Find Similar Tweets Within Health Related Topics

By
Danny Gilberto Villanueva Vega
A thesis submitted in partial fulfillment of the requirements for the degree
of
MASTER OF SCIENCE
in
COMPUTER ENGINEERING
UNIVERSITY OF PUERTO RICO
MAYAGÜEZ CAMPUS
2019

Approved by:

Manuel Rodríguez Martínez, Ph.D. President, Graduate Committee	Date
Wilson Rivera Gallego, Ph.D. Member, Graduate Committee	Date
Bienvenido J Velez Rivera, Ph.D. Member, Graduate Committee	Date
Graduate School Graduate School Representative	Date
Chairperson, Ph.D. Department Chairperson	Date

Abstract of Thesis Presented to the Graduate School

of the University of Puerto Rico in Partial Fulfillment of the

Requirements for the Degree of Master of Science in Computer Engineering

Find Similar Tweets Within Health Related Topics

here abstract of thesis.

Here abstract of thesis.

Here abstract of thesis.

Resumen de tesis presentada a la Escuela Graduada

de la Universidad de Puerto Rico como requisito parcial de los

requerimientos para el grado de Maestría en Ciencias en Ingeniería de Computadoras

Encontrar tweets similares en temas relacionados con la salud

Aquí resumen de tesis.

Aquí resumen de tesis.

Aquí resumen de tesis.

DEDICATION

To my Mom, Carin Vega Pérez. To my sister, Emyli S Rodriguez Vega.

Acknowledgments

to thank family

to thank advisor

This research is supported by the United States (US) National Library of Medicine of the National Institutes of Health (NIH) under award number R15LM012275. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. Some results presented in this thesis were obtained using the Chameleon Cloud supported by the National Science Foundation (NSF).

Contents

List of Abbreviations

AI Artificial Intelligence

CNN Convolutional Neural Network

DL Deep Learning

ML Machine Learning

MLP Multilayer Perceptron

NIH National Institutes of Health

NLP Natural Language Processing

NSF National Science Foundation

RNN Recurrent Neural Network

THS Twitter Health Surveillance

UPRM University of Puerto Rico Mayagüez Campus

US United States

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Social networks have become a very important means to share ideas, discuss news, and opinions on many topics. They also provide real-time information on sales, marketing, politics, natural disasters, and crisis situations, among others. These networks include Facebook, Twitter, WhatsApp, and Instagram, to name a few.

In this work, we shall focus our efforts on the Twitter social network. This network provides a mechanism for people to express their views using short messages (i.e., 280 characters) called *tweets*. Users of this network can find each other messages without the need of becoming "friends", as happens in other networks. The analysis of these tweets can enable us to understand the current situation regarding certain topics, for example, discussions related to medical topics (e.g., "flu"). Using the tweets, users can monitor and find patterns that give information about some type of disease being discussed in the social network. In addition, it is possible to detect the position, "mood", or sentiment of the people around some topic.

For the analysis of all this available information it is necessary to group or categorize the text along similarities in structure and or meaning. However, this is a challenging task, due to the complexity/ambiguity introduced by spelling errors or the use of informal language

("slang"). In the case of tweets, the small size of the message often makes it difficult to analyze without the context provided by previous messages or user interactions. Making use of the data stored in the Twitter Health Surveillance (THS) System at /acUPRM , as one of our sources, it is possible to process all the information more easily and quickly, and use it to analyze and process the data using Machine Learning (ML) algorithms.

The view of the world has changed with the presence of Artificial Intelligence (AI) in our lives, we live in a new world surrounding by ML (e.g. Amazon Alexa, writing correctors). Companies like Google, Amazon, Netflix and others are using AI algorithms to obtain value and insight of large amount of data that in otherwise will been impossible to analyze. The value of the information has been ever the key for the growth of companies; therefore, text analysis is a rough task aim to extract value information to use in business decisions, however this is challenging job due to complexity of Natural Language Processing (NLP) a field of ML focuses on analyzing the human language.

The detection of similarity in texts in their meaning of semantics content is a topic present in many researches because the need to obtain valuable and reliable information from the amount of available data over internet like, communication services (e.g. "Twitter"), feedback user, system log files, customer reviews to mention a few; and the data present in the same company about employees, clients and others.

