INSTITUTO TECNOLÓGICO DE AERONÁUTICA

Heládio Sampaio Lopes

NATURAL LANGUAGE PROCESSING FOR TREND FORECASTING

Final Paper
2020

Course of Computer Engineering

CDU 621.38

Heládio Sampaio Lopes

NATURAL LANGUAGE PROCESSING FOR TREND FORECASTING

Advisor

Prof. Dr. Filipe Alves Neto Verri (ITA)

COMPUTER ENGINEERING

São José dos Campos
Instituto Tecnológico de Aeronáutica

2020

Cataloging-in Publication Data
Documentation and Information Division

Lopes, Heládio Sampaio
Natural Language Processing for Trend Forecasting / Heládio Sampaio Lopes.
São José dos Campos, 2020.
23f.

Final paper (Undergraduation study) – Course of Computer Engineering– Instituto Tecnológico de Aeronáutica, 2020. Advisor: Prof. Dr. Filipe Alves Neto Verri.

1. Natural Language Processing. 2. Deep Learning. 3. Machine Learning. I. Instituto Tecnológico de Aeronáutica. II. Title.

BIBLIOGRAPHIC REFERENCE

LOPES, Heládio Sampaio. **Natural Language Processing for Trend Forecasting**. 2020. 23f. Final paper (Undergraduation study) – Instituto Tecnológico de Aeronáutica, São José dos Campos.

CESSION OF RIGHTS

Heládio Sampaio Lopes
H8A St., 113
12228-460 – São José dos Campos–SP

NATURAL LANGUAGE PROCESSING FOR TREND FORECASTING

This publication was accepted like Final Work of Undergraduation Study

---

Heládio Sampaio Lopes

Author

Filipe Alves Neto Verri (ITA)

Advisor

Inaldo Capistrano Costa
Course Coordinator of Computer Engineering

São José dos Campos: JUNE 19, 2020.

Acknowledgments

Thank you

"That's All Folks"
— Looney Tunes,

Resumo

Resumo

Abstract

Abstract

List of Figures

FIGURE 3.1 – Stemming process for connect variations (VIJAYARANI *et al.*, 2015). . 17
FIGURE 3.2 – Bag of Words example. 19

List of Tables

List of Abbreviations and Acronyms

BoW	Bag of Words
NLP	Natural Language Processing
ML	Machine Learning
TF-IDF	Term Frequency Inverse Document Frequency

List of Symbols

Contents

1 Introduction 15

1.1 Motivation 15

1.2 Objective 15

1.3 Organization of this work 15

2 Background and Literature Review 16

2.1 Definition 16

2.2 Applications 16

3 Natural Language Processing 17

3.1 Text Processing Techniques 17

3.1.1 Normalization 17

3.1.2 Tokenization 18

3.1.3 Bag of Words 18

3.1.4 TF-IDF 19

3.2 Word Embedding 19

3.3 Topic Modeling 20

3.4 Topic Classification 20

4 Deep Learning 21

4.1 Neuron 21

4.2 Perceptron and Activation Functions 21

4.3 Loss Functions 21

4.4 Optimization 21

5 The Proposal 22
5.1 Hypothesis 22
5.2 Objective 22
5.3 Research method 22
5.4 Schedule 22
Bibliography 23

1 Introduction

1.1 Motivation

Every kind of expression, verbal or in writing, brings us a lot of information to be interpreted. Whether the topic is chosen, the tone used or the choice of words, everything can be interpreted, and then generate some useful information. Over the years, more and more knowledge is generated and we humans are not able to process such an amount of information. Natural language processing emerges as a technology capable of assisting us in this hard task.

(KHURANA *et al.*, 2017) defines natural language processing, abbreviated by NLP, as a branch of artificial intelligence capable of making computers understand and extract information from human language. NLP can perform a lot of tasks, such as identifying different topics for a set of documents, classifying texts on predefined subjects, and beyond that extract the sentiment to know what people are saying about something.

1.2 Objective

Curious about the fast world's evolution, this works aims to explore and compare a several Natural Language Processing techniques to model the topic's evolution over time. With this in mind, evaluate te ability of those models to make predictions about future trends.

