**UNIVERSIDADE FEDERAL DE PERNAMBUCO**

**CENTRO DE INFORMÁTICA**

**PÓS - GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO**

CLOVES ALVES DA ROCHA

FLÁVIO DA SILVA NEVES

MILTON VINICIUS MORAIS DE LIMA

**ANÁLISE DE SENTIMENTOS NO TWITTER**

Trabalho apresentado ao curso de Mestrado em Ciência da Computação como requisito parcial à avaliação da disciplina de **Recuperação Inteligente de Informação (IN1152),** ministrada pela professora **Flávia de Almeida Barros**.

**RECIFE – PE**

**DEZEMBRO -2015**

**RESUMO**

Neste projeto foi realizado uma análise de sentimentos de várias empresas multinacionais de serviços online (*Google, Apple, Microsoft e Twitter*) utilizando o *CoreLNP*. O método de pesquisa aplicado foi o experimento controlado, sendo organizados em dois experimentos e executados em paralelo, o primeiro experimento tem a sua base de dados rotulada, já o segundo não tem, sendo assim...

**SUMÁRIO**

1. **INTRODUÇÃO**
   1. **Contextualização**

Na atualidade as opiniões têm grande influência sobre o comportamento das pessoas e o mundo corporativo não foge desta afirmação. Várias decisões, por mais que sejam simples, são frequentemente baseadas em opiniões de pessoas próximas, por exemplo, a compra de um carro, qual roupa comprar, o melhor filme entre outros. Devido a alta concorrência no mundo dos negócios, muitas empresas estão se baseando em várias estratégias de negócios, dentre essas, apontamos para a opinião de seus clientes para seus produtos e serviços. A opinião dos clientes esta tão grande que várias empresas têm focado ações para obter esse tipo de informação. Embora este tipo de ação possa trazer resultados satisfatórios, envolve muito custo e seu retorno pode ser demorado muitas vezes. O tempo de resposta também e alta, devido ao longo período de coleta dessas informações, em sua maioria em seu estado bruto.

No mundo a *web* é o maior repositório de informações existentes nos dias atuais. Pessoas podem interagir todos os dias numa enorme quantidade de dados com conteúdos diversos.

**1.2. Motivação**

Com o grande volume de informações produzidas diariamente, implica a necessidade de métodos e ferramentas capazes de processar automaticamente essa demanda, não apenas publicações, mas também as opiniões e o que os usuários têm expressado. Como então filtrar essas informações já que ocorrem num fluxo constante? Como recuperar apenas o que se deseja? Como resumir com clareza os dados imensos que foram encontrados?

A mineração de sentimento ou analise subjetiva [20, 34, 23] é uma disciplina que une mineração de dados, linguística computacional, recuperação de informações, inteligência artificial entre outras. Ela é definida como uma disciplina computacional que envolvem opiniões, sentimentos, emoções, subjetividade, entre outros.

**1.3. Objetivo do Trabalho**

O objetivo principal deste é analisar os sentimentos com ênfase em empresas multinacionais de serviços online (*Google, Apple, Microsoft e Twitter*) utilizando o *CoreLNP,*  ferramenta desenvolvida na linguagem *Python*, com estudo de caso à partir do Twitter.

**1.4. Estrutura do Trabalho**

Para atender ao objetivo descrito na seção anterior, além desta Introdução o presente relatório é apresentado com a seguinte distribuição dos assuntos:

