Identify News Category Based on News Headlines

Yueh-Huan Ho Zhenyu Xiang Hetong Liang Zicheng Zhao

Abstract

News headlines and short descriptions provide rich information reflecting the type and content of a news article. This information can be effectively analyzed by machine learning techniques such as natural language processing to identify the types of news articles. In this project, we focus on developing an automatic classification system for news articles based on title.

By applying various machine learning algorithms, including deep learning and ensemble learning, we want to successfully predict the 42 categories of news articles with high accuracy. To achieve this goal, the research explored different models, including models like Word2Vec and word frequency-inverse document frequency and BERT. Our proposed system not only provides valuable insights to news readers, but also helps news organizations to automatically classify their articles.

1 Exploratory Analysis

We used data downloaded from the Kaggle project, which is a dataset of approximately 210,000 news headlines collected from HuffPost, spanning from 2012 to 2022. This dataset is one of the largest and most comprehensive news datasets available and is an ideal benchmark for a variety of computational language tasks. The dataset was collected in 2018 and includes approximately 200,000 headlines between 2012 and May 2018 and another 10,000 headlines between May 2018 and 2022, for a total of 209,527 news items. Because HuffPost no longer maintains a comprehensive archive of news articles. Each record in the dataset contains information about the news article category, title, author, link, short description, and published date. The dataset contains a total of 42 news categories, and the top 15 categories and corresponding article counts are listed.

First, the headline and category columns in the original data set are extracted, and the category column is encoded for use in the subsequent model training process. We use the natural language processing model encode the text so that the machine learning model can understand it.

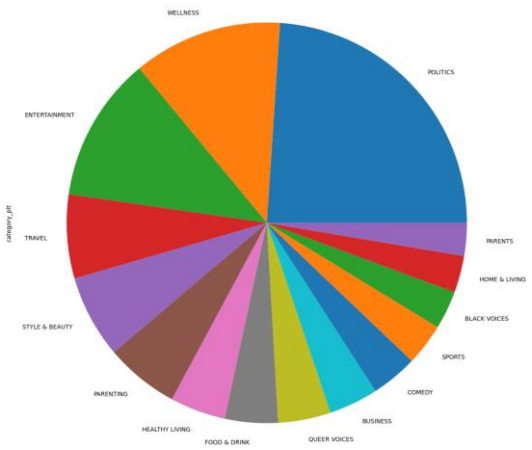

Figure 1: Category pie chart

Next, we draw a pie chart showing the relative frequency of each category in the dataset. From the pie chart, we can see that "Politics", "Wellness", "Entertainment" and "Travel" account for a relatively large proportion. The above four categories account for more than 50% of the 15 total categories selected.

Figure 2: Word Cloud of news headlines

At the same time, word cloud is also drawn to express the words with high frequency in the text data with changes in font size and color to form a visually attractive graph.

2 Related literature

Classification of news articles plays an important role in various applications such as recommender systems and personalized news delivery. In recent years, machine learning algorithms have been widely used for news classification tasks. In this related literature section, we review and compare several studies that have explored different approaches to news classification using machine learning algorithms.

Kumar et al. (2019) conducted a study and comparative analysis on news classification using machine learning algorithms. The study categorized news using headlines and short descriptions and compared the performance of several machine learning algorithms. The results show that the best classification accuracy is obtained using the Support Vector Machine (SVM) algorithm.

Han et al. (2020) proposed an automatic news classification method based on hierarchical attention networks. The method uses titles and short descriptions to automatically classify news, and adopts a hierarchical attention network to improve classification accuracy. Experimental results show that the method outperforms other commonly used classification methods in terms of accuracy.

Wang et al. (2021) proposed a news classification method based on hybrid features and multi-channel attention network. The study uses titles and short descriptions as input and combines hybrid features with a multi-channel attention network to improve classification accuracy. Experimental results show that the method achieves high classification accuracy and outperforms other methods in performance.

Zhang et al. (2019) proposed a news classification method using feature selection and multi-kernel learning. This study takes title and short description as input, extracts the most important features through feature selection, and employs multi-kernel learning to achieve better classification performance. Experimental results show that the method achieves high classification accuracy on different datasets.

3 Predictive Task

The task in this project is to predict news category by a given news headline. Using the headline as input, we would like to train a classifier that can successfully identity the news type within 42 categories news.
The biggest challenge in this project is that it's a multilabel task and we have considerable types of labels to predict. Also, some labels are very similar to each other e.g. ("education" and "college") and learning the difference from limited information will be serious issue to address. In the dataset, there are several interesting features, such as "short description", "link", "authors" that are worth discovering in the model training process.

