EasyChair Preprint

№ 7529

IMDB Movie Reviews Sentiment Classification Using Deep Learning

Ravi Kumar and Angeline Benitta

March 11, 2022

IMDB MOVIE REVIEWS SENTIMENT CLASSIFICATION USING DEEP LEARNING

Ravi Kumar J [1]
Ravikumar94051@gmail.com
Dept. of Computer Applications
Hindustan Institute of Technology
Chennai, India

Angeline Benitta [2]
dangeline@hindustanuniv.ac.in
Asst. Professor, Dept. of Computer Applications Hindustan Institute of Technology Chennai, India

Abstract- Sentiment analysis is the most commonly used method for predicting user evaluations. Various machine-learning approaches have been used to make accurate predictions regarding the data. Long-term reliance and maximum pooling are not taken into account by these classifiers. In this research, we use Deep Learning technologies to classify reviews to enhance predictions utilizing these features. In this work with the Convolution Neural Network and the Long Short Term Memory Recurrent Neural Network to get higher accuracy with less loss and less time. The performance of six machine-learning algorithms in terms of sentiment analysis in the IMDB review dataset was tested in this research. One of these algorithms is based on neural networks, whereas the others are not. For sentiment analysis in IMDB, Binary classification it has been employed.

Keywords: movie reviews; sentiment analysis; KNN algorithm natural language toolkit.

I. INTRODUCTION

By the way, of definition sentiment analysis or opinion mining is the usage of text evaluation, or natural language processing (NLP) to get semantic quantification of the studied statistics. Using sentiment analysis, select textual content can be interpreted as a response to certain viewpoints (e.g. a tweet, or a product evaluation). Selection makers use these indicators consequently in making plans and taking appropriate actions along with advertising decisions, clients searching or enterprise expansion in specific geographic locations. Because of the large records evolution and the quantity of data being exchanged and produced every 2d, the urge to understand, mine, and examine this fact has remarkably multiplied. As a result, since the ordinary device learning strategies and neural networks did not suffice to be able to acquire this kind of big data, there was a shift to deep studying. Movie critiques are important for determining whether a movie worth seeing is worth their time.. A precis of all reviews for a movie can assist users in making this selection by way of no longer wasting their time studying all evaluations. Critics to put up remarks often use film-rating websites and fee movies that help visitors determine if the movie is worth watching. Sentiment evaluation can determine the mind-set of critics depending on their critiques. Sentiment evaluation of a movie evaluation can see how advantageous or poor a movie assessment is and subsequently the general rating for a film. As the technology is rising, users can directly give reviews about products, brands, etc. These reviews play a vital role in online shopping as well as help people to determine whether a product is good or not Obtained. The study concludes with several suggestions for a better model prediction. So, an automated process must be applied to mine these text data and analyze the sentiment effectively as the companies need to use these numerous amounts of data to improve their businesses by drawing more effective marketing analysis, product reviews, public relations, etc.

II. LITERATURE SURVEY

Title: The Parsimonious Rule-Based Version of Social Media Sentiment Analysis. Author: *Hutto, C. & Gilbert.* Abstract: Earlier, various rule-based approaches have been used for sentiment analysis. For example, Hutto and Gilbert presented a simple rule-based model for general sentiment analysis and found better performance than the benchmarks used in their study. However, the performance of their proposed model was not compared with neural network-based approaches. Popular Social Media website like Twitter has also been used for the sentiment

Title: Sentiment analysis of Twitter data. Author: *Agarwal, A., Xie, B., Vovsha, I., Ram bow, O., & Passonneau, R.* (2011) Abstract: Agarwal et al. Examined sentiment analysis on Twitter data by introducing Parts of Speech (POS) features.

Title: Recognizing contextual polarity in phrase-level sentiment analysis. Author: - *Wilson, T., Wiebe, J., & Hoffmann, P.* (2005). Abstract:-Wilson et al. Offered a new technique for word-stage sentiment evaluation that first decided whether an expression turned neutral or polar, after, which disambiguated the polarity of the polar expressions.

Title: Joint sentiment/topic model for sentiment analysis. Author: - Lin, C., & He, Y. (2009). Abstract: - They developed an unsupervised probabilistic modeling framework based on Latent Dirichlet Allocation (LDA), called the joint sentiment/topic model (JST).

