PLANO DE TRABALHO

1. IDENTIFICAÇÃO DA AÇÃO

Título da ação: Identificação Automática de Classes Gramaticais em Textos Provenientes de Comunicadores *Online*.

Tipo de ação: Projeto de Pesquisa, Desenvolvimento Tecnológico e Inovação.

Área de conhecimento: Engenharia/Tecnologia.

Temática: Tecnologia e Produção.

Linha: Tecnologia da Informação.

2. COORDENAÇÃO DA AÇÃO JUNTO AO CEFET-MG

Coordenador: Prof. Daniel Hasan Dalip.
SIAPE: 2312571 - **Telefone**: (31) 99983-9046 - **E-mail**: hasan@cefetmg.br
Setor de lotação: Departamento de Computação.
Chefe do Setor de lotação: Kécia Aline Marques Ferreira.

3. CARACTERIZAÇÃO DA AÇÃO

3.1 Objetivos

O **objetivo geral** deste projeto consiste na implementação de um método de identificação automática de classes gramaticais em português em conversas de usuários com *ChatBots*[1] por meio de técnicas de Aprendizado de Máquina (ramo da Inteligência Artificial). Nesse projeto, algumas classes gramaticais poderão ser mais discriminadas, por exemplo: verbos de posse (ex: comprar, vender, possuir), verbos de movimento (ex: ir, vir, andar, correr, dirigir, nadar, caminhar) e de cognição (ex: analisar, pensar, julgar, duvidar). Além disso, substantivos poderão ser discriminados em entidades[2], por exemplo: lugar, empresa, artefato/objeto, comida, valor, temporal (ex: data, hora, dia da semana), comunicação (ex: email, telefone).

Essas classes gramaticais/entidades serão definidas na etapa de criação da base de dados.

Para que o objetivo geral proposto seja alcançado, serão perseguidos os seguintes **objetivos específicos**:

- Estudo de métodos estado-da-arte para identificação de classes gramaticais e reconhecimento de entidades;
- A partir de bases de dados de conversas fornecidas pela Take, criar uma amostra de dados manualmente rotulada de acordo com sua entidade/classe gramatical. Essa rotulação manual é importante para o método aprender padrões nos dados que serão usados para a identificação automática de classes gramaticais/entidades.

[1] *ChatBots* são *softwares* capazes de se comunicar, por meio de linguagem natural, com usuários. Tais *softwares* podem ser usados para o atendimento ao público e integrados à comunicadores onlines como WhatsApp e o Messenger.
[2] Entidade é qualquer elemento do mundo que possui sua existência independente de qualquer outro elemento

Além disso, por meio da rotulação manual, é possível realizar a avaliação do método proposto.
- Avaliar, por meio das bases rotuladas, qual é o melhor método para identificação automática de classes gramaticais/entidades;
- Durante a avaliação, realizar possíveis ajustes nos métodos e no pré-processamento para cada contexto fornecido;
- Avaliar o impacto do uso de entidades e classes gramaticais em tarefas como: análise de sentimento e sumarização de texto.

No final do projeto, espera-se como contribuição: (1) a rotulação manual de milhares de termos de acordo com sua classe gramatical/entidade a partir de bases fornecidas pela Take; um estudo para entender qual é a (2) melhor abordagem para preprocessamento e rotulação automática de tais entidades e classes gramaticais; (3) e quais são as melhores técnicas de reconhecimento de classes gramaticais para este contexto estudado.

3.2 Importância para a Sociedade e para os Partícipes Envolvidos

__ Associação com o ensino	X Associação com a pesquisa
__Inovação Tecnológica	__ Associação com política pública

O presente projeto apresenta associação com atividades de pesquisa e inovação tecnológica.

A **Take** é uma empresa que oferece soluções integradas para comunicação entre empresas e seus clientes por meio de comunicadores *online* (como o WhatsApp, Messenger). Tais soluções utilizam meios textuais para o cliente se comunicar com a empresa, inclusive, por meio de ChatBots. Um dos principais objetivos de um *ChatBot* é interpretar a necessidade do usuário - expressa em linguagem natural (ex. português) - para, assim, responder ao usuário da melhor forma possível.

