Major Project Synopsis

on

Product Review Analysis Using Sentiment Analysis

In partial fulfillment of requirements for the degree

of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE & ENGINEERING

Submitted by:

ADITI SUGANDHI [18100BTCSE02655]

AKSHAT KOTHARI [8100BTCSE02659]

PRERNA BANGAD [18100BTCSE02717]

Under the guidance of

PROF. Mr. JUBER MIRZA

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

SHRI VAISHNAV INSTITUTE OF INFORMATION TECHNOLOGY

SHRI VAISHNAV VIDYAPEETH VISHWAVIDYALAYA , INDORE

JULY-DEC 2021

SHRI VAISHNAV INSTITUTE OF INFORMATION TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

Abstract

Sentiment analysis is defined as the process of mining of data, view, review or sentence to predict the emotion of the sentence through natural language processing (NLP). The sentiment analysis involve classification of text into three phase “Positive”, “Negative” or “Neutral”. It analyzes the data and labels the ‘better’ and ‘worse’ sentiment as positive and negative respectively. Thus, in the past years, the World Wide Web (WWW) has become a huge source of raw data generated custom or user. Using social media, e-commerce website, movies reviews such as Facebook, twitter, Amazon, Flipkart etc. user share their views, feelings in a convenient way. In WWW, where millions of people express their views in their daily interaction, either in the social media or in e-commence which can be their sentiments and opinions about particular thing. These growing raw data are an extremely high source of information for any kind of decision making process either positive or negative. To analysis of such huge data automatically, the field of sentiment analysis has turn up. The main aim of sentiment analysis is to identifying polarity of the data in the Web and classifying them. Sentiment analysis is text based analysis, but there are certain challenges to find the accurate polarity of the sentence. To find polarity or sentiment of, user or customer reviews we are going to use some of the data analysis techniques.

The Two basic techniques for sentiment analysis

1. Rule-based sentiment analysis

The first technique is rules-based and uses a dictionary of words labelled by sentiment to determine the sentiment of a sentence. Sentiment scores typically need to be combined with additional rules to mitigate sentences containing negations, sarcasm, or dependent clauses.

2. Machine Learning (ML) based sentiment analysis

Here, we train an ML model to recognize the sentiment based on the words and their order using a sentiment-labelled training set. This approach depends largely on the type of algorithm and the quality of the training data used.

In our project we are going to use Deep Learning LSTM model.

Deep Learning techniques are also known as Artificial Neural Networks. These techniques have given great advances in Natural Language Processing in the last few years.

One particular model known as the LSTM (Long Short-Term Memory) has been dominating most NLP tasks in the last few years achieving state of the art results. An LSTM approach reads text sequentially and stores relevant information to the task at hand.

So, we this approach we are trying to build a system which can predict the sentiments of reviews posted on public platforms which can help buyers to choose among the pool of brands and products which is best suited for them.

1. INTRODUCTION

Sentiment analysis is a kind of text classification that classifies texts based on the sentimental orientation (SO) of opinions they contain. Sentiment analysis of product reviews has recently become very popular in text mining and computational linguistics research. The following example provides an overall idea of the challenge. The sentences below are extracted from a movie review on the Internet Movie Database:

"It is quite boring..... the acting is brilliant, especially Massimo Troisi."

In the example, the author stated that "it" (the movie) is quite boring but the acting is brilliant. Understanding such sentiments involves several tasks. Firstly, evaluative terms expressing opinions must be extracted from the review. Secondly, the SO, or the polarity, of the opinions must be determined. For instance, "boring" and "brilliant" respectively carry a negative and a positive opinion. Thirdly, the opinion strength, or the intensity, of an opinion should also be determined. For instance, both "brilliant" and "good" indicate positive opinions, but "brilliant" obviously implies a stronger preference. Finally, the review is classified with respect to sentiment classes, such as Positive and Negative, based on the SO of the opinions it contains.

Sentiment analysis can be used to focus on the customer feedback verbatims where the sentiment is strongly negative. Likewise, we can look at positive customer comments to find out why these customers love us. Only after these sentiment analysis have been conducted successfully, we can focus on increasing the number of our promoters.

When used in combination with Thematic analysis, we can further narrow down this information to find precisely which themes are talked about with positive/negative sentiment. This provides actionable insights for your business.

Evidently, sentiment analysis is being used by taking a source of text data that has a narrow scope of context and then gauging the polarity of the text.