1.3 Organization of this work

The remaining of this work is organized as follows: Chapter 2 will describe the general views of the concepts on which this work is based, (enfatizando no problema abbordado qu ainda não defini) Chapter 3 will cover the Natural Processing Language theory user over this work. Chapter 4 will cover the Deep Learning concepts. Finally, Chapter 5 will bring the proposal with the objectives and next steps for the work's development.

2 Background and Literature Review

2.1 Definition

2.2 Applications

3 Natural Language Processing

3.1 Text Processing Techniques

The key task to several machine learning problems consist in make a good data processing before applying any model. A clean data set can allow a model to increase its performance in the learning process, making a better identification in the patterns present in the variables. Hence, in the next sections, it will be discussed a few techniques to clear the text and prepare it for ML algorithms.

3.1.1 Normalization

There is no right way to normalize text, this process has it is really important to put all text in the same level. A normalization process has a series of steps to be followed sequentially, all of then can be seen as 4 big tasks: stemming, lemmatization, stop words removal and everything else.

1. Stemming: Is the process of reduce inflected words to a primitive form, the stem. This method is able to remove the word's affixes to capture its base meaning, and still reducing the number of variations to save memory space. Figure 3.1 shows how some inflections for "connect" can be converted to its root form.

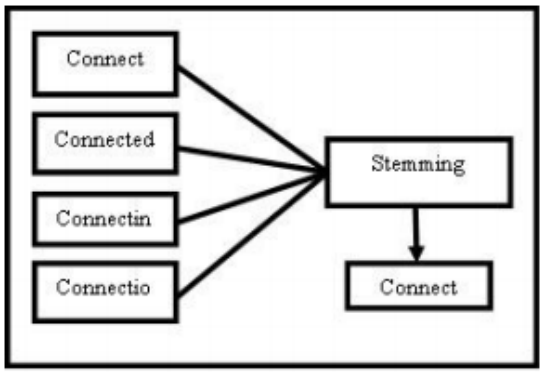


FIGURE 3.1 – Stemming process for connect variations (VIJAYARANI *et al.*, 2015).

2. Lemmatization: similar to stemming, this step also reduce words to some primitive form, but with a little improvement. Lemmatization can returns the words to his dictionary form, based on its part of speech context. So it is possible to discriminate words with the same spelling but different meanings depending on the context.

3. Remove stop words: Many words can occurs a several times in a document without add any meaningful information, such as *the*, *is*, *at*, *which*, and *on*. Their high frequency can be seen as an obstacle to perform good results on NLP models, (KANNAN; GURUSAMY, 2014).

   There are some types to remove stop words, most of then based on evaluating the frequency of words in text, for more information see (**??**). But the classic and easier method is based on using a pre-compiled list of know words and removing then from text.

4. Everything else: Differently from the previous steps, the last one doesn't need any grammar rules or even a frequency analysis, it's purely text manipulation. It involves set all character to lowercase; remove numbers or convert then to word form; remove punctuation; expand contractions; convert special characters to ASCII form; and any other conversion needed.

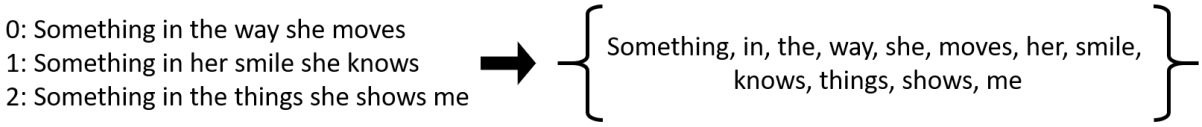
3.1.2 Tokenization

Once the data is normalized, we need to know how to represent it. The tokenization process consists in splitting longer strings into meaningful small pieces called tokens. The most common way to tokenize a text is chunking it the into words, ie, given a piece of text the tokenize process will return a list of words.

3.1.3 Bag of Words

The machine learning algorithms take numerical features as input, hence, it will bee necessary to represent the text in numerical form. With the Bag of Words model we can represent in matrix form a set of documents.