In this project, we investigate and implement text similarity algorithms in such a way that we can: 1) know if they are related or not with a disease, 2) group similar tweets to those that we have already captured, analyzed or stored and, 3) find similarity index between tweets using different learning algorithms. We based our work on, semantic similarity approaches and text similarity measures using Deep Learning (DL) algorithms to deliver reliable information to the end user about health-related topics.

1.2 Objectives

- **Collect and filter the data file:** We select necessary tweets and filter all them which are related to health disease thereby, we use a clean data as input to train the algorithms to be implemented. Tweets were collected using THS System at University of Puerto Rico Mayagüez Campus (UPRM). In this way, it is convenient to describe the steps through the process of data selection until deliver of final cleaned up inputs. Part of inputs is labeled data by hand; thus, it is necessary a group of people to classify a measure of similarity between tweets that will be used in training sample.
- **Investigate and implements DL algorithms for text similarity measures** It is necessary investigate the state-of-the-art methods and techniques related to text analysis, and then build a robust architecture using DL algorithms like, Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) on NLP approaches, focusing in text similarity measures. The output of our trained models will be one of two of next options a) there is an acceptable similarity measure between the pair of tweets and, b) no exist enough similarity between the pair of tweets.
- **Test algorithms in a Big Data environment:** The algorithms will be tested to measure the performance and accuracy of results in THS cluster located in the Electrical and Computer Department at UPRM. Because DL algorithms consume a lot of resources, we also use virtual environments as Chameleon Cloud Platforms to test algorithms with better Graphic Processing Unit (GPU) resources than physical machines.

1.3 Contributions

- **Use social networks to get valuable information about health topics:** All data present on internet through social networks, entertainment apps and others, can be used in health-related research. In our case twitter is a source of amount of data of different topics that can be transform in important information relevant for medical issues. Here is showed how we can use data of medical conditions to build models able to compute the similarity in tweets therefore these could be used in future to support on medical applications.
- **Present DL models for text similarity analysis:** DL models are very powerful for analysis of data like images, sound, and text. In this project, we try to use these excellent tools of ML to figure out a better text similarity model.
- **Employ Supervised Learning in text similarity tasks:** Many studies about sentence representation ("encoding models") is based in Unsupervised Learning because there is not enough labeled data about a specific topic to train a model like in our case, with data diseases related. We show the trained models with labeled data in sentence similarity has enough performance to be widely adopted in others NLP tasks.
- **Use different measure of similarity:** In this project we used three methods to calculate similarity text. Most known is cosine similarity, also Frobenius Distance and our own distance measure called Triangular UL Distance based in part of linear algebra, in a special kind of square matrix called triangular matrix.
- **Describe the evaluation of similarity models:** We built models with CNN, RNN and merged approaches to get better results. They were tested with different setting to find the best model possible, we used the next metrics F1-Score, Precision and Recall.

1.4 Outline

The outline of this thesis is as follows. Chapter 2 contains the literature review about concept of ML and DL to contextualize the presented solution. In this chapter also in describe topics of NLP and text similarity methods. The problem description and the methodology followed to get the similarity models is described in Chapter 3. In Chapter 4, explains the ML architecture and describe the different DL models built using CNN and RNN approaches. Chapter 5 shows the result of performance and accuracy of all models described on above chapter. Finally, Chapter 6 shows the conclusions and future work to follow.

Chapter 2

Literature Review

2.1 Introduction

Advance of technology has been many changes in the world and our life is now surrounding for ML algorithms hence, companies of all sizes are following the large business' success using AI to draw insight that can be used to take better decisions.

The field of AI seeks to understand how humans think ("intelligence") and how build intelligent entities. Then what we use AI today for? Because the wide field, there is no a simple answer, but we can mention a few applications next, robotic vehicles, speech recognition, autonomous planning and scheduling, game playing, spam fighting, logistics planning, robotics, machine translation, to mention a few applications that exist today, combining efforts of science, engineering and mathematics [1].This applications need process a lot of data, hence it is necessary automate the process of analysis, here ML appears like a subfield of AI that automate the process of learning extracting patterns from the raw data to get insight [2].