Some existing work involves analysis at different levels. Specifically, the SO of opinion words or phrases can be aggregated to determine the overall SO of a sentence or that of a review. Most existing sentiment analysis algorithms were designed for binary classification, meaning that they assign opinions or reviews to bipolar classes such as Positive or Negative. Some recently proposed algorithms extend binary sentiment classification to classify reviews with respect to multi-point rating scales, a problem known as rating inference. Some sentiment analysis algorithms aim at summarizing the opinions expressed in reviews towards a given product or its features. Sentiment analysis is closely related to subjectivity analysis (Wiebe et al., 2001; Esuli and Sebastiani, 2005). Subjectivity analysis determines whether a given text is subjective or objective in nature. It has been addressed using two methods in sentiment analysis algorithms. The first method considers subjectivity analysis a binary classification problem, for exam ple, using Subjective and Objective as class labels. Pang and Lee (2005) adopted this method to identify subjective sentences in movie reviews. The second method makes use of part-of-speech (POS) information about words to identify opinions (Turney, 2002; Hu and Liu, 2004a; Leung et al., forthcoming) because previous work on subjectivity analysis suggests that adjectives usually have significant correlation with subjectivity (Bruce and Wiebe, 1999; Wiebe et al., 2001).

2. PROBLEM DOMAIN

In today's environment where we're suffering from data overload, companies might have mountains of customer feedback collected. Yet for mere humans, it's still impossible to analyze it manually without any sort of error or bias.
Oftentimes, companies with the best intentions find themselves in an insights vacuum. We know we need insights to inform your decision making. And we know that we're lacking them. But we don't know how best to get them.
Sentiment analysis provides answers into what the most important issues are. Because sentiment analysis can be automated, decisions can be made based on a significant amount of data rather than plain intuition that isn't always right.
As it is impossible to analyze large amounts of data without error. Let's Imagine this scenario: we're the owner of a small delivery business and we receive about 20 responses to your email surveys every month. We should read these ourself and perform our own analysis by hand.
Now, imagine receiving 30,000 responses per month. That's more than a thousand responses each day! Needless to say this is impossible as a part of a business owner's day job.
Sentiment analysis is important because companies want their brand being perceived positively, or at least more positively than the brands of competitors.
Beside this issue faced by companies, it is also difficult for buyers to select a product from the pool of companies. Before buying anything we always try to check that is that the right choice I am going to make? What are the reviews of this product? How people are liking it? etc, etc…..
Now instead on searching on different platforms for reviews for the product, with the interface provided by us people can easily find out which product is best for them.

Main objectives of our project are-
1 To provide a convenient and easy way to judge a product on the basis of customers review published on public platforms.

2. To provide an interface to the companies to check the value of their product in the market. As well as to the buyers to select the best product available in the market for them.

3. SOLUTION DOMAIN

The solution for the problem proposed above can be implemented according to the following steps.