Neste contexto, a identificação de classes gramaticais e entidades no texto auxilia na organização da informação textual dessa interação. Tal organização contribui para um melhor entendimento da necessidade do cliente dessas empresas. Além disso, com essa informação, a Take poderá desenvolver soluções para entender melhor se um cliente está satisfeito ou não (por meio da análise de sentimentos) além de utilizar tais termos e a rotulação automática para uma possível sumarização textual.

As empresas atendidas pela Take abrangem uma diversa gama de serviços oferecidos para a sociedade: telefonia, bancos, comércio e eventos. Dessa forma, esse projeto também poderá gerar um impacto positivo **para o público em geral,** pois, tais soluções poderão aprimorar o atendimento ao público de tais empresas.

Além disso, esse projeto irá propiciar a oportunidade para que um aluno do CEFET-MG desenvolva uma pesquisa aplicada para a necessidade de uma empresa. **Para o aluno**, será uma experiência enriquecedora tanto do ponto de vista acadêmico quanto pelo ponto de vista profissional. No ponto de vista da **comunidade do CEFET-MG**, este projeto estreita os laços entre o CEFET-MG e a Take aumentando, consequentemente, as

possibilidades de futuras parcerias entre tais instituições tais como: oportunidade de estágio para alunos, palestras, oficinas e mini-cursos.

3.3 Público Alvo

O público alvo deste projeto é constituído por clientes de empresas de diversos ramos. Considera-se que o presente projeto tem o potencial de impactar, em média, 2.917.919 milhões de usuários por mês (direta e indiretamente) na plataforma com os resultados obtidos no projeto[3].

3.4 Locais de Realização

Este projeto será executado no CEFET-MG, no Laboratório de Pesquisas Interdisciplinares em Informação Multimídia (Piim-Lab)[4], durante todo o período do projeto, de 8 às 22 horas. O coordenador do presente projeto é pesquisador desse laboratório. Além disso, eventuais reuniões poderão ser executadas tanto nas dependências do Departamento de Computação do CEFET-MG quanto nas dependências da Take.

3.5 Metodologia

A **metodologia** para execução deste projeto pode ser descrita pelo seguinte conjunto de ações:

1. **Analisar os requisitos** da Take. Em específico: quais são as entidades/classes gramaticais relevantes a serem extraídas do texto? Quais são as principais necessidades e aplicações da abordagem a ser desenvolvida? ;

2. **Fazer o levantamento** de métodos e ferramentas disponíveis para rotulação automática de classes gramaticais/entidades em texto;

3. **Realizar uma análise da base de dados** com o objetivo de entender suas peculiaridades e também formas de preprocessamento a serem realizadas. Tal base de dados é composta interações entre o público e os ChatBots;

3. **Realização do preprocessamento** da base com o objetivo de obter uma melhor performance do método automático de identificação de entidades/classes gramaticais;

4. **Criação de uma amostra manualmente rotulada** de acordo com as classes gramaticais e entidades preestabelecidas, considerando uma base fornecida pela Take;

4. **Avaliação e análise** do erro do método realizado, para eventuais ajustes;

5. **Realização de eventuais ajustes** no método e no processamento para melhoria dos resultados. São exemplos de melhorias de processamento uma maior padronização

[3] Levando em consideração que a plataforma de ChatBots da Take atendeu, em média, 2.917.919 de usuários únicos por mês desde julho de 2018

[4] http://www.piim-lab.cefetmg.br/

do vocabulário (ex.: transformar abreviaturas como “vc” para “você”). Considerando o método, pode-se fazer combinações de métodos e parametrizações e, além disso, utilizar o dado não rotulado por meio de técnicas de aprendizado semi-supervisionado;

6. **Avaliar o uso desta ferramenta** para algumas aplicações como análise de sentimentos e sumarização de texto;

7. Depositar eventuais pedidos de proteção intelectual, oriundos do desenvolvimento da plataforma proposta, junto ao Instituto Nacional de Propriedade Industrial (INPI) e, eventualmente, no exterior.

A metodologia acima descrita será respaldada pela realização de reuniões técnicas com os membros da equipe envolvida no projeto, as quais serão de periodicidade quinzenal, com o objetivo de avaliar o desenvolvimento das tarefas e apresentar os progressos nos trabalhos realizados. Durante essas reuniões serão realizados a apresentação dos resultados (Seção 3.7) de acordo com a fase de execução do projeto, que pode ser prevista no cronograma (Seção 5) .

