INSTITUTO FEDERAL DO ESPÍRITO SANTO
CURSO SUPERIOR DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO

HARÃ HEIQUE DOS SANTOS

SIMILARIDADES DE ESTILOS LITERÁRIOS BASEADAS EM APRENDIZADO PROFUNDO

Serra
2019

HARÃ HEIQUE DOS SANTOS

SIMILARIDADES DE ESTILOS LITERÁRIOS BASEADAS EM APRENDIZADO PROFUNDO

Trabalho de Conclusão de Curso apresentado à Coordenadoria do Curso de Bacharelado em Sistemas de Informação do Instituto Federal do Espírito Santo, Campus Serra, como requisito parcial para a obtenção do título de Bacharel em Sistemas de Informação.

Orientador: Prof. Dr. Fidelis Zanetti de Castro

Serra
2019

RESUMO

Neste trabalho de conclusão de curso é proposto o estudo sobre a identificação de autoria de textos literários baseadas em medidas de similaridade com aprendizado profundo de máquina de uma Rede Neural Convolutiva Siamesa. Serão utilizadas três tipos de medidas de similaridade cardinal conhecidas na área e através de experimentos computacionais na Rede Neural Siamesa será apresentado um estudo análitico da perfomance dos modelos adotados especialmente em quesito de acurácia. A linguagem de programação para manipulação da Rede Neural Siamesa será Python devido sua notoriedade e difusão na área científica e facilidade de manipulação.

Palavras-chave: redes neurais, medidas de similaridade, Siamesa, textos literários, identificação, autoria.

SUMÁRIO

1 INTRODUÇÃO

Nos dias atuais é comum utilização de ferramentas tecnológicas que possuem a capacidade de interpretar a linguagem natural humana, tais como: *chatbots*, tradutores de documentos, corretores automáticos em *smartphones*, editores de textos e afins, mecanismos de buscas por um assunto específico em páginas *webs* entre outros. Esta variedade de ferramentas que utilizamos no nosso dia a dia só é possível devido aos avanços científicos da área de pesquisa da Inteligência Artificial denominada ***Processamento da Linguagem Natural*** (PLN).

Segundo a cientista (LIDDY, 2001), Universidade de Siracusa, na Itália, o Processamento de Linguagem Natural consiste em um conjunto de "teorias motivadas por uma série de técnicas computacionais para análise e representação de textos decorrentes da linguagem natural. Essas técnicas são utilizadas com o objetivo de processar linguagens humanas para diversas aplicações". Nesse sentido o PLN possibilita que computadores analisem, entendam, manipulem e interpretem a linguagem natural humana, estabelecendo assim uma espécie um modus operandi que permite a comunicação entre máquina e ser humano.

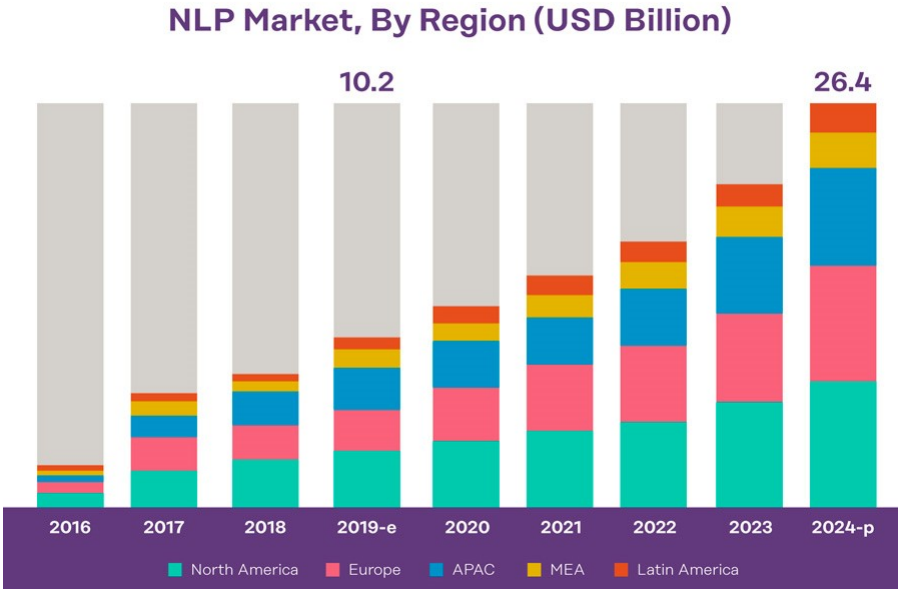
Figura 1 – Aplicações do Processamento de Linguagem Natural.

Fonte: Elaborado pelo autor (2020).

Ao longo dos anos a geração de dados não estruturados foi crescendo significativamente, no cenário em que a tendência é que cerca de 80% dos dados gerados por empresas são não estruturados, os quais são provenientes principalmente de mídias sociais e conversas com representantes de atendimento (Borges, 2018).