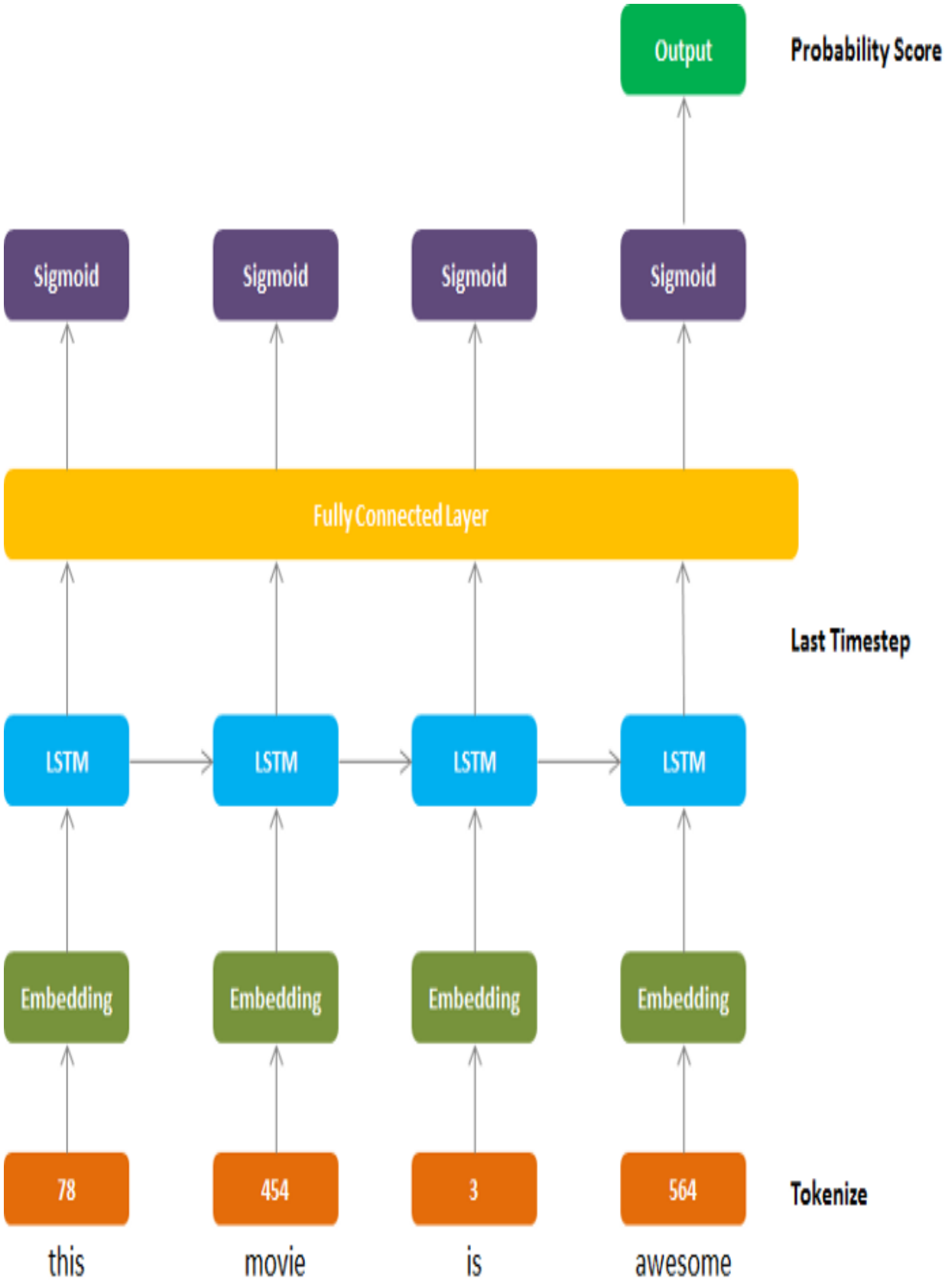
1) Data gathering- The most crucial task of the project would be to efficiently collect data from open platforms, this task needs to be automated and should be time efficient. To perform this we plan to build a Web Crawler using Python.

2) Data cleaning- The gathered data will then be processed so that it can be further used for training. Data cleaning will involve :
    - Converting reviews to lower case
    - Removing special characters and punctuations.
    - Discarding unwanted reviews.
    - Creating list of reviews .

3) Tokenization- In this firstly, we will create an index mapping dictionary in such a way that our frequently occurring words are assigned lower indexes. One of the most common way of doing this is to use Counter method from Collections library.
   Next, we will encode the words by replacing the words with integers. This will create a list of lists. Each individual review is a list of integer values and all of them are stored in one huge list. Lastly , we will encode the labels as positive, negative and neutral.

4) Training, Validation, Test Dataset Split- Once we have got our data in nice shape, we will split it into training, validation and test sets
   train= 80% | valid = 10% | test = 10%
   After creating our training, test and validation data. Next step is to create dataloaders for this data. We can use generator function for batching our data into batches

5) The LSTM Network Architecture-
   The layers are as follows:
    - Tokenize : This is not a layer for LSTM network but a mandatory step of converting our words into tokens (integers)
    - Embedding Layer: that converts our word tokens (integers) into embedding of specific size
    - LSTM Layer: defined by hidden state dims and number of layers
    - Fully Connected Layer: that maps output of LSTM layer to a desired output size
    - Sigmoid Activation Layer: that turns all output values in a value between 0 and 1
    - Output: Sigmoid output from the last timestep is considered as the final output of this network



LSTM Architecture for Sentiment Analysis.

