Detecting Fake News with Python and Machine Learning

Yatharth Mittal[1], Dr. Neha Aggarwal[2]
[1] B-tech scholar, Department of IT Maharaja Agrasen Institute of Technology[2]
[2]Assistant Professor, Department of IT Maharaja Agrasen Institute of Technology

---

Date of Submission: 25-11-2021 Date of Acceptance: 11-12-2021

Detecting Fake News with Python and Machine Learning

Problem Statement

Do you trust all the news you hear from social

media?All news are not real, right?

How will you detect fake news?

The answer is Python. By practicing this advanced python project of detecting fake news, you willeasily make a difference between real and fake news.

Before moving ahead in this machine learning project, get aware of the terms related to it like fakenews, tfidfvectorizer, Passive Aggressive Classifier.

TF-IDF is an abbreviation for Term Frequency Inverse Document Frequency. This is very **commonalgorithm to transform text into a meaningful representation of numbers** which is used to fit machine algorithm for prediction.

What is Fake News?

A type of yellow journalism, fake news encapsulates pieces of news that may be hoaxes and is generally spread through social media and other online media. This is often done to further or impose certain ideas and is often achieved with political agendas. Such news items may contain falseand/or exaggerated claims, and may end up being viralized by algorithms, and users may end up in afilter bubble.

What is a TfidfVectorizer?

TF (Term Frequency): The number of times a word appears in a document is its Term Frequency. A higher value means a term appears more often than others, and so, the document is a good matchwhen the term is part of the search terms.

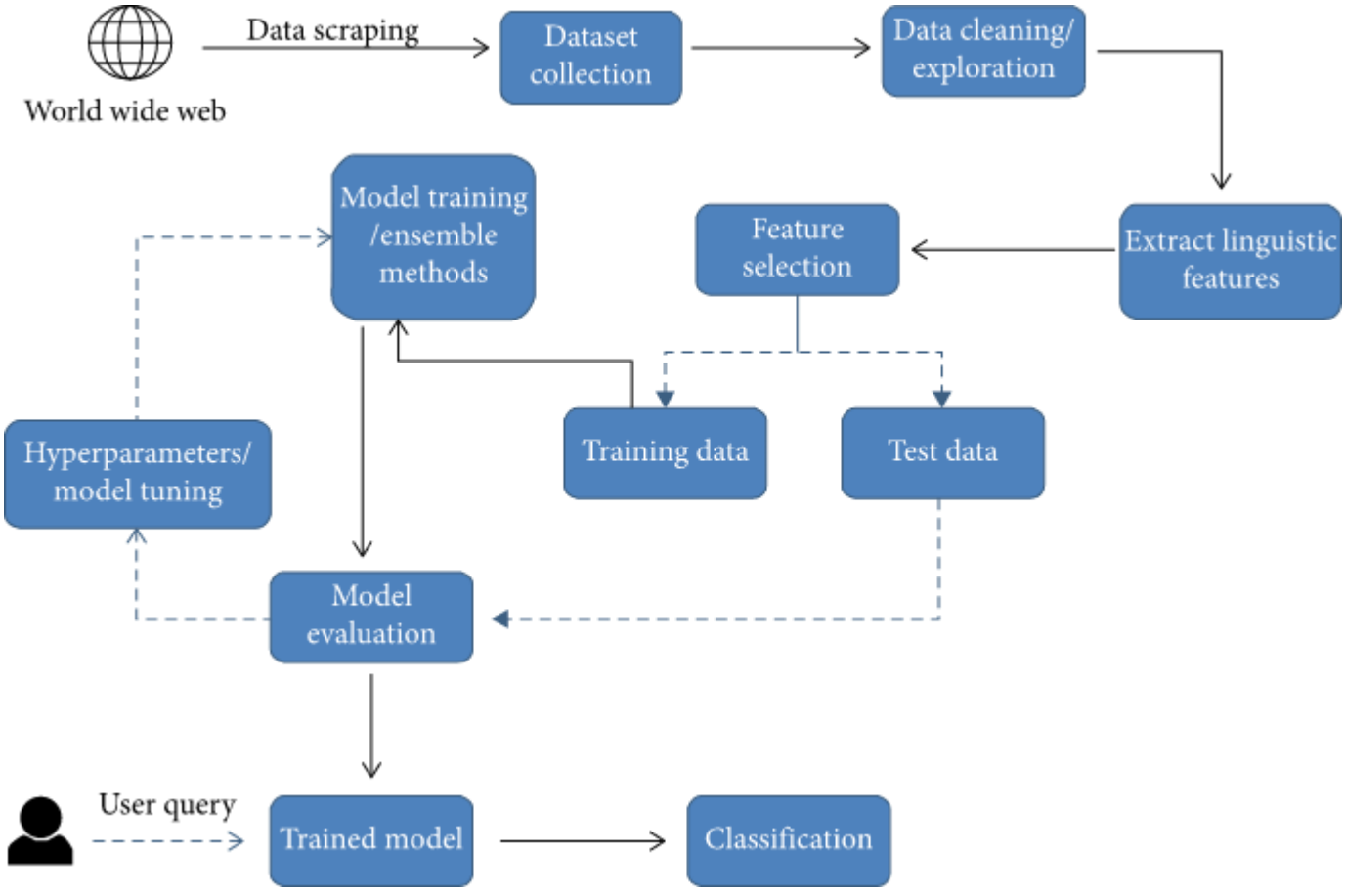
IDF (Inverse Document Frequency): Words that occur many times a document, but also occur manytimes in many others, may be irrelevant. IDF is a measure of how significant a term is in the entire corpus.