Artificial Neural Networks are simply a collection of connected units that represent abstractly the human brain ("neurons") to aim achieve learning a specific task [1]. In this project we are working on similarity tasks. Today exist many applications of similarity like handwritten digits recognition, similarity images detection from text or an image in the web searchers (e.g. Google, Bing). Generic Neural Network techniques can be successfully applied for these problems, but to achieve better result and scale to large applications we need used techniques specialized on certain domains, for example in case of NLP ("text analysis") we need methods to process sequential data, like RNN, that is aDL technique part of ML field. [3]

2.2 Machine Learning

ML also knows as automated learning is associated with the concept of "to learn", this learning is composed of an input data that represent experience, using a learning algorithm is achieved an output with some expertise [4]. This learning is focus in gain knowledge, understanding, experience and skills [5] in such a manner its performance improves significantly.

Today we look that ML is in many practical applications that we use in frequently in our daily life like: movie recommendation, text translation, speech recognition [3], robotic vehicles, autonomous planning and scheduling, diagnosing diseases [1]. In essence, we are speaking about ML when an artificial intelligence system has the ability or capability to get knowledge or find patterns from raw data [3]. ML study data to detect patterns to be able to categorize, predict, identify unknown pattern and detect anomalies or unknown patterns, because Big Data we now have the advantage to process an amount of data, and using ML algorithms we can identify new opportunities to solve complex problems like self-driving cars, fraud detection, virtual assistants, resource optimization and more applications [6].

ML is a very wide field, for this reason it has branched into several sub-fields related to distinct tasks [4] and approaches to solve problems. For the study and application of these algorithm exist many ways to classify the learning paradigms, the best known is supervised learning and unsupervised learning, this distinction is by what kind of experience they are allowed to have during the learning process [3], others kinds of ML are semi-supervised learning and reinforcement learning.

2.2.1 Supervised Learning

Supervised Learning is based in an set of input-output pairs, these models learn automatically of the relationship between input features and target features, through a function that maps from the input data to output feature [1] [2]. This kind of learning is only possible if we know the target of the output data [5], and if we have the enough input labeled data to train the ML algorithm.

More abstractly, Supervised Learning describe a scenery where the training examples is a set of data that contains the necessary information to identify and associate it to the output value. This information is not available in the test data to which the learned model is being applied. The aim is that the acquired expertise will can predict the expected output [4].

It is called "supervised" because the environment provides an extra information, commonly known as labels, the model is trained with input data and target data [4]. The target value ("label") is provided by a supervisor who teach to the system what to do [3] , providing the correct output ("desired output") as a feedback for reduce error.

Supervised learning identify correlation and a logical pattern in data from the from state A to state B, after these pattern are learnt, we can transfer learning to solve similar problems [6] [7]. Common techniques used in this kind of ML are decision trees, forecasting, neural networks, support vector machines to name a few; and their applications are risk assessment; personalizing interaction; image, speech and text recognition; fraud detection; customer segmentation like the most knowns [6].

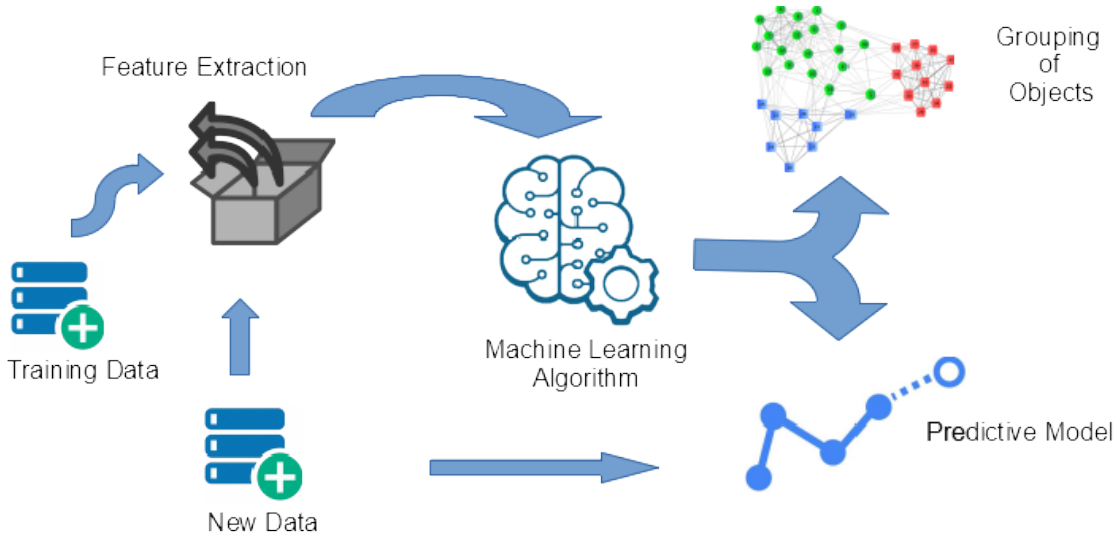

Fig 2.1: Supervised Learning Workflow.