4. **SYSTEM DOMAIN**

The project is planned to build using python environment with a set of libraries offered in python programming language, the project structure and further development will be done using Pycharm IDE, reason for considering Pycharm over other IDE is because code formatting and debugging is fairly easier and Pycharm also offers a plethora of extension exclusively for python. Pycharm also allows easy deployment on cloud and git making it easier for project members to contribute easily. Since the project is planned to use Deep Learning the hardware requirements expected to have a System running the algorithm contain good specifications especially a mid to high end CPU or either a GPU that can easily train our ML algorithm.
An alternative of a high end PC would be using the services offered by the cloud for example Google Colab that can be used to train our ML algorithm on any available computer system, Google Colab allot GPUs to users for their use case. A case can also be made for firebase as it offers deployment services also.
There are various tools on the market for text analytics and sentiment analysis. At Thematic, we're focused on staying up to date with the latest NLP research and the most successful models used in academia, where there has been a huge amount of progress in the last 4-5 years. Our team at Thematic implements these models and then trains them on a specific dataset for customer feedback. Thereby, we can create a reliable, and accurate analysis.
The model that we are going to use is known as the LSTM (Long Short-Term Memory). It has been dominating most NLP tasks in the last few years achieving state of the art results. An LSTM approach reads text sequentially and stores relevant information to the task at hand.

Within the LSTM there are cells which control what information is remembered and what is forgotten. In the case of sentiment analysis negation is very important. For example, the difference between "great" and "not great". An LSTM trained to predict sentiment will learn that this is important and get good at understanding which words should be negated. By reading large amounts of text an LSTM can be thought of as 'learning' grammar rules.

Deep learning architectures continue to advance with innovations such as the Sentiment Neuron which is an unsupervised system (a system that does not need labelled training data) coming from Open.ai. Google has developed the Transformer and recently added pretraining (pre-training is where you train a model on a different task before fine tuning with your specialised dataset) to the transformer with a technique known as BERT , achieving state of the art results across many NLP tasks.

5. **APPLICATION DOMAIN**

Sentiment analysis is a uniquely powerful tool for businesses that are looking to measure attitudes, feelings and emotions regarding their brand. To date, the majority of sentiment analysis projects have been conducted almost exclusively by companies and brands through the use of social media data, survey responses and other hubs of user-generated content.

The future of sentiment analysis is going to continue to dig deeper, far past the surface of the number of likes, comments and shares, and aim to reach, and truly understand, the significance of social media interactions and what they tell us about the consumers behind the screens. This forecast also predicts broader applications for sentiment analysis – brands will continue to leverage this tool, but so will individuals in the public eye, governments, nonprofits, education centers and many other organizations.

Sentiment analysis is on the verge of breaking into new areas of application. While we will likely always think of it first in terms of the traditional marketing sense, the world has already seen a few ways that sentiment analysis can be used in other areas. Social media analytics helped predict and explain the emotions of concerned parties behind Brexit and the 2016 US election, which has spurred a number of non-brand organizations to investigate how sentiment analysis can be used to predict outcomes and map out the emotional landscape of people, voters and the like. Additionally, businesses are looking at ways that sentiment analysis can be used outside of their marketing and PR departments. Sentiment analysis simply looks more popular in the future.

The most popular applications of sentiment analysis in real life in today's time are:

- Social media monitoring
- Customer support
- Customer feedback
- Brand monitoring and reputation management
- Voice of customer (VoC)
- Voice of employee
- Product analysis
- Market research and competitive research