Title: Sentiment analysis of blogs by combining lexical. Author: - Melville, P., Gryc, W., & Lawrence R. D (2019) Abstract: - Among supervised models, Melville et al. developed an effective framework for incorporating lexical knowledge and successfully applied the tools.

EXISTING SYSTEM

An invaluable concept for online users is Sentiment Analysis. These days, users tend to evaluate products, movies, health care systems, etc. Based on user reviews. The field of sentiment analysis examines the sentiments expressed by online users in the form of reviews or tweets. Sentimental Analysis concentrates on categorizing reviews based on their clarity. The sentiment or opinion expressed in reviews is analyzed using machine learning algorithms such as Support Vector Machines, Naive Bayes, k Nearest Neighbors, k Means, Random Forests, Dimensionality Reduction Algorithms, etc. The algorithms have their Limitation.

PROPOSED SYSTEM

In the proposed work, Deep Learning techniques are used for the classification of reviews based on their sentiment. The core of the proposed work is a combination of long short memory (LSTM) and recurrent neural network (RNN) used to classify the sentiments with a high degree of accuracy in a short period. CNN and LSTM work together to consider long-term dependency in order to improve classification accuracy and speed.

ADVANTAGES:

- Result is better by using Deep Learning
- Comparatively take lesser computing time

MODULES

- DATA COLLECTION
- DATA PRE-PROCESSING
- FEATURE EXTRACTION
- EVALUATION MODEL

DATA COLLECTION

Data used in this paper is a set of movie reviews collected of movies. This step involves selecting the subset of the data that you will be working with. A ML problem begins with data, preferably lots of data (examples or observations) for which you already know the answer.

DATA PRE-PROCESSING

By formatting, cleaning, and sampling your selected data, you can organize it .Three common data pre-processing steps are

STEP 1: FORMATTING

It is possible that the data you selected is not in a format that can be manipulated by you. The data may be in a relational database and you would like it in a flat file, or the data may be in a proprietary file format and you would like it in a relational database or a text file.

STEP 2: CLEANING

Cleaning data is the removal or fixing of missing data. There may be data instances that are incomplete and do not carry the data you believe you need to address the problem. These instances may need to be removed. Additionally, there may be sensitive information in some of the attributes and these attributes may need to be anonymized or removed from the data entirely.

STEP 3: SAMPLING

There may be far more selected data available than you need to work with. A much larger dataset can mean much longer running times for algorithms and greater computational requirements. You can take a smaller representative sample of the selected dataset that may be faster during the exploratory stage of prototyping solutions before you look at the whole dataset.

FEATURE EXTRACTION

Feature extraction is an attribute reduction process. Unlike feature selection, which ranks the existing attributes according to their predictive significance, feature extraction transforms the attributes into features. We use the labeled dataset gathered. The rest of the unlabeled data will be used to evaluate the models. Some machine learning algorithms were used to classify pre-processed data.

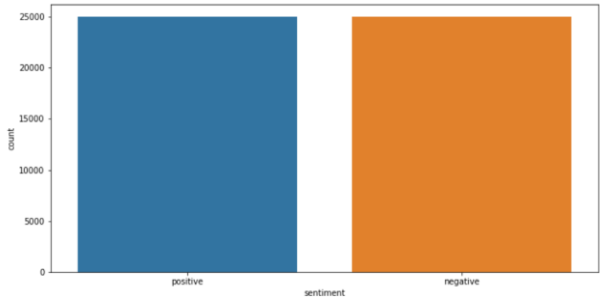
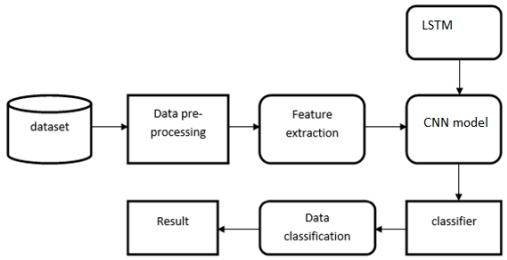


Fig. 1. Review chart

EVALUATION: We scored the classifier on unseen test data and calculated the R squared values for both the training and test data.

ACCURACY: There is defined as the percentage of correct predictions for the test data. It can be calculated easily by dividing the number of correct predictions by the number of total predictions.



