****

**UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO**

PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO

COORDENADORIA GERAL DE PROGRAMAS ESPECIAIS

**Relatório das Atividades de Aluno de Iniciação Científica**

PIBIC/CNPq/UFRPE

PIC/UFRPE

1. **IDENTIFICAÇÃO**

**ALUNO**: DIEGO DE FREITAS BEZERRA

**CURSO**: BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

**PROGRAMA**: PIC

**ORIENTADOR**: ROBSON SANTOS DE OLIVEIRA

**DEPARTAMENTO/ÁREA**: UAG/BCC

1. **TÍTULO DO PROJETO**: Identificação do *Self* nas Redes Sociais Virtuais: Ferramentas de Análise Semióticas e Computacionais.
2. **RESUMO DO RELATÓRIO**

Este projeto de Pesquisa de Iniciação Científica objetiva desenvolver uma ferramenta computacional de identificação do *Self* nas Redes Sociais Virtuais (RSV) a partir do modelo de investigação MEVERSI (Modelo Etnográfico Virtual de Estudo das Representações do *Self* na Internet), permitindo gerar informações baseadas em dados obtidos dos vários ambientes virtuais de autorreferencialidade (RSV) observando-se o comportamento de mudança e permanência do *Self*. Inicialmente, foi realizada a revisão de literatura referente aos conceitos abordados pelo MEVERSI (etnografia nos ambientes virtuais e noções do *Self*), definindo uma metodologia para descrever perfis, grupos, ambientes e a interação que ocorre entre eles. Em seguida, foram avaliadas as atividades que ocorrem no Facebook e Twitter visando identificar, com base no MEVERSI, os diversos espaços de interação (chat, *timeline*, grupos, campos de comentários, etc.) e quais os dados que melhor representam o *Self* nesses ambientes (RSV). Nessa primeira etapa de desenvolvimento, o projeto focou na coleta de dados a partir do Twitter com a biblioteca Java *Twitter4J.* Posteriormente, foi realizado o pré-processamento dos dados coletados utilizando Processamento de Linguagem Natural (PLN) por meio do etiquetador morfossintático LX-Tagger. Por fim, os dados foram indexados e minerados com base nas características do *Self* apresentadas por George Mead (1934) e Hubert Hermans (1987, 2001), permitindo avaliar padrões de cada perfil.

1. **INTRODUÇÃO**
2. **OBJETIVOS**
   1. **Geral**: Integrar num único modelo de investigação do *Self* duas ferramentas para análise da pessoa nas Redes Sociais Virtuais (RSV): a Mineração de Dados na Web (Análise de Sentimentos, Mineração Textual) e o MEVERSI (Psicologia Cognitiva, o Modelo Etnográfico Virtual de Estudo das Representações do *Self* na Internet).
   2. **Específicos**: Realizar pesquisa etnográfica no Facebook e no Twitter; Estudar e testar o Modelo Etnográfico Virtual de Estudo das Representações do *Self* na Internet; Desenvolver uma ferramenta computacional de mineração de dados web baseada nas RSV: sistema de busca, coleta e classificação de dados.
3. **METODOLOGIA**

Métodos utilizados na metodologia deste projeto de pesquisa:

**Etnografia virtual:** é a metodologia de pesquisa que busca a coleta de dados através do ambiente virtual, utilizando os diversos materiais disponíveis na rede, para descrever pessoas, grupos e ambientes (HINE, 2000; OLIVEIRA, 2011).

**MEVERSI:** é um Modelo Etnográfico Virtual de Estudo das Representações do *Self* na Internet que permite gerar informações baseadas em dados obtidos dos vários ambientes virtuais de autorreferencialidade (RSV) observando-se o comportamento de mudança e permanência do *Self* (OLIVEIRA, 2012).

**Estabelecer Requisitos:** Um requisito consiste em uma declaração sobre um produto pretendido que especifica o que ele deveria fazer ou como deveria operar (ROGERS; SHARP; PREECE, 2011).

**Prototipagem:** Um protótipo é modelo em escala menor, pode ser qualquer coisa, desde um *Storyboard* de papel a uma parte complexa de um software. Ele possibilita que os *Stakeholders* interajam com um produto imaginado podendo visualizar e avaliar de forma prévia algumas de suas funcionalidades (ROGERS; SHARP; PREECE, 2011).