O **CEFET-MG** (http://www.cefetmg.br) e a **Take** (https://take.net/) integram diferentes saberes e competências, os quais são essenciais para lidar com os desafios inerentes à natureza interdisciplinar deste projeto.

O **CEFET-MG**, é reconhecido como centro de excelência na formação tecnológica de profissionais que atuam, em especial, no setor produtivo, na pesquisa aplicada e no magistério do ensino tecnológico. Assim sendo, exerce um papel que vai além da formação profissional e assume o diálogo crítico e construtivo com a sociedade, para geração de conhecimentos e de novas tecnologias.

A **Take** é uma empresa que oferece soluções integradas com o objetivo de aprimorar a comunicação das empresas com seus clientes. Para isso, a Take possui profissionais qualificados e uma plataforma, de desenvolvimento próprio, para integrar diversas tecnologias para comunicação com clientes por meio de ChatBots e, também, comunicadores online.

3.6 Direitos Intelectuais, Autorais e Patentes sobre Produtos, Bens e Processos

Conforme já informado, a Take já possui uma plataforma Web de sua exclusiva propriedade (i.e. um software disponibilizado via Web) em que é possível, além de outras funcionalidades, a criação e uso de *Chatbots.* Tal plataforma licenciada para utilização de seus clientes (empresas) que criam ambientes para atendimento ao público. Este projeto irá fazer um estudo para descobrir, dentre as abordagens de preprocessamento e técnicas de reconhecimento de entidades e classes gramaticais existentes, qual é a melhor abordagem no contexto da Take.

Como esse estudo irá usar abordagens já existentes criadas e publicadas por terceiros com direito de utilização até para fins comerciais, a Take poderá usar/adaptar qualquer uma dessas abordagens e comercializá-las sem qualquer ônus ou restrição.

Apesar de já existirem soluções (i.e. softwares e métodos) para identificação de classes gramaticais e entidades, tais soluções nem sempre estão adaptadas para um contexto específico e nem para um idioma. Além disso, não é garantido que tais soluções funcionem em qualquer situação, principalmente no contexto da Take em que os textos são provenientes de comunicadores online possuindo, assim, uma linguagem informal e uma grande quantidade de erros gramaticais e ortográficos. Assim, o propósito deste trabalho é investigar quais métodos/ferramentas já existentes é melhor para a linguagem informal oriundo de tais comunicadores, em específico, utilizando os dados e contextos da Take. Para isso, será verificado qual é o melhor método e a melhor forma de tratamento do texto para esta tarefa. Assim, a princípio não será gerado um produto inovador.

Eventualmente, iremos usar softwares que implementam métodos de terceiros para a rotulação automática e Processamento de Linguagem Natural. Durante a procura de tais software para serem usados, as licenças serão analisadas e utilizaremos apenas softwares que possuem licença para uso irrestrito, inclusive para fins comerciais. Como, por exemplo, a licença Apache 2.0[5] e a MIT[6].

Para o depósito dos eventuais pedidos de proteção intelectual associados ao projeto, será solicitado o apoio do Núcleo de Inovação Tecnológica do CEFET-MG. A titularidade da propriedade intelectual oriunda deste projeto e a participação nos resultados de sua exploração comercial deverão estar devidamente previstas em instrumento jurídico específico, assegurando aos partícipes o direito à exploração, ao licenciamento e à transferência de tecnologia, observado o disposto nos §4º e §5º do art. 6º da Lei nº 10.973/2004.

Sem prejuízo do exposto, o projeto em questão não afetará nem representará qualquer interferência no direito da Take de exercer a titularidade exclusiva da plataforma de sua propriedade.