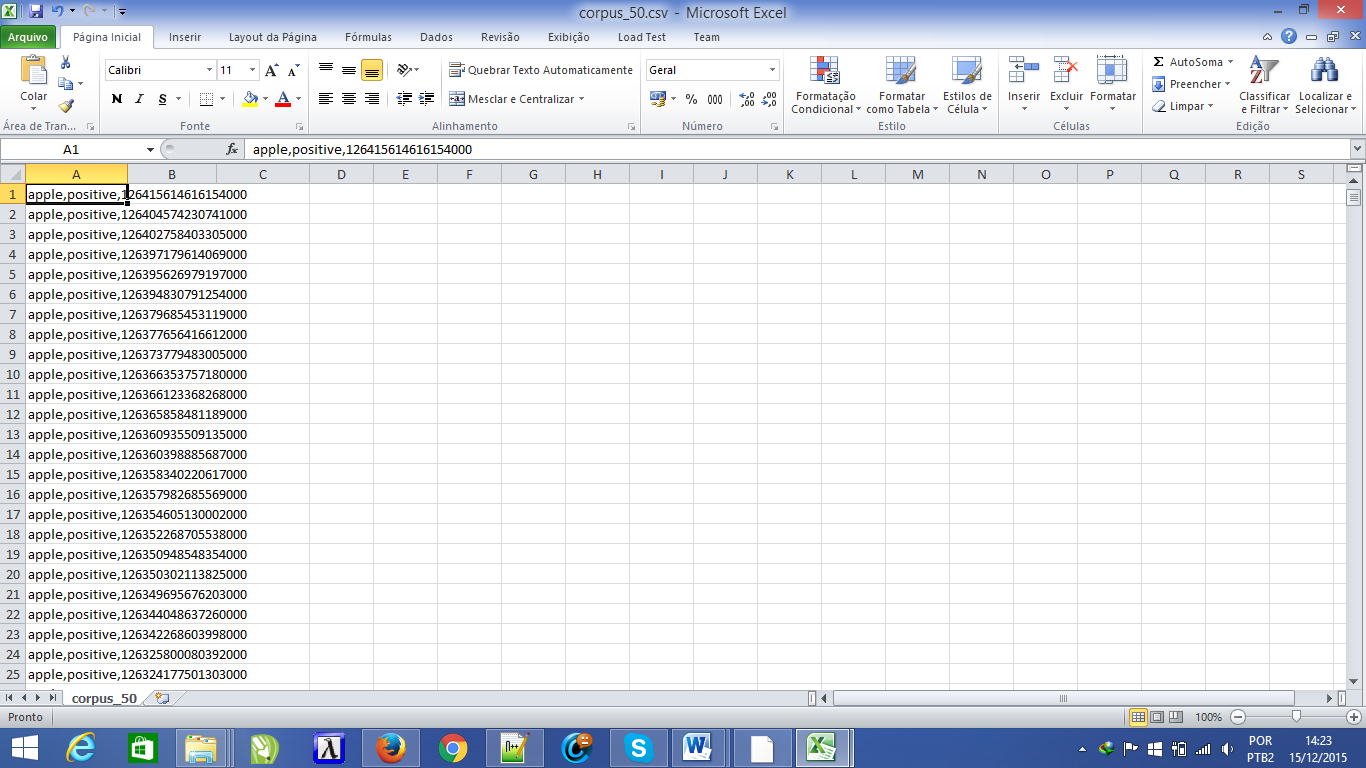
* O Capítulo 2 constituem as *Descrições do Sistema* apresentado e utilizado neste trabalho, tendo com premissas os objetivos e a base documentos utilizada.
* O Capítulo 3 descreve o *Protótipo* implementado, sua *Arquitetura do Sistema*, Representação dos documentos e o sistema em uso.
* O Capítulo 4 contém a Avaliação do Sistema e os resultados obtidos.
* O Capítulo 5 apresenta as conclusões acerca da validade do trabalho e possibilidades que ainda podem ser exploradas.

**2. DESCRIÇÃO DO SISTEMA**

**2.1 Objetivos e Funcionamento**

O sistema realiza a análise de sentimentos no Twitter buscando a opinião dos usuários sobre uma determinada empresa. Desta forma é possível que uma empresa possa identificar o nível de aceitação na rede social mencionada anteriormente. O funcionamento do sistema é descrito com os seguintes passos:

* É utilizado um corpus previamente rotulado com os tweets, este corpus serve para teste e treinamento (CoreNLP, 2015);
* O corpus é salvo em formato “csv”;
* Em seguida os dados são estruturados nas seguintes descrições: (“NomeEmpresa”,”Polaridade”,”Id”), respectivamente, conforme figura 1.