Figure 2.1 shows the process of a supervised learning model. The input of the model is data that contains each element with its respective label, next the model extracts the most relevant features of the data, to find the relation between input and output, and construct a logical pattern. Finally, the model after the training step can predict new outputs given a new data without labels.

2.2.2 Unsupervised Learning

Today Big Data is a great tool that makes possible analyze a great amount of information, faster and easier. Raw data in many cases is difficult to analyze and many cases we do not have an answer key for our trainable data. In this cases we can use unsupervised learning

like an alternative, therefore we can determine correlations and identify pattern parsing the available data [6].

This kind of algorithms learns patterns without an explicit feedback ("label") [1], we only have a training data without target values for them [5]. There is no separation of training data and test data [4]; all the data is the input for the algorithm. Here we have a similitude with the human behavior, when they observe the world, usually they do inferences and group things based on their interaction with the environment, and they are guided by their observations and intuition, this learning is refined exposing to experience and a lot of observations [6].

A the most common technique for this learning is called clustering, [1] which classify data in meaningful categories, dividing the data into groups with similar characteristics called clusters [5] [3]. Other techniques are nearest neighbor mapping, singular value decomposition to name a few. Their applications are related to market basket analysis, anomaly/intrusion detection, text classification, identifying like things, sentiment analysis, and so on [6] [8].

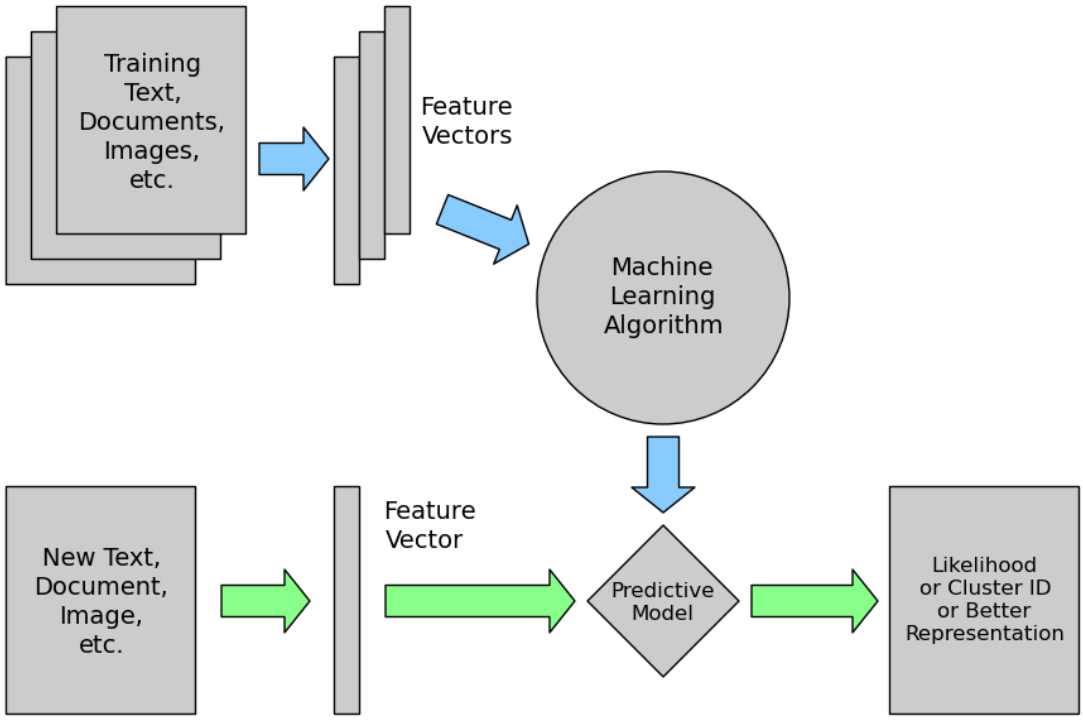

Fig 2.2: Unsupervised Learning Workflow.