**Mineração de dados:** Envolve a execução de algoritmos para a busca de padrões relevantes à dados coletados e pré-processados, podendo executá-los diversas vezes, até que sejam alcançados os resultados esperados (CARVALHO FILHO, 2014).

**Análise de dados:** Envolve a identificação de diversos padrões entre dados minerados, alguns não relevantes, outros interessantes para o domínio do problema (CARVALHO FILHO, 2014).

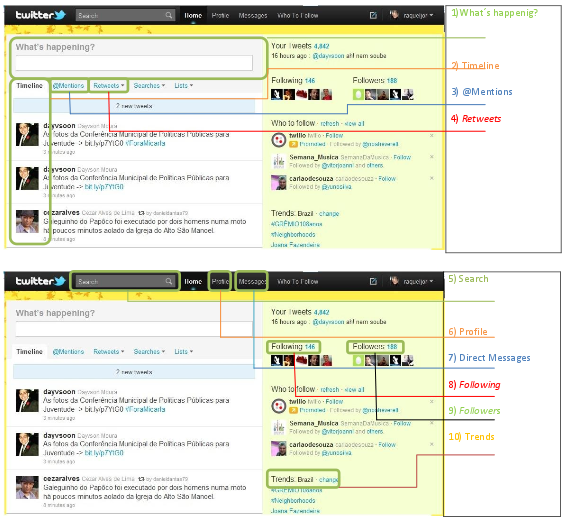
A metodologia desse projeto envolveu uma análise etnográfica e a definição do Twitter como ambiente inicial para a coleta de dados, observando que o conteúdo textual da *timeline* e os dados básicos dos perfis dos usuários (nome, descrição e localização) apresentavam as características que definiam o *Self*.

1. **RESULTADOS E DISCUSSÃO**

Foi realizado no mês de agosto a revisão da literatura sobre a aplicação da etnografia nos ambientes virtuais: Etnografia Virtual (HINE, 2000), Netnografia (BRAGA, 2010) e Webetnografia (POLIVANOV, 2013). No mês de setembro foi realizada a revisão de literatura sobre noções do *Self* baseadas em William James (1890/2007), George Mead (1934) e Hubert Hermans (1987, 2001) apresentadas por Oliveira (2012). No mês de outubro foi realizado o estudo sobre o MEVERSI (Modelo Etnográfico Virtual de Estudo das Representações do *Self* na Internet) (OLIVEIRA, 2012) e definido o Twitter como RSV a ser avaliada nesta primeira etapa. No mês de novembro foram levantados modelos de mineração de dados (DIAS & PACHECO, 2005), seleção de perfis, o levantamento e a identificação de requisitos, funcionais e não-funcionais do sistema a ser desenvolvido. No mês dezembro foi selecionada a biblioteca Java *Twitter4J* como ferramenta de coleta dos dados no Twitter, realizado o estudo da documentação da biblioteca e dado início ao processo de prototipagem (definição de layout e estrutura do sistema). No mês de janeiro foram definidas e iniciadas as etapas de mineração de dados: coleta, pré-processamento, indexação, mineração de dados e análise. Para o pré-processamento dos dados coletados foi definido o uso do Processamento de Linguagem Natural (PLN) por meio do etiquetador morfossintático LX-Tagger. No mês de fevereiro estão sendo implementados os requisitos funcionais apresentados no documento de requisitos do sistema utilizando a linguagem de programação Java, a IDE (*Integrated Development Environment****)*** Eclipse, a ferramenta de automação de compilação Apache Maven, o etiquetador morfossintático LX-Tagger e a biblioteca Java Twitter4J para a Twitter API.

**7.1. Etnografia virtual no Twitter**

O twitter é uma rede social e um serviço de microblog, desenvolvido por Jack Dorsey na cidade de São Francisco - EUA e posto está disponível desde agosto de 2006. A interação entre os usuários tem como ponto de partida a os *tweets*, mensagens que contém no máximo 140 caracteres, característica que privilegia a rapidez na comunicação por meio de dispositivos móveis, e que podem ser compostas por textos, imagens, vídeos e *emoticons[[1]](#footnote-1)* . Silva (2012), descreveu o microblog da seguinte maneira:

