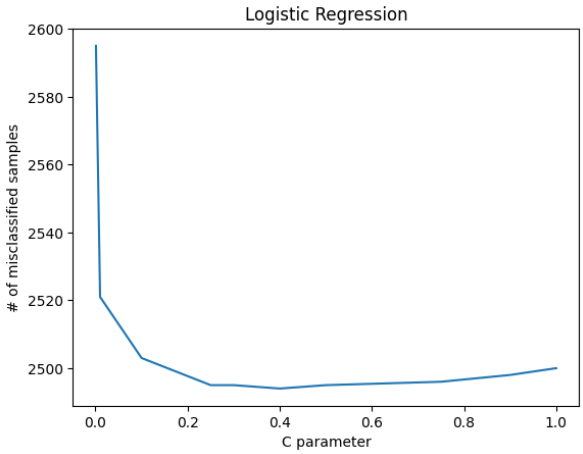
Linguistic Inquiry and Word Count (LIWC) is an easy-to-use text processing tool that can effectively measure various language characteristics within text to reveal intentions and identify patterns with minimal bias. For our project, we will be utilizing LIWC-22, which generates over 100 language features from a given text, all of which have been validated by over 20,000 scientific articles that utilized LIWC.

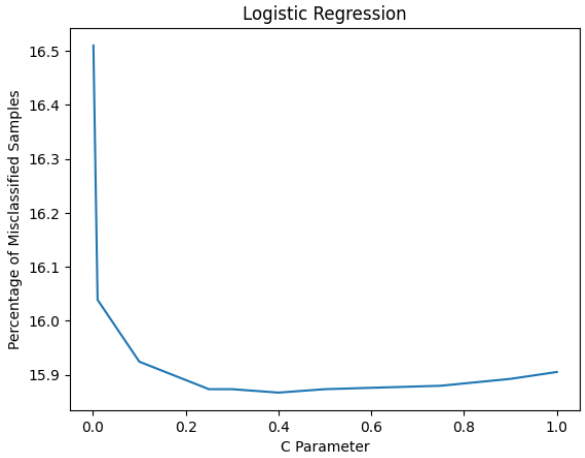
In our approach to classifying fake news, LIWC plays a significant role. We believe that relying on language characteristics, rather than just textual pattern recognition, results in a more robust language model. By training exclusively on textual data, a model may simply correlate specific keywords with certain labels, without considering the context in which those words are used. For example, if a model only utilizes the text in its most basic form, it may mistakenly classify any story containing the word "Trump" as fake news. This correlation has no basis in detecting fake news and may lead to the model miscategorizing stories related to Trump.

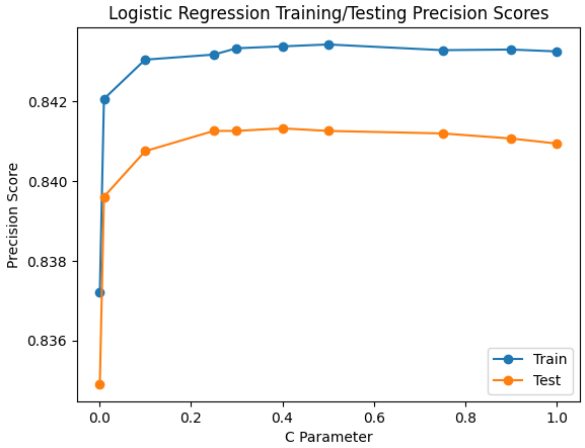
Using LIWC enables us to leverage widely recognized patterns of speaking and writing that correlate with disingenuous or truthful dialogue. Thus, our model is less biased and rooted in psychology. The result is a more accurate and reliable classification of fake news that considers the nuances of language and its underlying psychological cues.