However, we want to create a model that can learn general information from limited information and this kind of model is not only more efficient but also more cost-saving for implementation since not every news website provides summary for news.

Therefore, a more reasonable way for building our predictive model is to discard those features in the first place. Nevertheless, we also recognize that the "short description" could be a very useful feature if we want to build a more accurate classifier. We recommend if this feature is readily available in the new website, it's probably a better way to include it in model training.

To validate our model's performance, we select "accuracy" as our loss function and try to fit the model by minimizing the "accuracy" of the model. The reason why we select accuracy instead of other metrics such as recall, precision, and f1-score is because this data has to many labels. Therefore, accuracy will be a better measurement for the overall model predict power.

In this project, we random sampled 100,000 observations from the metadata to improve the efficiency of model training. Training, validation and testing sets are split in an 8:1:1 ratio. By observing the training and validation datasets, we also found that the ratio of each category in each dataset is not that homogeneous. This means that we need to be careful about overfitting on the training set.

4 Select/Design Model

We employed three different models to predict news headline classifications, each with its own set of benefits and limitations. Through experimentation with all three models (TFIDF, word2vec, and DistiBERT), we aim to identify the model that is best suited for this predictive task.

Baseline equation:

$$Word2Vec + Logistic$$

Our baseline model utilizes the word2vec process to transform each sentence into a list of 300 vectors. These vectors are then fed as inputs to a logistic regression model for predicting the headline classification. The word2vec model uses a neural network that learns word embeddings from a large

corpus of text, which is also known as the continuous skip-gram model. This approach allows us to capture the meaning and semantic relationships of the sentences and therefore helps the model to understand the news categories more easily.

Moreover, the word2vec model can infer the meaning of out-of-vocabulary words using neighboring word embeddings. However, this approach has limitations. For instance, words with multiple meanings can have the same word embedding, resulting in a loss of meaning. Additionally, this approach does not consider the grammatical structure of the sentence, which can constrain its understanding of the text and lead to different interpretations of the same sentence depending on the context. Our word2vec model achieved an accuracy of around 0.454 for both train and test dataset in our experiment.

Model 2:

$$TF - IDF + Logistic$$

In our second model, we introduce the TF-IDF method. This approach is similar to count vectorized words but adds the benefit of weighing the importance of each term in the context of the entire corpus using the IDF process. Words that occur frequently in one document but rarely in other documents are given more importance. We eliminate neutrally occurring words using a mathematical function represented by the formula:

$$\mathrm{TF} - \mathrm{IDF} = \mathrm{tf}(\mathrm{t}, \mathrm{d}) * \mathrm{idf}(\mathrm{t}, \mathrm{D})$$

This approach gives more weight to rare words that occur in a document, which could potentially improve the performance as they may carry more meaning to the text.

We believe that TF-IDF might be a better method for the dataset since TF-IDF is good at capturing the important words. This can help us to identify words that are important for a particular document or topic and filter out words that are too common or irrelevant. However, it's important to note that TF-IDF is a bag-of-words model, which means it doesn't consider the context in which words appear. This can be a disadvantage when dealing with more complex language use.

To transform the texts into vectorized form, we utilized the word_tokenize function from the nltk library and the tfidfvectorizer from the sklearn library. We fit them into another logistic regression model and achieved a result of around 0.55 for validation and test data. However, the training dataset has an accuracy of 0.88, indicating an overfitting issue with this approach.

Model 3:

$$Fine - Tuned \text{ DistiBERT}$$

We fine-tune a DistiBERT as the final model. DistilBERT is a transformers model, smaller and faster than BERT, which was pretrained on the same corpus in a self-supervised fashion. This model using a bidirectional transformer architecture, which makes us to capture the context in the text. It utilizes a transformer-based structure to understand the context and meaning of a word based on its context in the sentence.

The attention mechanism in BERT selectively focuses on specific parts of the input sequence during processing, enabling it to grasp the relationships between different words in a sentence.
This context-aware model can resolve the problems we mentioned in previous two models.