**Figura **1**: Etnografia virtual Twitter (SILVA, R.S., 2012, p.14-15)**

* **What's happening?** - A interação entre os usuários da rede tem como ponto de partida esse campo, que tem o limite de postagem de 140 caracteres;
* **Timeline** – Apresenta os *posts* dos usuários seguidos. Ela também é apresentada no Profile do próprio usuário;
* **@Mentions** – *Posts* que mencionam nomes de usuários;
* **Retweets** – As postagens podem ser disseminadas por meio do *Retweet*, onde a mensagem de determinado autor pode ser reproduzida por seus seguidores;
* **Search** – Campo de busca do Twitter;
* **Profile** – Apresenta todas as postagens do próprio usuário. São observados ainda campos com dados pessoais: nome de usuário, descrição, idade e localização e a própria *timeline*, registrando as mensagens publicadas ou *retweets*;
* **Direct Messages** – Permite a comunicação por meio de mensagens privadas entre os usuários;
* **Following** – Usuários seguidos;
* **Followers** – Usuários seguidores;
* **Trends** – Assuntos mais populares no microblog em um determinado espaço/tempo, identificados por uma #*hashtag*

**7.2. Coleta de dados do Twitter**

Para a coleta de dados, foi definida uma seleção de quatro perfis, de maneira aleatória: dois perfis de usuários que representassem figuras anônimas e dois perfis de usuários que representassem figuras públicas. Os usuários que representavam as figuras anônimas foram chamadas de e e os que representavam figuras públicas foram chamadas de e . Utilizando o Twitter4J, foram extraídos os últimos 1.000 *tweets* de suas respectivas *timelines* e armazenados em uma base de dados.

**7.3. Pré-processamento dos dados**

O pré-processamento dos dados coletados foi realizado por meio do Processamento de Linguagem Natural, utilizando a ferramenta LX-Tagger. A ferramenta, inicialmente, realiza a segmentação do *tweet* em unidades linguísticas como palavras, pontuações, números, alfanuméricos, etc., gerando tokens Este processo é conhecido como tokenização. Após esse processo, é feita atribuição de categorias a porções do texto, ou seja, para cada palavra presente no *token*, é atribuída uma classificação morfossintática. Abaixo é apresentado um exemplo do processo do *tweet*.

**Tweet**: Melzinha em um estado supremo de fofura \*-\* https://t.co/iTQMcwJAp9/

**Token**: [Melzinha] [em] [um] [estado] [supremo] [de] [fofura] [\*-\*] [https://t.co/iTQMcwJAp9/]

**Etiquetagem morfossintática**: Melzinha/**PNM** em/**PREP** um/**UM** estado/**CN** supremo/**ADJ** de/**PREP** fofura/**CN** \*-\*/**PNT** https://t.co/iTQMcwJAp9/**PNT**

Na etiquetagem morfossintática, cada palavra recebe **/TERMO** que define sua classificação. Os termos adicionados acima significam:

* **PNM** – Nomes próprios;
* **PREP** – Preposições;
* **UM** – Artigos indefinidos;
* **CN** – Nomes comuns;
* **ADJ** – Adjetivos;
* **PNT** – Caracteres de pontuação.

O LX-Tagger é capaz de atribuir 47 termos na classificação morfossintática para o idioma português brasileiro. Termos desconhecidos podem ser classificados com o termo PNT, assim como ocorreu com o *hiperlink* presente no *tweet* utilizado no exemplo.

Todos os *tweets* foram submetidos ao processo de etiquetagem morfossintática. Após esse processo, foram armazenados em uma base de dados para a realização do processo mineração.

**7.4. Mineração de dados**

A etapa de mineração dos dados foi realizada com base na seleção de características do *Self*, apresentado por Hermans (2001), onde a expressão de qualidades e sensações próprias (‘eu como pai’, ‘eu como responsável’, ‘eu como brincalhão’, etc.) e a contemplação de pessoas e objetos que têm profunda e relevante relação com o indivíduo (‘meus filhos’, ‘meus pais’, ‘minhas amigas’, ‘meu carro’, ‘meu emprego’, etc.), estão estreitamente ligadas à definição do *Self*.

Abaixo são apresentados tais parâmetros que podem caracterizar o *Self*. Foram relacionadas ocorrências do MEU/MINHA expressas pelos usuários, bem como a recorrência dessas expressões no Twitter. O mesmo foi feito para as interações com as Hashtags e os usuários.

Visando melhor exibição dos dados, são apresentados apenas os 10 casos que mais ocorreram durante o processo, quando o número total for maior que 10.

.