**Figura 1: Exemplo do *corpus* de entrada para o sistema**

**2.2 Base de Documentos Utilizada**

Como mencionado no Item anterior, foi utilizado um corpus já rotulado com 5.513 tweets, divididos nas polaridades: Positive, Negative, Neutral e Irrelevant, a respeito de 4 empresas , *Google*, *Apple*, *Twitter* e *Microsoft*. Este corpus é foi etiquetado no ano de 2011. Abaixo é descrito na tabela 1 o corpus rotulado com as empresas citadas.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Topic** | **# Positive** | **# Neutral** | **# Negative** | **# Irrelevant** | **Twitter Search Term** |
| Apple | 191 | 581 | 377 | 164 | @apple |
| Google | 218 | 604 | 61 | 498 | #google |
| Microsoft | 93 | 671 | 138 | 512 | #microsoft |
| Twitter | 68 | 647 | 78 | 611 | #twitter |

**Tabela 1: Corpus Rotulado com 5.513 tweets**

[**3.**](#h.2et92p0) **PROTÓTIPO**

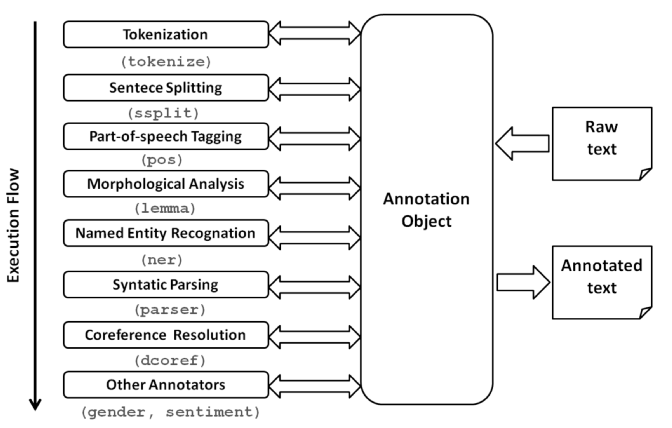
Como descrito na seção anterior exploramos a ferramenta Stanford CoreNLP[18]. O *CoreNLP* fornece um conjunto de linguagem natural, que integra a maior parte das etapas de processamento de linguagem natural, assim analisando a estrutura gramatical das sentenças, sistemas de resolução de correferências, análise de sentimento e ferramentas de aprendizagem de máquina. A ferramenta possui suporte para análise de textos em Inglês.

[**3.1. Arquitetura do Sistema**](#h.tyjcwt)

Segundo Josiane 2015[xx] O processamento de texto é feito através das seguintes etapas:

* O texto de entrada passa por um processo de anotação.
* O processo de anotação é representado por uma sequência de tokens, que em seguida são agrupados em sentenças. Os tokens são rotulados com suas partes do discurso, são gerados os lemas, é feito o reconhecimento das entidades (se são nomes de empresas, pessoas, lugares etc.) e é fornecida uma análise sintática completa, incluindo uma representação de dependências baseada em análise probabilística. Com base nestas informações, é possível fazer análise de sentimento aplicando um modelo composicional baseado em um classificador e implementar detecção de menções e resolução de correferências.
* Na saída é obtida uma anotação contendo todas as informações analisadas pelos anotadores, estruturadas em um arquivo XML.
* O texto de entrada passa por um processo de anotação.
* O processo de anotação é representado por uma sequência de tokens, que em seguida são agrupados em sentenças.
* Os tokens são rotulados com suas partes do discurso, são gerados os lemas,
* É feito o reconhecimento das entidades (se são nomes de empresas, pessoas, lugares etc.)
* É fornecida uma análise sintática completa, incluindo uma representação de dependências baseada em análise probabilística.
* Com base nestas informações, é possível fazer análise de sentimento aplicando um modelo composicional baseado em um classificador e implementar detecção de menções e resolução de correferências.
* Na saída é obtida uma anotação contendo todas as informações analisadas pelos anotadores, estruturadas em um arquivo XML.

Abaixo é mostrada a figura da *Arquitetura do Sistema* conforme *Stanford:*



**Figura 2 - Arquitetura do CoreNLP**

**Fonte: http://nlp.stanford.edu/software/corenlp.shtml#Abou**[**t**](http://nlp.stanford.edu/software/corenlp.shtml#About)

[**3.2. Representação dos Documentos**](#h.3dy6vkm)

Os documentos foram representados no formato csv e classificados em relação a um dos 4 diferentes tópicos. Cada entrada contém:

* Tweet ID
* Texto Tweet
* Tweet data de criação
* Tópico usado para sentimento
* Rótulo Sentiment: “positivo”, “neutro”, “negativo”, ou “irrelevante”.