The TfidfVectorizer converts a collection of raw documents into a matrix of TF-IDF

features.

What is a PassiveAggressiveClassifier?

Passive Aggressive algorithms are online learning algorithms. Such an algorithm remains passive fora correct classification outcome, and turns aggressive in the event of a miscalculation, updating andadjusting. Unlike most other algorithms, it does not converge. Its purpose is to make updates that correct the loss, causing very little change in the norm of the weight vector.



Methodology

Detecting Fake News with Python

To build a model to accurately classify a piece of news as REAL or FAKE.

About Detecting Fake News with Python

This advanced python project of detecting fake news deals with fake and real news. Using sklearn, we build a TfidfVectorizer on our dataset. Then, we initialize a PassiveAggressive Classifier and fit themodel. In the end, the accuracy score and the confusion matrix tell us how well our model fares.

The fake news Dataset

The dataset we'll use for this python project- we'll call it news.csv. This dataset has a shape of 7796×4. The first column identifies the news, the second and third are the title and text, and the fourth column has labels denoting whether the news is REAL or FAKE.

Importing the libraries

We imported some libraries which can be useful, we imported as fewer libraries aspossible because initially, we wanted to make models from scratch instead of using sklearn.

- **Deleting some columns**

We tried to see the variables which have the least effect on the data so we dropped those usingthe dropna functionality. This reduced the useless information, from the dataset and made surethat our model was not overfitted.

- **Filling**

The missing values: Instead of just taking median values to fill missing values, we used a betteridea. If a doctor does not have a report, he/she would assume that report to be normal, so we googled the average values of those features for normal human being and used that value to fillmissing features.

- **Normalization:**

We had to normalize the values in each column to treat every column equally. This helped us inavoiding giving extra weightage to the features which generally have a larger value than the others.

Summary

Today, we learned to detect fake news with Python. We took a political dataset, implemented a TfidfVectorizer, initialized a PassiveAggressiveClassifier, and fit our model. We ended up obtainingan accuracy of 92.82% in magnitude.

Conclusion and Future Scope

As we can see around us as Facebook whistle blower and other social media platform has been detecting and working on fake news which can effect billions of people we need a robust mechanismto detect fake news in different languages, ethnicities, religions and much more. So this is the need of the future and by implementing projects like this on a large scale we can actual reduce the chaos that is caused by social media and reduce the spread of fake news on social media.

References

- Abdullah-All-Tanvir, Mahir, E. M., Akhter S., & Huq, M. R. (2019). Detecting Fake News using
- Machine Learning and Deep Learning Algorithms. 7th International Conference on Smart
- Computing & Communications (ICSCC), Sarawak, Malaysia, Malaysia, 2019, pp.1-5,
- https://doi.org/10.1109/ICSCC.2019.8843612
- Ahmed, H., Traore, I., & Saad, S. (2017). Detection of online fake news using n-gram analysis and
- machine learning techniques. Proceedings of the International Conference on Intelligent,
- Secure, and Dependable Systems in Distributed and Cloud Environments, 127–138, Springer,
- Vancouver, Canada, 2017. https://doi.org/10.1007/978-3-319-69155-8_9
- Ahmed, H., Traoré, I., & Saad, S. (2018). Detecting opinion spams and fake news using text
- classification. Secur. Priv., 1(1), 1-15. https://doi.org/10.1002/spy2.9
- Al Asaad, B., & Erascu, M. (2018). A Tool for Fake News Detection. 2018 20th International
- Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC),
- Timisoara, Romania, 2018, pp.379-386. https://doi.org/10.1109/SYNASC.2018.00064
- Aphiwongsophon, S., & Chongstitvatana, P. (2018). Detecting Fake News with Machine Learning
- Method. 2018 15th International Conference on Electrical Engineering/Electronics,
- Computer, Telecommunications and Information Technology (ECTI-CON), 528-531.