Usuário:

|  |  |  |
| --- | --- | --- |
| **Meu/Minha: 49** | **Hashtags: 44** | **Usuários: 224** |
| minha vida: 4  minha irma: 4  meu celular: 3  minha casa: 3  minha mae: 2  meu pais: 2  minha cara: 2  meu tablet: 1  meu único: 1  meus planos: 1 | #oscars: 6  #cerimoniadeabertura: 6  #g1: 6  #gameofthrones: 5  #rio2016: 3  #dexter: 3  #ff: 2  #oscaromelete: 2  #rip: 1  #tbt: 1 | @amanda\_olb: 50  @raquelmorais: 27  @eteduardo: 27  @jornaloglobo: 24  @luscas: 15  @g1: 14  @jornalcrer: 12  @paulomatew: 10  @diimabr: 9  @kellisonj: 8 |

**Tabela 1: Dados obtidos para**

Na Tabela 1, para o usuário , foram obtidas 49 referências para termos Meu/Minha, interação com 44 hashtags e 224 usuários em seus últimos 1.000 tweets.

Usuário:

|  |  |  |
| --- | --- | --- |
| **Meu/Minha: 44** | **Hashtags: 60** | **Usuários: 184** |
| minha vida: 2  meu celular: 2  meu aquario: 1  meu cartaoclique: 1  minha cama: 1  meu note: 1  meu coraçao: 1  meu arduino: 1  minha sugestao: 1  meu pc-nao: 1 | #masterchefbr: 16  #xboxclips: 4  #pokemongo: 3  #oscars: 2  #gameofthrones: 2  #injustice: 2  #justiça: 2  #vr: 1  #169: 1  #brasil: 1 | @youtube: 260  @botanicgnome: 60  @mr\_\_\_\_\_\_\_d: 53  @reidosmafagafos: 46  @kawaii\_bagarai: 28  @nerdologia: 21  @omelete: 21  @pipocaenanquim: 18  @beatriz\_cnd: 12  @dimirsu: 11 |

**Tabela 2: Dados obtidos para**

Na Tabela 2, para o usuário , foram obtidas 44 referências para termos Meu/Minha, interação com 60 hashtags e com 184 usuários em seus últimos 1.000 tweets.

Usuário:

|  |  |  |
| --- | --- | --- |
| **Meu/Minha: 7** | **Hashtags: 6** | **Usuários: 567** |
| meu contrato: 3  meu nome: 3  minha epoca: 1  meus videos: 1  minhas fotos: 1  meu começou: 1  meu celular: 1 | #fitacrepe: 1  #carnavalsemassédio: 1  #educadosempre: 1  #ressaca: 1  #transtorno: 1  #carnavalacaboue: 1 | @shewillbecarol: 13  @jornalhoje: 10  @ifavsdemetria: 5  @fitzlovers: 5  @ney\_divo: 5  @racostalopes: 5  @walktheness: 5  @isaiasalvez: 5  @raizacunha: 5  @silvanasodre: 4 |

**Tabela 3: Dados obtidos para**

Na Tabela 3, para o usuário , foram obtidas 7 referências para termos Meu/Minha, interação com 6 hashtags e com 567 usuários em seus últimos 1.000 tweets.

Usuário:

|  |  |  |
| --- | --- | --- |
| **Meu/Minha: 23** | **Hashtags: 333** | **Usuários: 126** |
| meu amigo: 6  minha vida: 4  meu dia: 3  minha pagina: 3  meu snap: 3  meu querido: 3  meu irmao: 2  minha folia: 2  meu amor: 2  meu facebook: 2 | #caldeirão: 302  #latavelha: 18  #dandovalor: 18  #caldeirão…: 14  #umportodosetodosporum: 13  #caldeirao: 13  #ferias: 9  #fifthharmony: 9  #saltibum: 8  #timebrasil: 8 | @: 254  @angelicaksy: 9  @lucasranngel: 8  @…: 8  @marceloadnet: 8  @brumarquezine: 7  @luansantana: 6  @thiaguinhocomth: 6  @neymarjr: 4  @redeglobo: 4 |

**Tabela 4: Dados obtidos para**

Na Tabela 4, para o usuário , foram obtidas 23 referências para termos Meu/Minha, interação com 333 hashtags e com 126 usuários em seus últimos 1.000 tweets.

1. Caracteres especiais ou imagens que expressam sentimentos. [↑](#footnote-ref-1)