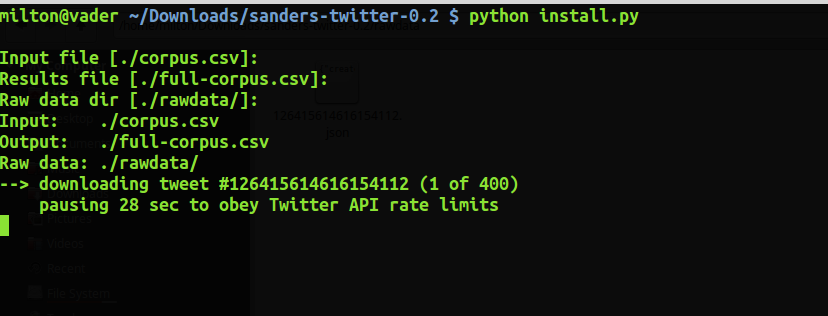
A quebra dos tópicos e os dados está representada na tabela 1.

Todos os dados do Twitter (tweets, datas de criação, tweet IDS) é coberto por Twitter [Termos de Serviço: Ref.<https://dev.twitter.com/terms/api-terms>]

[**3.3. Sistema em Uso**](#h.1t3h5sf)

Para uso do sistema utilizamos o terminal Linux para executar a aplicação escrita na linguagem *Python,* onde fizemos adaptações necessárias em seu código fonte em que o *CoreNLP* necessitava criar uma conexão via https com a API do Twitter, além disso, utilização dos módulos *NLTK*, para processamento de linguagem natural, *Json*, *csv*, *getpass*, *time*, *os*, *urllib* e o *Phyton* para o Twitter.

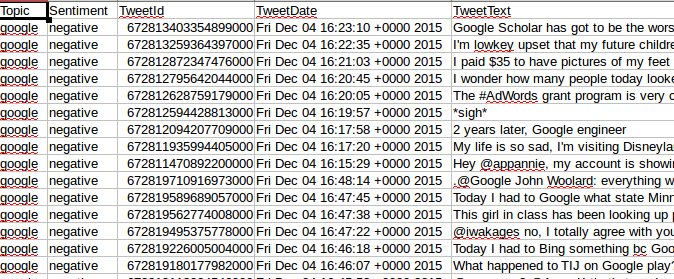
A seguir é visualizada o sistema em execução utilizando um terminal Linux para a aplicação em *Phyton*. A figura 3 representa o início da execução, a figura 4 o fim da execução e a figura 5 ilustra o corpus de saída com os resultados respectivamente.



**Figura 3 - Inicio da Análise de Sentimento**

# print_todos_final.png

**Figura 4 - Fim do processo da Análise de Sentimento**



**Figura 5 - Exemplo do *corpus* de saída criado pelo sistema.**

# 

**4. AVALIAÇÃO DO SISTEMA**

**4.1. Descrição dos Experimentos**

Foram feitos dois experimentos controlados para realização dos testes do sistema, foram utilizados dois corpus, um com 400 tweets retirados do corpus disponibilizado no site do *CoreNLP*, que será chamado de “Experimento 1” ele estava dividido da seguinte maneira: Google 100, Apple 100, Twitter 100 e Microsoft 100 divididos nas quatro polaridade igualmente. A tabela 2 apresenta um exemplo de documento de entrada.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Topic** | **# positive** | **# Neutral** | **# Negative** | **# Irrelevant** | **Twitter Search Term** |
| Apple | 25 | 25 | 25 | 25 | @apple |
| Google | 25 | 25 | 25 | 25 | #google |
| Microsoft | 25 | 25 | 25 | 25 | #microsoft |
| Twitter | 25 | 25 | 25 | 25 | #twitter |