NATURAL LANGUAGE PROCESSING

A) FEATURE SELECTION AND PREPROCESSING

In this part of CRISP-DM is important for feature selection and preprocessing, which represent the main part of data preparation in Artificial Intelligence. The success of text analysis may be greatly impacted by feature selection and preprocessing in NLP. Most of this is due to the unstructured and arbitrary nature of text data. Furthermore, machines need structure and numerical data. Couples of approaches for this transformation task e.g. word embedding or the vector space model exist. In this section, a discussion of the theoretical basis of different preprocessing and feature selection techniques that provided accompanied by an English phrase that illustrates the application of preprocessing. Nevertheless, every routine should be with care. It is not always true that a reasonably good preprocessing method leads to better results. In the practical section of this thesis, we must assess every suggested method separately.

B) TOKENIZATION

In order to process natural language texts, it is necessary to segment them into smaller units, called tokens. Tokenizing textual content allows computer systems to create the distinction between individual entities of content. Normally tokens represent simple words, which can be the smallest independent units of herbal language. Furthermore, token experiment includes idioms or hyphens, e.g. „person-generated ". Tokenization breaks running texts into brief text entities and is the first actual assignment in any text preprocessing cycle. Besides the partition of small units, complete sentences can also be the output of a tokenizer. An easy word tokenizer can be found in many languages using splitting the textual content on the occurrences of space symbols. This simple baseline approach does have a couple of downsides, due to the lack of figuring out phrases that semantically belong together. however, an easy tokenizer divides the word, which became introduced above, is into the following five tokens using tokens, so-referred to as n-grams can be created, which imply a token set with the length of n. „Gramma "is the Greek word for letter or token. When talking approximately a set of n letters in phrases, it is far approximately character grams. Prevent word elimination a very essential method to lessen the massive uncooked input space in NLP is prevent word elimination (SWR). Maximum languages have unique words, which do appear more regularly than others do or do no longer encompass a great deal of information approximately the content material of the text, e.g. auxiliary verbs or articles. Due to this, it often takes experience to exclude this so-referred to as prevent words in similar analysis. In English such phrases will be "the", "a" or "an" and for German regular prevent words are the articles "der", "die" and "das". The elimination may be accomplished by using checking the words in opposition to a standardized stop phrase listing. Those lists are to be had in literature and are often applied in different software applications. In our instance, „the "and „ is "are removed. Stop phrase removal should be used with care, mainly in sentiment evaluation, which attempts to predict a tremendous or poor goal of a text. The elimination might exclude words that can trade a whole declaration, together with „now not "or „none".

C) STEMMING

Besides, forestall phrase elimination, stemming is a beneficial technique to map phrases to their word stems and further reduce the entered measurement. This helps to extract the real that means contemporary a text and makes the unstructured information better on hand for a machine. The first stemming set of rules based on deleting longest suffixes and spelling exceptions changed into advanced in 1968. Through now, the porter-stemming algorithm is a state-of-the-art method and strips suffixes from phrases to hold the word stem. The approach plays well in English, however there are some disadvantages when it comes to the German language, including the fact that German is not strictly built through suffixes .However, there is a German equivalent based on Porter's concept and the Snowball string processing language. The English Porter Stemmer assigns the following phrases to those phrases that describe quality foxes and going for walks

ALGORITHM

K-MEANS CLUSTERING: Puts data into some groups (k) that each contains data with similar characteristics (as determined by the model, not in advance by humans) and it is the Clustering type. The K-NN algorithm stores all the available data, and classifies a new data point based on the similarity. This means that when new data appears, it can be assigned to a category that is well suited with the old data. A large amount of training data can be more effective, since it is simple to implement and robust to noisy training data.

Applications of deep learning applications

AI IN FINANCE: The financial generation area has already started out using AI to store time, lessen expenses, and upload value. Deep studying is converting the lending enterprise by the usage of credit that is more robust scoring. Credit decision-makers can use AI for sturdy credit lending packages to reap quicker, greater accurate risk assessment, the usage of machine intelligence to factor within the person and capability of candidates.

AI IN HR: Below amour, a sports clothing company revolutionizes hiring and modernizes the candidate revel in with the assist of AI. In fact, below amour Reduces hiring time for its retail shops by 35%.

Under Armor faced a growing reputation interest back in 2012. They'd, on average, 30000 resumes a month. Analyzing all of those applications and begin to begin the screening and interview system become taking too long.