With the tokenization output we will have the lists representations for all documents in the data set. These lists can be interpreted as vectors over the vector space of all unique tokens, also called by vocabulary. So, for a given sentence, we mark how many times its words appears in the list indexes where each entry corresponds to a word in the vocabulary. The Figure 3.2 show a simple example of how three sentences can be represented with BoW model.



	Something	In	The	Way	She	Moves	Her	Smile	Knows	Things	Shows	Me
0	1	1	1	1	1	1	0	0	0	0	0	0
1	1	1	0	0	1	0	1	1	1	0	0	0
2	1	1	1	0	1	0	0	0	0	1	1	1

FIGURE 3.2 – Bag of Words example.

3.1.4 TF-IDF

Term Frequency Inverse Document Frequency, TF-IDF for short, it is applied to a BoW to determine the relative frequency for words in a specific document when compared to the inverse proportion of that word over all documents in the collection. So, it can be determined how important are the words in a specific document.

From BoW, for the i^{th} vocabulary's word in the j^{th} document, its TF-IDF weight, $w_{i,j}$, can be calculated with Equation 3.1.

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right) \tag{3.1}$$

Where, the term frequency, $tf_{i,j}$, is how many time i^{th} word appears in the j^{th} document. The document frequency, df_i, is the number of documents in which th i_{th} vocabulary words is present. And, finally, N is the size of the document collection, with a large number of documents this term can explodes, so the logarithmic function is applied to dampen this effect.

3.2 Word Embedding

The vectorization methods such as BoW and TF-IDF can be very useful, but they can not represent the words context. This means that the same words used in different contexts have the same representation, just as different words used with the same meaning are represented differently. Besides that, an one-hot encoding method, like BoW, presents a very sparse representation with high dimensionality.

The Word Embedding is a technique to represent words in vectors capable of capture the words context in a document. It is also able to smooth the high dimensionality effect by using much more compact vector to represent the words.

3.3 Topic Modeling

3.4 Topic Classification

4 Deep Learning

4.1 Neuron

4.2 Perceptron and Activation Functions

4.3 Loss Functions

4.4 Optimization

5 The Proposal

5.1 Hypothesis

5.2 Objective

5.3 Research method

5.4 Schedule

Bibliography

KANNAN, S.; GURUSAMY, V. Preprocessing techniques for text mining. **International Journal of Computer Science & Communication Networks**, v. 5, n. 1, p. 7–16, 2014.

KHURANA, D.; KOLI, A.; KHATTER, K.; SINGH, S. Natural language processing: State of the art, current trends and challenges. 08 2017.

VIJAYARANI, S.; ILAMATHI, M. J.; NITHYA, M. Preprocessing techniques for text mining-an overview. **International Journal of Computer Science & Communication Networks**, v. 5, n. 1, p. 7–16, 2015.

FOLHA DE REGISTRO DO DOCUMENTO

1. CLASSIFICAÇÃO/TIPO	2. DATA	3. DOCUMENTO N°	4. N° DE PÁGINAS
TC	June 19th, 2020	DCTA/ITA/DM-018/2015	23

5. TÍTULO E SUBTÍTULO:
Natural Language Processing for Trend Forecasting

6. AUTOR(ES):
Heládio Sampaio Lopes

7. INSTITUIÇÃO(ÕES)/ÓRGÃO(S) INTERNO(S)/DIVISÃO(ÕES):
Aeronautics Institute of Technology – ITA

8. PALAVRAS-CHAVE SUGERIDAS PELO AUTOR:
Natural Language Processing; Deep Learning; Machine Learning

9. PALAVRAS-CHAVE RESULTANTES DE INDEXAÇÃO:
Natural Language Processing; Deep Learning; Machine Learning

10. APRESENTAÇÃO: (**X**) **Nacional** ( ) **Internacional**
ITA, São José dos Campos, 2020. Trabalho de Graduação. 23 páginas.

11. RESUMO:
Resumo

12. GRAU DE SIGILO:
(**X**) **OSTENSIVO** ( ) **RESERVADO** ( ) **SECRETO**