3.7 Resultados Esperados

Metas Quantitativas	Indicadores
Rotular termos de, pelo menos, 5 mil mensagens de acordo com sua classe gramatical/entidade	Número de frases e termos rotulados. Uma apresentação à Take demonstrando características da base rotulada.
Um estudo de qual é a melhor abordagem para preprocessamento do texto levando em consideração o vocabulário da base de dados fornecida pela Take com o objetivo de remover ruídos e padronizar o vocabulário e, com isso, melhorar a performance do método.	Verificação se, ao usar um determinado preprocessamento houve uma melhoria na performance na tarefa. Tal performance será mensurada por meio de técnicas comumente usadas em Recuperação de Informação como o F1-Score. Este indicador pode ser analisado considerando apenas entidades/classes gramaticais que a Take julgar mais relevantes (ex.: substantivos de lugar, tempo e pessoa). Uma apresentação à Take demonstrando quais foram as técnicas de preprocessamento adotadas e o impacto no vocabulário.

[5] https://www.apache.org/licenses/LICENSE-2.0
[6] https://opensource.org/licenses/MIT

	Uma documentação da abordagem criada.
Um estudo de quais são as melhores formas de preprocessar o texto - para essa tarefa - e também quais são as melhores técnicas de reconhecimento de classes gramaticais para este domínio	Verificação de qual método é possível obter um melhor F1-Score. Um relatório com os resultados de cada método/preprocessamento e discussão dos mesmos. Uma apresentação à Take demonstrando quais foram os resultados obtidos e qual foi a melhor forma de preprocessar o texto.
Um estudo de qual é a melhor abordagem para identificação de classes gramaticais/entidades que utiliza o melhor método (dentre os testados, e com a melhor adaptação possível) e o melhor preprocessamento.	A implementação desta abordagem e a documentação da mesma.
Orientar, ao menos, 1 (um) bolsista de extensão oriundo da Eng. de Computação do CEFET-MG.	Número de alunos orientados
Publicar 1 (um) artigo em periódico	Número de artigos publicados

3.8 Formas de Divulgação e Publicação de Resultados

O presente projeto prevê, conforme descrito neste plano de trabalho, o estudo da melhor abordagem para a identificação de entidades e classes gramaticais em textos. Assim sendo, após a realização dos eventuais pedidos de proteção intelectual, as base de dados utilizadas relativos ao projeto serão mantidos em sigilo. Contudo, pretende-se apresentar o estudo e suas aplicações em revistas científicas, feiras e eventos que contribuam para divulgá-la, bem como utilizar as abordagens estudadas em parcerias com representantes de organizações públicas e privadas, visando-se atender as demandas da sociedade.

4. RECURSOS ENVOLVIDOS NA AÇÃO

4.1 Recursos Humanos

Nome	SIAPE ou Matrícula ou CPF	Cargo e Instituição	Função	Disponibilidade (Horas)		
				Semanal	2019	2020
Daniel Hasan Dalip	2312571	Professor CEFET-MG	Coordenador	2	32	64
Alexandre Augusto Aguiar Gomes	201722040084	Estudante de Eng. de Computação CEFET-MG	Desenvolvedor de Software	20	320	640
Karina Tiemi Kato	040.437.599-59	Participante Externo Take	Analista de Inteligência Artificial	2	32	64

4.2 Recursos Financeiros

O presente projeto prevê o aporte de recursos financeiros no total de R$ 37.200,00. Além dos custos operacionais do CEFET-MG e da Fundação CEFET-MG. Esse aporte será feito pela Take.

4.3 Instalações e Equipamentos

Será usado uma estação de trabalho do Piim-Lab para o estudante de Engenharia de Computação a ser contratado como estagiário neste projeto. Além disso, eventualmente, será usado o espaço do Departamento de Computação e da Take para reuniões e apresentações.

5. CRONOGRAMA

Período de Realização	Início	Mês	Ano	Fim	Mês	Ano	Parâmetro de tempo (X)					
		setembro	2019		agosto	2020	**Mensal**	X	**Bimestral**		**Trimestral**	