**Tabela 2: Corpus Rotulado com 400 tweets**

No segundo teste que chamaremos de “experimento 2” montamos nosso próprio corpus, fazendo a classificação manualmente, com 103 tweets mencionando a empresa Google, divididos da seguinte maneira, 36 positivos, 34 negativos e 33 neutros, como está descrito na tabela 3.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Topic** | **#Positive** | **# Neutral** | **# Negative** | **Twitter Search Term** |
| Google | 36 | 34 | 33 | #google |

**Tabela 3: Corpus Rotulado Manualmente com 103 tweets**

**4.2. Descrição do Resultado dos Testes**

Após a execução do experimento 1, dos 400 tweets usados para treinamento, 18,25% não foram retornados. Foram retornados os seguinte dados:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Topic** | **#Positive** | **# Neutral** | **# Negative** | **# Irrelevant** | **Twitter Search Term** |
| Apple | 17 | 20 | 20 | 19 | @apple |
| Google | 22 | 23 | 22 | 21 | #google |
| Microsoft | 19 | 22 | 21 | 23 | #microsoft |
| Twitter | 18 | 22 | 14 | 13 | #twitter |

**Tabela 4: Resultados dos testes do experimento 1**

Após a execução do experimento 2, dos 103 tweets usados para treinamento, 6,18% não foram retornados. A seguir estão descritos os seguinte dados retornados:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Topic** | **#Positive** | **# Neutral** | **# Negative** | **Twitter Search Term** |
| Google | 34 | 32 | 31 | #google |

**Tabela 5: Resultados dos testes do experimento 2**

Depois da execução da aplicação foi feita a comparação dos dados retornados (resultados) com para avaliar a eficiência do algoritmo, então foi elaborada a matriz de confusão para o experimento 2.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | **Automático** | | |
| Classe 1 (+) | Classe 2 (-) | Classe 3 (0) |
| **Manual** | Classe 1 (+) | 24 | 7 | 3 |
| Classe 2 (-) | 6 | 19 | 7 |
| Classe 3 (0) | 5 | 8 | 18 |

**Tabela 6: Matriz Confusão para avaliar classificadores.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | **Automático** | | |
| Classe 1 (+) | Classe 2 (-) | Classe 3 (0) |
| **Manual** | Classe 1 (+) | 71% | 22,79% | 6,25% |
| Classe 2 (-) | 18,1% | 59,75% | 22,13% |
| Classe 3 (0) | 16,43% | 23,74% | 58,81% |

**Tabela 7: Precisão da análise.**

**5. CONCLUSÃO**

O corpus realmente foi projetado para treinamentos e testes Twitter sentimento algoritmos de análise. Na etapa de instalação devido a restrição no Twitter Termos de Serviço, os tweets reais não podem ser distribuídos com o corpus sentimento. Um script na linguagem Python foi incluído para baixar todos os tweets rotulados.

Devido às limitações na API do Twitter, o processo de download levou cerca de 43 horas, assim tivemos que reduzir o escopo do corpus como apresentado na tabela 2.

Com o escopo do corpus reduzido o processo levou cerca de 4 horas no experimento 1 (tabela 2), já no experimento 2 (tabela 3) cerca de 50 minutos.

Esses dois experimentos foram controlados de formas semelhantes, todavia vale ressaltar que no experimento 1 obtivemos os dados previamente rotulados, enquanto no experimento 2 os dados formam rotulados “manualmente” sendo necessário uma análise individual de cada texto.

**REFERÊNCIAS**

[]SILVA. Josiane. **Detecção de Opiniões e Análise de Polaridade em Documentos Financeiros com Múltiplas Entidades**.. MANAUS-AM, Março de 2015.

[] Recuperação de Informação - **Conceitos e Tecnologia das Máquinas de Busca** - 2ª Ed. 2013 - Ricardo Baeza-yates, Berthier Ribeiro-neto Solicitação da professora

[] Stanford CoreNLP – **a suite of core NLP tools**, Acesso no link: http://stanfordnlp.github.io/CoreNLP/

[34]Tsytsarau, M. and Palpanas, T. (2012). Survey on mining subjective data on the web.

Data Mining and Knowledge Discovery, 24(3):478–514.

[20] Liu, B. (2012). Sentiment analysis and opinion mining. Morgan & Claypool Publishers.

[23] Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. Foundations

and trends in information retrieval, 2(1-2):1–135.