MODEL RESULTS & EVALUATION
(from Google Colab,
'Fake News Classification.ipynb')

Logistic Regression Classifier







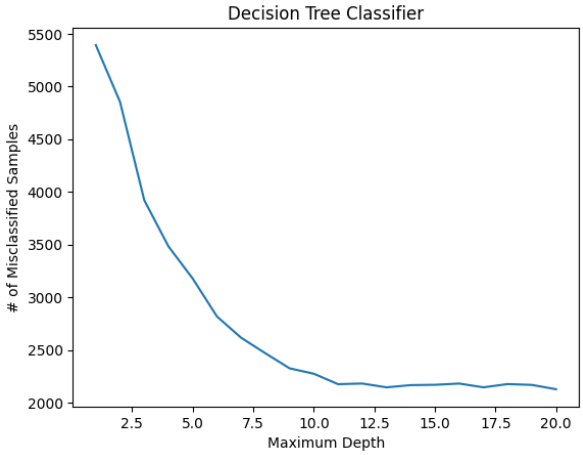
Training time: 4.269660711288452
Testing time: 0.0037016868591308594
# of samples tested: 15718
Misclassified samples: 2494
Percentage of Misclassified samples: 15.867158671586715
Training Precision Score: 0.8433886335077702
Testing Precision Score: 0.8413284132841329

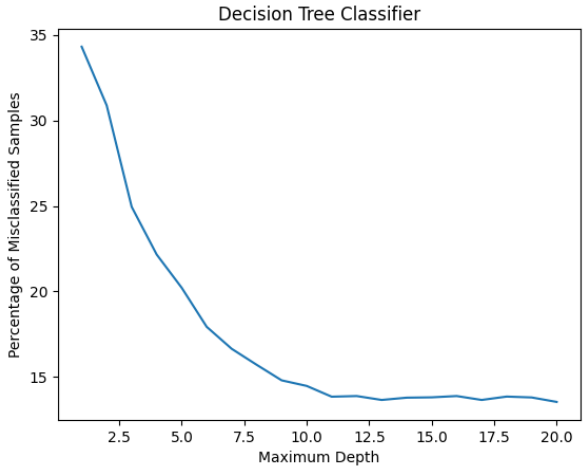
When tuning the hyperparameters for logistic regression, using C = 0.4 gave the best results as it produced the least number of misclassified samples and had the lowest percentage of misclassified samples. The precision scores for the training data might look significantly higher than the testing data when plotted, but the difference is only 0.002 which shows that the model is not overfitting (the y-axis is just very skewed).

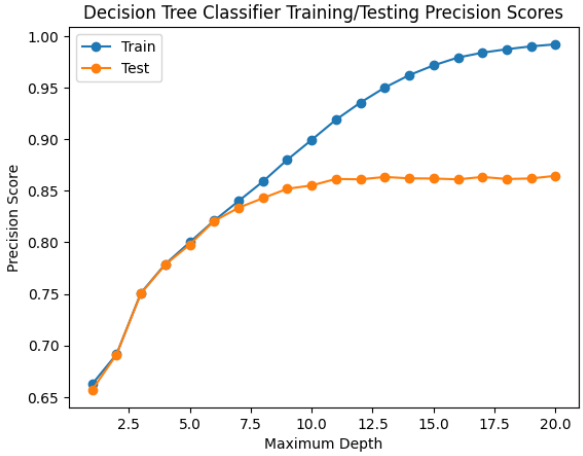
Parameter Values for Logistic Regression at optimized conditions:

- random state = 5
- C = 0.4
- max_iteration = 1000

Decision Tree Classifier







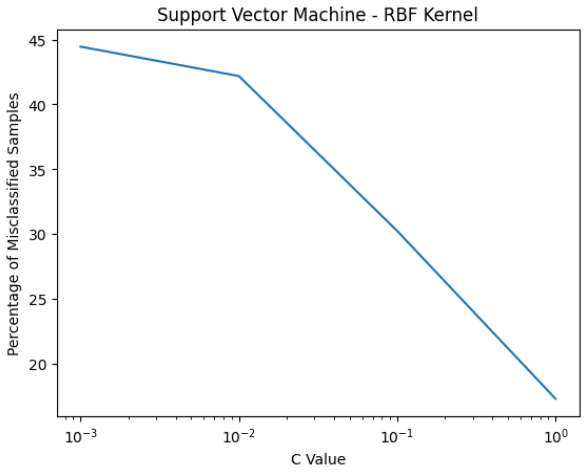
When tuning the hyperparameters for the decision tree classifier, using max_depth = 8 gave the best results (without showing evidence of overfitting) as it produced the least number of misclassified samples and had the lowest percentage of misclassified samples while not showing a large gap between the training data precision score and the testing data precision score. When increasing the max_depth to a value larger than 8, the training precision score would become a lot higher than the testing precision score, which showed that the model was overfitting at those values.