In Figure 2.2 the features or patterns are extracted of the input data (text, documents, images, etc) data, represented in dimensionality vector, with this data the algorithm is

trained. The model created is able use new data to predict outputs like group likelihood or cluster ID.

2.3 Neural Networks

An artificial neural network is a very simplified abstract model of a biological brain, an interconnection of processing units with capability to learn patterns to generalize and associate data. A significant aspect adopted from biological brain is the "learning capability" from the experience and transfer knowledge to find reasonable solutions to similar tasks [9] [10].

When we speak about network, it can be referred from a simple single node to a collection of nodes [10]. The structures of neural networks are basically the next 1) a set of nodes linked associated with a numeric weight that represent the strength of connection between them, 2) an input function for each node that computes the weighted sum of all inputs and 3) an activation function to control the neuron behavior and get the desired output [10] [1].

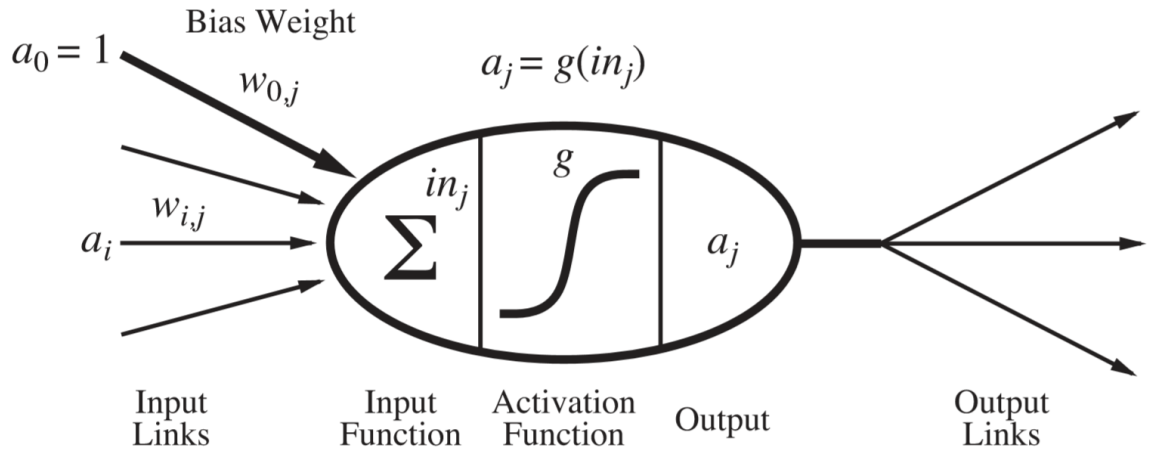

Fig 2.3: A Simple Mathematical Neuron Representation [1].

Figure 2.3 show the basic structure of a simple mathematical neuron, where output is result of apply an activation function (e.g. binary threshold, logistic function, Rectified linear unit function, etc.) to the weighted sum of inputs. The weights are important to minimize the cost of activation function and they are updated when the model is trained. The following equation represent the output after applying an activation function in a

neuron showed in Figure 2.3.

$$a_j = g\left(\sum_{i=0}^{n} w_{ij}\ a_i\right), \tag{2.1}$$

To form a network, we need connect all these individual neurons like a structure. There are many ways to build the network, feed-forward network and recurrent network are two network topologies very used. The first one showed in Figure 2.4, is a network with component clearly separated: an input layer, an output layer and one or more hidden layers also called processing layers. All connections are directed to the following layer and the internal states are just the weights themselves. The second network design the neurons have extra connections adding to a classic feed-forward network, they can be a direct recurrence when the neurons are connected to themselves, indirect recurrence is a connection to previous layers and a lateral recurrence exist when neurons have connections with another ones at the same layer. These connections influence the neurons itself and their influence depends of the kind of recurrent design. This mean that the network is a dynamic system, the fact that this network feeds its outputs back in its own inputs permit a short-term memory, a model more seem to a brain and by the way, more difficult to understand and build [10] [1].