AI IN MARKETING: AI is a precious device for customer service control and personalization demanding situations. Improved speech reputation in call-center management and phone routing due to the software of AI strategies permits an extra seamless enjoy for clients. For example, deep getting to know the evaluation of audio permits systems to evaluate a patron's emotional tone. If the customer is responding poorly to the AI Chabot, the machine can be rerouted the conversation to actual, human operators that take over the issue. Other than the examples above, AI is extensively used in other sectors/industries

RESULTS:

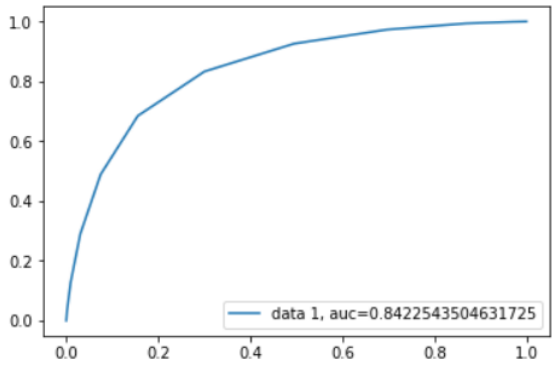


Fig. 2. Sentiment analysis graph

DISTANCE, TITLE

1. 0.38934644457877743, American Beauty (1999)
2. 0.3886146017673616, American History X (1998)
3. 0.38623530666805705, Pulp Fiction (1994)
4. 0.3716217352393141, "Lord of the Rings: The Return of the King, the (2003)"
5. 0.35016653951075516, Kill Bill: Vol. 1 (2003)
6. 0.34835832183548976, "Lord of the Rings: The Two Towers, the (2002)"
7. 0.34619612676017253, Eternal Sunshine of the Spotless Mind (2004)
8. 0.32621547353684777, "Matrix, the (1999)"
9. 0.3167773296626084," Lord of the Rings: The Fellowship of the Ring, the (2001)"
10. 0.27238038222579664, Fight Club (1999)

CONCLUSION

The proposed methods evaluate all the machine learning algorithms' prediction of accuracy level and thus it represents that neural network is having highest accuracy prediction nearly 98%. Sentiment Analysis is a sensitive topic, which requires efficient evaluation. Deep Learning Neural Networks efficiently classify reviews compared with machine learning algorithms. Neural Networks work great on large data. In order to achieve efficient classification, the LSTM considers the long-term dependency of sentences. Convolution Neural Networks classify with better accuracy in less time. Multi-layer Perceptron Neural Network classifies in a short time. Collaborative LSTM+CNN performs efficiently by considering long-term memory dependency in less time.

REFERENCES

[1] S. Mathapati, S. H. Manjula, and K. R. Venugopal, "Sentiment Analysis and Opinion Mining from Social Media: A Review," Global Journal of Computer Science and Technology, vol. 16, no. 5, pp. 77–90, 2017.

[2] T. P. Sahu and S. Ahuja, "Sentiment Analysis of Movie Reviews: A Study on Feature Selection & Classification Algorithms," in IEEE International Conference on Microelectronics, Computing and Communications (MicroCom), pp. 1–6, 2016.

[3] W. Zhao, Z. Guan, L. Chen, X. He, D. Cai, B. Wang, and Q. Wang, "Weakly-supervised Deep Embedding for Product Review Sentiment Analysis," IEEE Transactions on Knowledge and Data Engineering, vol. 30, no. 1, pp. 185–197, 2018.

[4] K. R. Venugopal, K. G. Srinivasa, and L. M. Patnaik, Soft Computing for Data Mining Applications. Springer, 2009.

[5] M. Dragoni and G. Petrucci, "A Neural Word Embeddings Approach for Multi-domain Sentiment Analysis," IEEE Transactions on Affective Computing, vol. 8, no. 4, pp. 457–470, 2017.

[6] Y.-P. Ruan, Q. Chen, and Z.-H. Ling, "A Sequential Neural Encoder with Latent Structured Description for Modeling Sentences," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 2, pp. 231–242, 2018.

[7] R. L. Rosa, G. M. Schwartz, W. V. Ruggiero, and D. Z. Rodriguez, "A Knowledge-based Recommendation System that Includes Sentiment Analysis and Deep Learning," IEEE Transactions on Industrial Informatics, 2018.