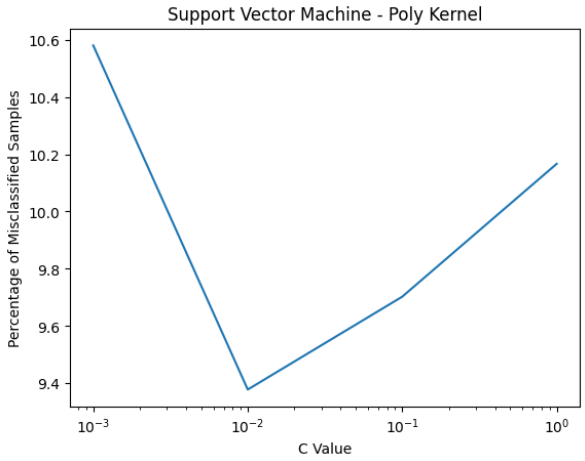
Parameter Values for Decision Tree Classifier at optimized conditions:

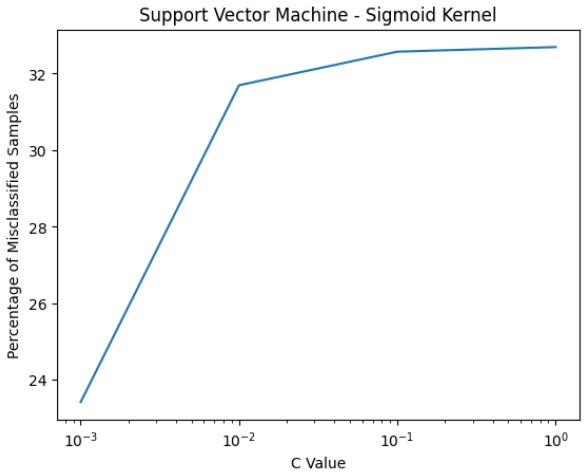
- random_state = 5
- max_depth = 8
- min_samples_split = 3

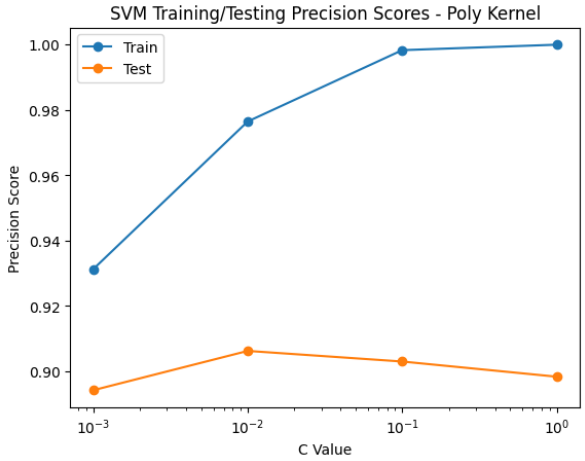
Training time: 3.802910804748535
Testing time: 0.006146430969238281
# of samples tested: 15718
Misclassified samples: 2470
Percentage of Misclassified samples: 15.714467489502482
Training Precision Score: 0.8589606960505177
Testing Precision Score: 0.8428553251049752

SVM Classifier









On the whole, the support vector machine produced quite accurate results, but was susceptible to overfitting issues. The poly kernel produced the best results overall, with the least amount of overfitting occurring at smaller values of C.

Parameter Values for Support Vector Machine at optimized conditions:

- random_state = 1
- kernel = poly
- C = 0.01
- gamma = 0.1

Training time: 1494.4344611167908
Testing time: 37.42927527427673
# of samples tested: 15718
Misclassified samples: 1474
Percentage of Misclassified samples: 9.377783433006744
Training Precision Score: 0.9764430800553532
Testing Precision Score: 0.9062221656699325

Naive Bayes Classifier
Training time: 0.16431736946105957
Testing time: 0.02162313461303711
# of samples tested: 15718
Misclassified samples: 6047
Percentage of Misclassified samples: 38.47181575264029
Training Precision Score: 0.6178561771302231
Testing Precision Score: 0.6152818424735972

Stratified 10-Fold Cross Validation was performed on the different models using their optimized parameters found from earlier analysis and the average scores were obtained. These were the results for each model (SVM was excluded due to runtime restrictions):

Model	**Average Score**
Decision Tree Classifier	0.8376147549491421
Logistic Regression	0.8407800874502234
Naive Bayes	0.6180312454817275

Based on these results, logistic regression gave the best results in terms of precision score and runtime efficiency.