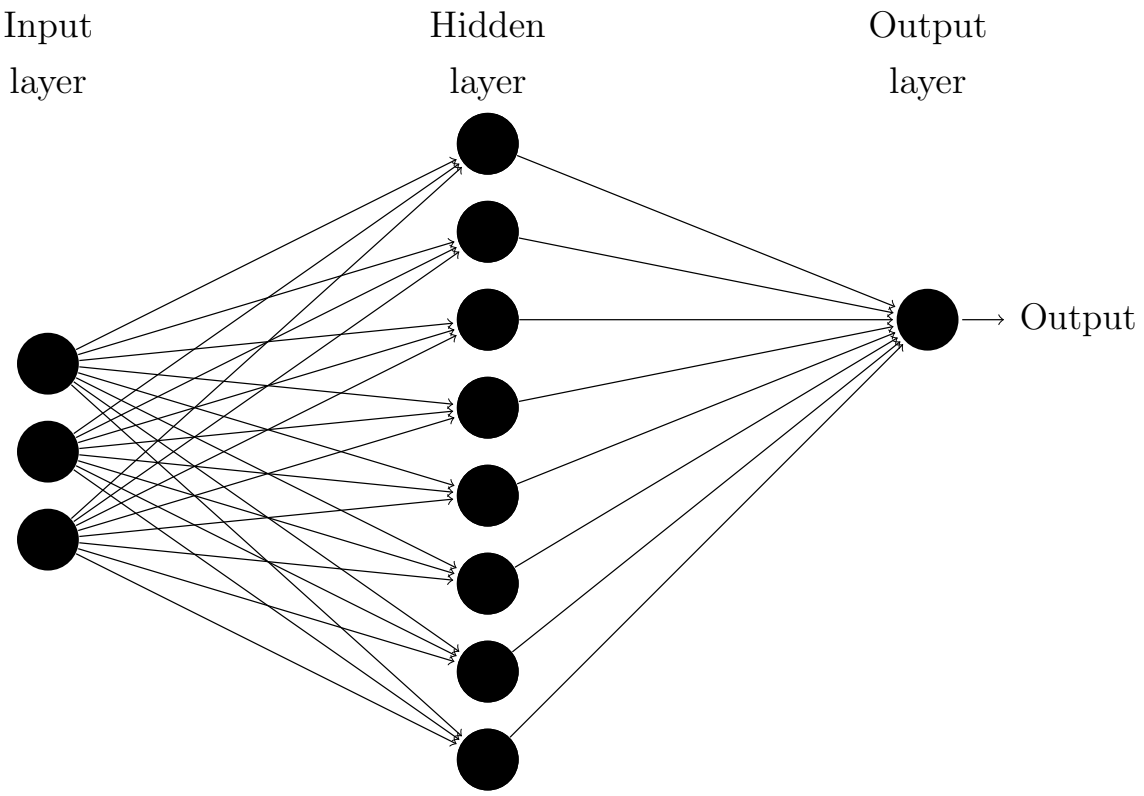

Fig 2.4: Feed Forward Network

2.4 Deep Learning

The artificial intelligence usually solves problems intellectually challenge for the humans relatively easy, but the task the human performs easy or solve intuitively are very difficult for the computers. The problems like speech recognition or applications to find faces in images are challenge task. For these problems we must allow computers gather knowledge from the experience, avoiding specify all the knowledge that the system needs. It means build a complicated task using more simpler concepts [3].

DL has great advancement in many fields of AI, like Natural Language Processing, Speech Recognition, Computer Vision, Biomedical Applications and so on [11]

These concepts form a deep graph, including many layers, the reason for that this approach is known how Deep Learning [3].

Multilayer Perceptron (MLP) or feedforward deep layer is the most typical example of deep learning approach. MLP consists in an input layer, that contains the input data for the model, next we have the hidden layers, a variable number of neurons and layers, and finally an output layer as is showed in the Figure **??**. That depth (hidden layers) allows

to learn a multi-step program.