Decorrente a isto é notável que as corporações estão recorrendo a esta subárea da ciência da computação através de aplicações, no intuito de extrair informações e transformá-las em conhecimento, consequentemente aumentando sua competitividade no mercado. Na figura 2 é demonstrado o crescimento estimado do mercado na área por regiões continentais entre os anos de 2016 e 2025.

Figura 2 – Crescimento do PLN por região.



Fonte: Natural Processing Market by Markets and Markets 2019 (2019).

Note que as regiões com maiores concentrações de países desenvolvidos possuem uma taxa de crescimento maior do que as regiões com maior número de países subdesenvolvidos. Este fato é decorrente de os países desenvolvidos terem uma melhor situação financeira e investismentos focados nas áreas tecnológicas e científicas, consequentemente aumento gradativamente a vantagem competitiva.

Baseado neste contexto a similaridade de textos *(text similarity)* é um estudo na área de Processamento de Linguagem Natural e Mineração de Textos, objetivando identificar o quão próximos são os textos com intuito de rastrear tópicos, classificar e resumir textos, agrupar documentos, detectar plágios (Sieg, 2018), visto que um dos pontos de enfoque do trabalho é na análise e classificação de textos literários.

A utilização de ferramentas deste ramo para detecção de plágio acaba sendo relevante, pois elas acabam sendo fundamentais para originalidade e autoria dos textos, assegurando assim a integridade e propriedade intelectual produzida pelo autor que o escreveu. Outro ponto em que a detecção de plágio auxilia é no desenvolvimento do pensamento crítico, principalmente em áreas acadêmicas, pois evita que acadêmicos pratiquem o ato.

Existem diversas áreas em que podem ocorrer o plágio, dentre elas estão: música, obras, fotografias, trabalhos acadêmicos e afins. Entretanto este trabalho tem como foco obras literárias, ou seja, na comparação e classificação de textos, a fim de auxiliar na identificação de autoria.

1.1 METODOLOGIA DO TRABALHO

Neste trabalho de conclusão de curso é realizado o estudo sistemático da *Rede Neural Convulucional Siamesa* para performar na tarefa de identificação de similaridades dos textos com base em medidas de similaridade cardinal. Os recursos utilizados para implementação do projeto baseado na metodologia investigativa são:

- Livros sobre redes neurais, principalmente a do tipo Profunda e Siamesa, e medidas de similaridade;
- Artigos científicos similares que utilizam redes neurais, especialmente a Convolutiva *Siamesa*, para tarefas de identificação;
- Estudo do funcionamento da *Rede Neural Siamesa* baseado em uma linguagem de programação para sua manipulação.

A pesquisa possui caracterização teórica-aplicada, dado que além dos desenvolvimentos matemática aplicados no projeto, devido às medidas de similaridade, será utilizado a *Rede Neural Siamesa* que recebe diferentes medidas similaridades e baseando-se nela é treinada para um conjunto de dados, no caso os textos literários de diferentes autores. Após isso é feito o processo de investigação na comparação entre textos literários distintos com o que foi treinado na rede a fim de identificar as similaridades.

A análise dos dados será quantitativa e qualitativa, pois mostrará os dados analisados de forma numérica, onde no caso deste trabalho serão os valores de acurácia, precisão e afins, referentes ao índice de similaridade entre duas instâncias diferentes.

A linguagem de programação utilizada para manipulação dos dados e da rede neural escolhida é *Python*, devido ser uma linguagem de alto nível e de simples manuseio, além de conter várias bibliotecas de suporte voltadas para área de ciência dos dados. Já em relação

a base de dados é utilizado textos literários provenientes de autores da língua inglesa, que estão disponíveis online.

1.2 OBJETIVOS

1.2.1 Objetivo Geral

Analisar a performance de uma rede neural convolutiva siamesa para classificação de textos literários usando medidas de similaridade.

1.2.2 Objetivos Específicos

Os objetivos específicos identificados para se atingir o objetivo geral proposto são:

- Organizar uma base de dados composta por textos literários de quatro escritores da língua inglesa;
- Converter os dados da base de dados para matrizes numéricas de números reais usando a técnica word2vec;
- Implementar uma rede neural convolutiva siamesa usando três medidas de similaridade baseadas em cardinalidade;
- Analisar os resultados de classificação texto/autor obtidos pela rede siamesa para cada medida de similaridade usada.

1.3 ESTRUTURA DO TRABALHO

O restante deste trabalho está dividido em quatro capítulos. O Capítulo 2 contém a fundamentação teórica. Nele, são apresentados os conceitos relativos a processamento de linguagem natural, medidas de similaridades e revisamos os modelos das redes neurais usadas para análise dos dados. O Capítulo 3 fornece os detalhes da captura e normalização dos dados para alimentar a rede neural utilizada, além de explicar detalhes da implementação do código realizado em *Python*. No Capítulo 4, são realizadas as análises dos resultados obtidos nos experimentos computacionais. No último capítulo, são descritas as considerações finais e as possibilidades de trabalhos futuros.