FURTHER IMPROVEMENTS

After determining that the Logistic Regression classifier outperformed other classifiers, we sought to improve its performance beyond parameter tuning. Two primary methods for extracting more features from our story text were identified. The first method involved utilizing user defined LIWC dictionaries to supplement the default dictionary. The second method to improve our model is adding the natural language processing technique of word embeddings.

To start, we will discuss the first method, user defined dictionaries. The standard LIWC dictionary extracts over 100 useful features, but there exist additional dictionaries created by users that can extract even more. To test this approach, we experimented with several user dictionaries to determine if the generated features had any significant correlation with our data. Two dictionaries, in particular, displayed a substantial correlation with our dataset labels: the “controversial-terms” dictionary and the “absolutist” dictionary.

To calculate the correlation between these generated features and our class label, we used the “Point-biserial correlation coefficient”. This coefficient measures the relationship between a continuous variable (generated by the dictionary) and a dichotomous variable (our class label). The “controversial-terms” dictionary generates three values from our news stories: high-controversy, medium-controversy, and low-controversy. The “absolutist” dictionary, on the other hand, generates only one value from our news stories: absolutist. A coefficient of -1 indicates perfect negative correlation, 0 indicates no correlation, and 1 indicates perfect positive correlation.

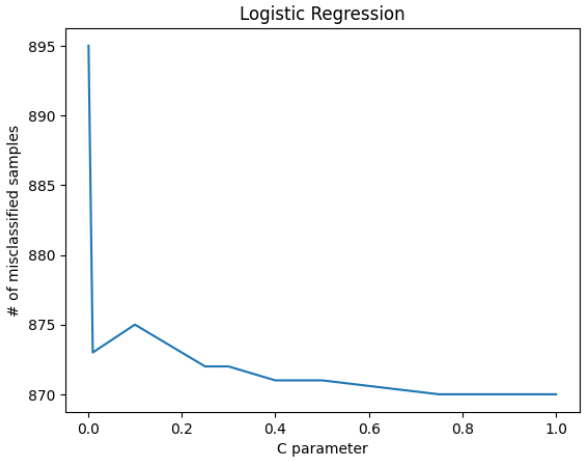
```
Point-biserial correlation value for high-controversy and real/fake news: 0.16876085574548944
Point-biserial correlation value for medium-controversy and real/fake news: 0.02367362818494699
Point-biserial correlation value for low-controversy and real/fake news: 0.12138143893061193

Point-biserial correlation value for absolutism and real/fake news: -0.1898928905557622
```

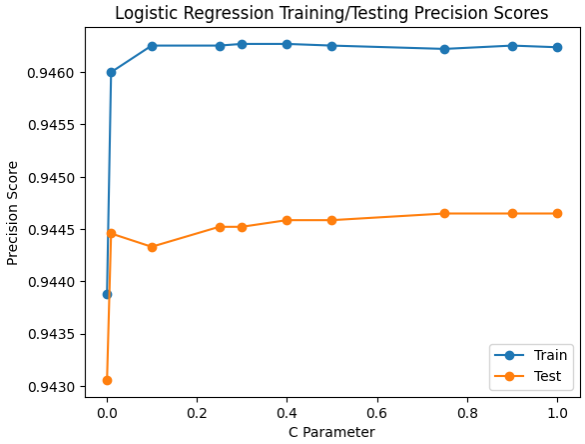
Here are the results (from correlation.py):

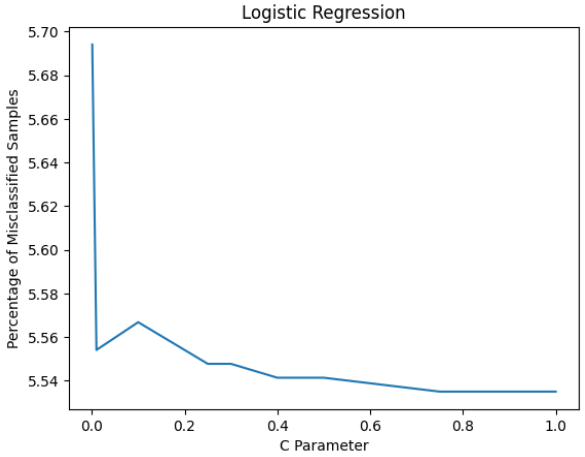
After running the correlation calculations, we found significant correlation between these generated features and the class label. Thus, it is recommended to include these dictionaries in our feature generation process to improve our model's performance.