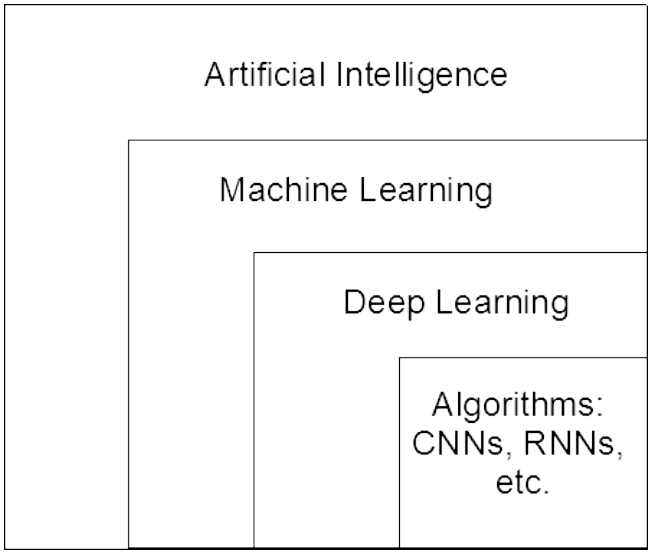

Fig 2.5: Artificial Intelligence Landscape

Figure 2.5 shows the AI landscape, when DL is a kind of ML, which is used for many approaches of AI.

There are many algorithms of Deep Learning, like CNN used commonly when is working with images and RNN to process sequences of data essentially.

2.4.1 Convolutional Neural Networks

CNN is applied to problems with a grid-like topology, a clear example of this type of data are the images, other are the regular time intervals. CNN formally is composed with a mathematical operation called convolution in at least one layer, instead of a general matrix operation [3]. CNN is very used in recognition, as a powerful visual model, to extract features in a hierarchy of concepts. [12].

Historically this type of network was some the firsts in solve important commercial problems [3] like the handwritten zip code recognition.

2.4.2 Recurrent Neural Networks

RNN is a kind of networks to processing sequential data. They are very powerful because they store efficiently information about the past. They are specialized in process a sequence of values [3], mapping all the previous input to each output. This allows the memory of old inputs can persist and influence in the next network output [13].

RNN are used for resolving many types of problems, [11] proposes a method to improving the use of RNN in classification of images. In other hand, RNN is also used to detection and diagnosis of a chemical process that show results with excellent performance [14].

2.5 Natural Language Processing

2.6 Text Similarity

2.6.1 Lexical Similarity

2.6.2 Semantic Similarity

Chapter 3

Methodology

3.1 Description

3.2 Formalization

3.3 Data Processing Pipeline

Chapter 4

Similarity Analysis

4.1 Deep Learning Framework

4.2 Model Description

4.3 Model Architecture

Chapter 5

Experiments

5.1 Hardware

5.2 Software

5.3 Experimental Results

Chapter 6

Conclusion and Future Work

In summary, more text here

Bibliography

[1] Stuart J. Russell and Peter Norvig. *Artificial intelligence: a modern approach*. Pearson, 2010.

[2] Aoife D'Arcy John D. Kelleher, Brian Mac Namee. *Fundamentals of Machine Learning for Predictive Data Analytics*. The MIT Press, 2015.

[3] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[4] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[5] Nils J. Nilson. *Introduction to Machine Learning*. Stanford University, 1998.

[6] Kimberly Nevala. *The Machine Learning Primer*. SAS Institute Inc., 2016.

[7] Daniel Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. Universal sentence encoder, 2018.

[8] A. S. Halibas, A. S. Shaffi, and M. A. K. V. Mohamed. Application of text classification and clustering of twitter data for business analytics. In *2018 Majan International Conference (MIC)*, pages 1–7, March 2018.

[9] Kevin Gurney. *An Introduction to Neural Networks*. Taylor and Francis e-Library, 2004.

[10] David Kriesel. *A Brief Introduction to Neural Networks*. dkriesel, 2005.

[11] B. Chandra and R. K. Sharma. On improving recurrent neural network for image classification. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 1904–1907, May 2017.

[12] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, June 2015.

[13] Alex Graves. Supervised sequence labelling with recurrent neural networks, 2010.

[14] G. M. Xavier and J. M. de Seixas. Fault detection and diagnosis in a chemical process using long short-term memory recurrent neural network. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, July 2018.

Appendices

Appendix A

GitHub Repositories

The GitHub repositories of the big data and machine learning deamon are avaliable upon request at danny.villanueva1@upr.edu. The following sections contain the links.

A.1 Big Data Platform

https://github.com/THSUPRM/bigdata/tree/master/python

A.1.1 Machine Learning Platform

https://github.com/THSUPRM/bigdata/tree/master/DetectDiseaseTHS/ths