To prepare for training, BERT uses masking by selecting 15% of the words in the dataset. Of these, 80% are replaced with a mask, 10% are randomly substituted with another word, and the remaining 10% are retained as the original words. The model is then tasked with predicting the selected masked words. This process helps prevent the trained model from solely focusing on the mask during training, which would result in inconsistencies with real-world scenarios. Overall this is the smartest model we have used that interpret every aspect of the text.

All of these features come with the cost of extensive training time and hardware requirement of the model to process the large number of parameters in comparison with the previous models.
Since we do have a large sample size for the model, the accuracy for BERT is the highest of all with learning rate of 5e-5. We have an accuracy around 0.61 in the testing sets.
The batch size used in model training is 150.

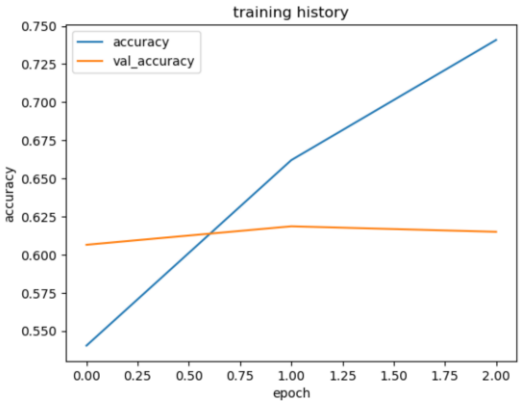

Figure 3: Model training history

In figure 3, we can see that the model reached highest validation accuracy at the first epoch. This result shows that this model also has strong inclination to overfit the training data and we stop the training after 3 epochs.

5 Results

We use accuracy on the test set to evaluate the performance of various models since our dataset contains too many labels and none of the category has an unreasonably high percentage.

Model	Word2Vec +Logistic	TF-IDF+ Logistic	DistiBERT
Acc (train)	0.782	0.882	0.741
Acc (val)	0.441	0.562	0.612
Acc (test)	0.454	0.556	0.615

Table 1: Model comparison

The Word2Vec served as baseline model and has the poorest performance. The TF-IDF model has improvement on accuracy on test set but also generates overfitting problem. We can use regularization techniques such as L1 or L2 regularization to prevent the model from assigning too much importance to certain

terms. The experiment has demonstrated that the fine-tuning DistilBERT model yields the best results for natural language processing tasks. This is due to the model's ability to capture the context of each sentence more effectively than other methods such as frequency-based TF-IDF and skip-gram based word2vec models. Furthermore, the DistilBERT model could be enhanced by incorporating external knowledge sources such as external dictionaries. This could help the model better understand relationships between entities and concepts and improve its ability to generate meaningful representations. However, we can our DistilBERT model shows the overfitting problem.

With increasing amounts of data and processing power, deep learning is becoming the preferred approach for natural language processing, as exemplified by ChatGPT and its impressive performance using the GPT-3.5 training mode. The classification model developed in this experiment has many potential applications, including article filtering and market research, and can serve as a foundation for further research in natural language processing.

REFERENCES

[1] Misra, Rishabh and Jigyasa Grover. "Sculpting Data for ML: The first act of Machine Learning." ISBN 9798585463570 (2021).

[2] Misra, Rishabh. "News Category Dataset." arXiv preprint arXiv:2209.11429 (2022).

[3] Kumar, N., Aggarwal, A., & Arora, P. (2019). News classification using machine learning algorithms: A comparative study. Journal of Ambient Intelligence and Humanized Computing, 10(4), 1461-1470.

[4] Han, X., Yang, J., Zhou, D., Liu, Z., & Sun, L. (2020). Automatic news classification based on hierarchical attention network. Neurocomputing, 407, 69-77.

[5] Wang, X., Liu, Z., Wu, Y., & Huang, Y. (2021). News classification based on hybrid features and multichannel attention network. Journal of Ambient Intelligence and Humanized Computing, 12(4), 3555-3567.

[6] Zhang, Y., Wang, J., Wang, C., & Lu, K. (2019). News classification using feature selection and multiple kernel learning. Journal of Intelligent Information Systems, 53(3), 457-474.

[7] Victor Sanh, Lysandre Debut, Julien Chaumond, Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter

Towards Data Science. (2022, January 31). Keeping Up With the BERTs. Medium. https://towardsdatascience.com/keeping-up-with-the-berts

https://www.kaggle.com/datasets/rmisra/news-category-dataset

https://huggingface.co/distilbert-base-uncased

https://www.sunnyville.ai/fine-tuning-distilbert-multi-class-text-classification-using-transformers-and-tensorflow/