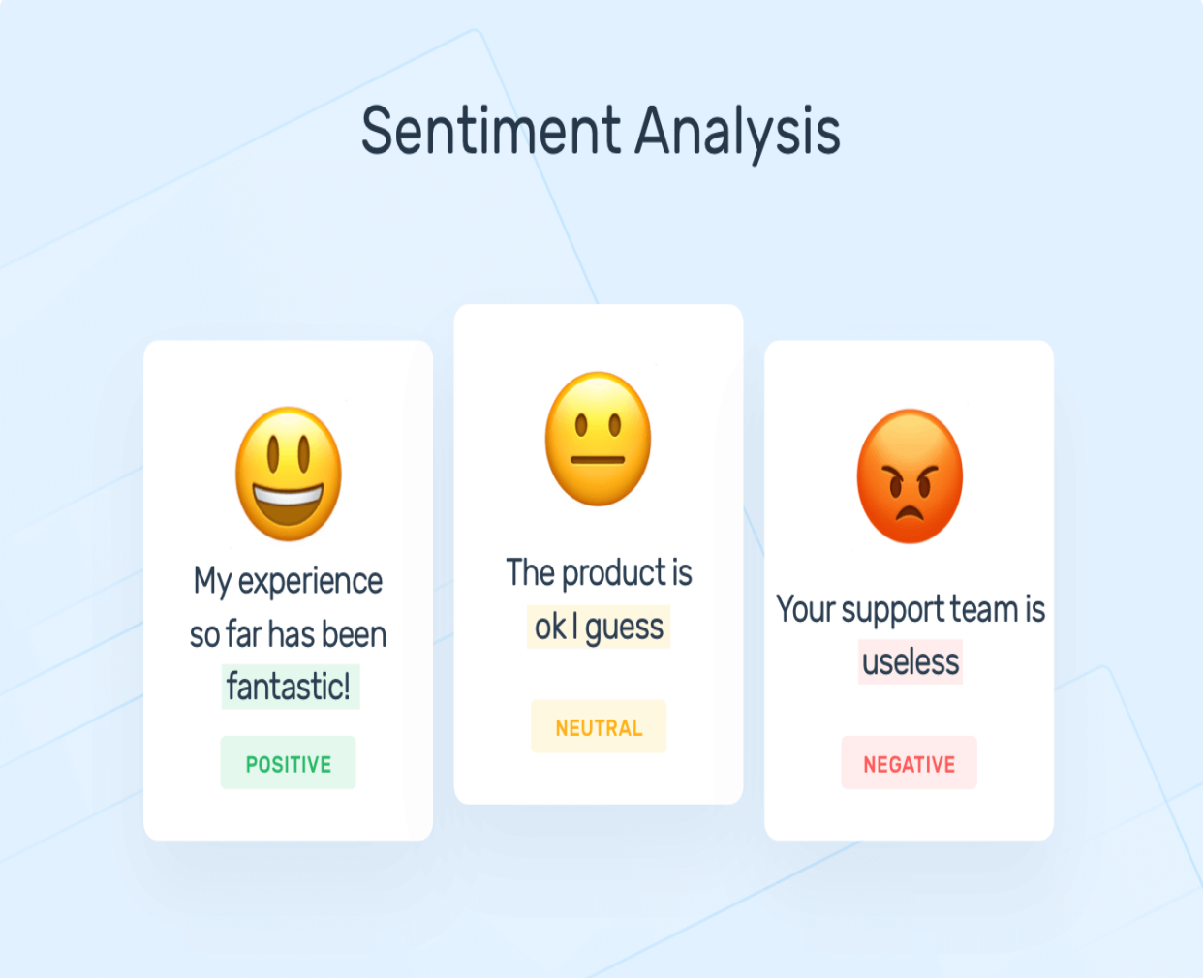
2022 is going to be another year that continues to drive the analytics machine forward. With more and more organizations turning to sentiment analysis to measure and predict outcomes, as well as better understand consumer behaviors, these tools are quickly building a reputation that is going to help propel it forward into the future and towards deeper and more accurate conclusions and insights.

Sentiment analysis has been an important tool for brands looking to learn more about how their customers are thinking and feeling. Additionally, the insights gained from these tools are becoming much deeper, as a result of emerging social media platforms and features.

6. **EXPECTED OUTCOME**

Predicted outcome of the system proposed by us are given as follows.

1. A website where people and company authorities can signup.
2. Users can select a particular company whose product they want to judge on or can simply select a product whose reviews one wants to analyze.
3. On selecting a product the buyer will get a generalized feedback on the basis of how positive, negative or neutral it's reviews are, represented in the form of a bar graph .
4. Along with the feedback it will also suggest a product with positive review percentage higher than the current product relevant to the product searched .
5. In the same way companies can also check reviews sentiment of people for their product specifying the detail which feature is liked by people the most, and which is disliked by many.



7. **REFERENCES**

1. https://www.researchgate.net/publication/228699116_Sentiment_Analysis_of_Product_Reviews

2. https://getthematic.com/insights/sentiment-analysis/

3. https://towardsdatascience.com/sentiment-analysis-using-lstm-step-by-step-50d074f09948

4. https://monkeylearn.com/blog/sentiment-analysis-of-product-reviews/#:~:text=Sentiment%20analysis%20is%20the%20automated,by%20Positive%2C%20Neutral%2C%20Negative.

5. Pang B, Lee L (2008) Opinion mining and sentiment analysis. Found Trends Inf Retr2(1-2): 1–135.**Article Google Scholar**

6. Stanford (2014) Sentiment 140. http://www.sentiment140.com/.

7. Sarvabhotla K, Pingali P, Varma V (2011) Sentiment classification: a lexical similarity based approach for extracting subjectivity in documents. Inf Retrieval14(3): 337–353. **Article Google Scholar**

8. (2014) Scikit-learn. http://scikit-learn.org/stable/.