Ações	Meses											
	1	2	3	4	5	6	7	8	9	10	11	12
1 - Análise de requisitos	X											
2 - Levantamento de métodos/ferramentas para a rotulação automática das entidades e classe gramaticais	X											
3 - Caracterização da base com o objetivo de verificar quais técnicas de preprocessamento serão usadas	X											
4 - Fornecimento da base de dados e adicionar a mesma possíveis dados complementares e remoção dados irrelevantes para tarefa	X	X										
5 - Preprocessamento: Criação de uma abordagem para tratar o texto		X	X									
6 - Documentação do preprocessamento			X									
7 - Criação de testes unitários para garantir a corretude e reprodutibilidade do preprocessamento. Ou seja, garantir que o programa, com uma determinada entrada e um conjunto de parâmetros definidos, sempre apresente o mesmo resultado (e que o resultado esteja correto).			X									
8 - Determinação das entidades a serem consideradas			X									
9 - Rotulação manual das entidades/classes gramaticais a serem usadas			X									
10 - Resultado preliminar da rotulação automática			X	X								
11 - Análise de erros do resultado preliminar				X	X							
12 - Adaptação da abordagem proposta considerando a análise de erros. Exemplo de ações a serem tomadas: - Melhoria no preprocessamento - Rotular manualmente mais mensagens - Uso/adaptação de base já rotulada existente (ex: MacMorpho[7]) - Utilização de outros métodos - Corrigir rotulações manuais incorretas - Adaptação do método de rotulação automática					X	X	X					
13 - Verificar a melhor forma de usar mensagens/termos não rotulados							X	X	X			
14 - Avaliação da abordagem para rotução automatica utilizada									X	X		
15 - Criação dos testes unitários para garantir a corretude e reprodutibilidade da rotulação automática										X		
16 - Documentação da abordagem										X	X	
17 - Escrita do Relatório			X							X	X	X
18 - Esclarecimentos e validações de cada resultado apresentado pelo CEFET.	X	X	X	X	X	X	X	X	X	X	X	X

6. DADOS DO PARTÍCIPE EXTERNO

6.1 Identificação do Partícipe Externo

Razão Social: Curupira S.A.

[7] http://nilc.icmc.usp.br/macmorpho/

CNPJ: 04.413.729/0001-40
Endereço: Rua Sergipe, 1440 – 9º andar (sala 109) e 10º andar – Savassi, Belo Horizonte, Minas Gerais, CEP: 30.130-174
Sítio Web: http://take.net/
Representante legal 1: Sérgio Cruz Passos
Cargo: Analista de Sistemas
CI: 7.752.887 expedida pela SSP/MG **CPF**: 040.454.566-19
Telefone: +55 31 3349-6271 **E-mail**: sergio@take.net
Representante legal 2: Daniel Rodrigues Costa
Cargo: Administrador de Empresas
CI: 19.604 expedida pelo CRA/MG **CPF**: 838.362.336-49
Telefone: +55 31 99618-2089 **E-mail**: daniel@take.net

6.2 Participação de Fundação de Apoio

[X]: **Sim**.	[ ]: **Não**.

Função:	Interveniente.

7. OBRIGAÇÕES DOS PARTÍCIPES

- **Caberá ao Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG)**:
  - Autorizar os servidores a ele vinculados, listados na Seção 4.1, a participarem do presente projeto, de acordo com a carga horária prevista neste plano de trabalho;
  - Disponibilizar as instalações, equipamentos e materiais de consumo a ele atribuídos nas Seções 4.3 deste plano de trabalho para a realização das ações planejadas;
  - Participar das execuções das ações listadas na Seção 5, garantindo o alcance dos resultados sob sua responsabilidade.

- **Caberá à Take**:
  - Alocar ao presente projeto os profissionais a ela vinculados, listados na Seção 4.1, de acordo com a carga horária prevista neste plano de trabalho;
  - Participar das execuções das ações 4 e 18 listadas na Seção 5, garantindo o alcance dos resultados sob sua responsabilidade.

Data: 09/07/2019

Coordenador da Ação - Daniel Hasan Dalip

PLANO DE TRABALHO Nº 240/2019 - DECOM (11.56.03)

(Nº do Protocolo: NÃO PROTOCOLADO)

(Assinado digitalmente em 09/07/2019 10:15)
DANIEL HASAN DALIP
PROFESSOR ENS BASICO TECN TECNOLOGICO
DECOM (11.56.03)
Matrícula: ###125#1

Visualize o documento original em https://sig.cefetmg.br/documentos/ informando seu número: ***240****, ano:* ***2019****, tipo:* ***PLANO DE TRABALHO****, data de emissão:* ***09/07/2019*** *e o código de verificação:* ***340d8ae3b1***