Secondly, creating word embeddings via Doc2Vec allows us to extract vector features from our text corpus. These vectors will signify the relation of words within the story on a less sophisticated level when compared to LIWC. However, even simple distributed representations of our text within a vector will help the model achieve better performance by grouping similar words [10].

The Doc2Vec model allows us great flexibility to trade training time and model performance. In the simple case, we can train a Doc2Vec model for only one epoch to extract 300 features from our text that improves our Logistic Regression model’s performance by 7 points. This training only adds 3 minutes of training time to our model and gives us a testing performance of around 0.916. In the extreme case we can train this same Doc2Vec model for 30 epochs over 48 minutes to achieve an all-time highest accuracy of 0.9446.







There are obviously diminished returns in this training time. However, the testing time remains relatively consistent with this increased performance, increasing from a testing time of 0.014 for the model with 1 epoch of training, to 0.018 for the model with 30 epochs of training. This 30-epoch trained model gives accuracy near that of the SVM model with substantially better runtime performance.

CONCLUSION

The spread of misinformation on social media is a growing concern that can lead to significant harm to individuals and society as a whole. Manual moderation is often difficult and expensive, but machine learning approaches can be used to identify and flag fake news effectively. This project aimed to develop a model to automatically classify news articles into fake or true categories. Using natural language processing techniques and machine learning algorithms, the project analyzed a substantially large and varied dataset and achieved over 94% accuracy in classifying fake news using LIWC, Doc2Vec, and the Logistic Regression classifier.

REFERENCES

[1]
Godfrey-Smith, Peter (December 1989). "Misinformation". Canadian Journal of Philosophy. 19 (4): 533–550. doi:10.1080/00455091.1989.10716781.

[2]
Caramancion, Kevin Matthe (2021). "The Role of Information Organization and Knowledge Structuring in Combatting Misinformation: A Literary Analysis". Computational Data and Social Networks. Lecture Notes in Computer Science. Vol. 13116. pp. 319–329. doi:10.1007/978-3-030-91434-9_28.

[3]
Barberá, Pablo, et al. (21 August 2015). "Tweeting from left to right: Is online political communication more than an echo chamber?". Psychological Science. 26.10: 1531-1542. doi:10.1177/0956797615594620

[4]
D Olivera Mesa, AB Hogan, OJ Watson et al. Quantifying the impact of vaccine hesitancy in prolonging the need
for Non-Pharmaceutical Interventions to control the COVID-19 pandemic. Imperial College London (24-03-2021),
doi: https://doi.org/10.25561/87096.

[5]
Scheufele, Dietram; Krause, Nicole (April 16, 2019). "Science audiences, misinformation, and fake news".

Proceedings of the National Academy of Sciences. 116 (16): 7662–7669. doi: https://doi.org/10.1073%2Fpnas.1805871115 .
[6]
Zeerak Waseem. 2016. Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. In Proceedings of the First Workshop on NLP and Computational Social Science, pages 138–142, Austin, Texas. Association for Computational Linguistics.
doi: https://doi.org/10.18653/v1/W16-5618 .
[7]
The online information environment: Understanding how the internet shapes people's engagement with scientific information (PDF). The Royal Society. January 2022. ISBN 978-1-78252-567-7. Retrieved 21 February 2022.
[8]
Poesio, M. , Fornaciari, T. (2018). Detecting deception in text using NLP methods. SIGNAL-AI https://research.signal-ai.com/assets/Deception_Detection_with_NLP.pdf
[9]
Steven. “Misinformation & Fake News Text Dataset 79K.” *Kaggle*, 9 May 2022, https://www.kaggle.com/datasets/stevenpeutz/misinformation-fake-news-text-dataset-79k.
[10]
Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems 26 (2013).
doi: https://doi.org/10.48550/arXiv.1310.4546 , pdf: https://arxiv.org/pdf/1310.4546.pdf .