ELC 2010
IX Encontro de Linguística de Corpus

MINERANDO TWEETS

Autores: Larissa Astrogildo de Freitas, Angélica Alves Fernandes, Ulisses Brisolara Corrêa

Emails: {larissaaf, angelicaalvesfernandes,ulisssesbcorrea}@gmail.com

Introdução

Tendo em vista que milhões de usuários interagem, se comunicam, criam, compartilham e organizam informações nos chamados softwares sociais (Orkut, Facebook, Youtube, Twitter e outros), fazem-se necessários trabalhos que buscam automatizar o processo de leitura e compreensão do que está sendo inserido nestes meios.

Segundo notícia apresentada em BBC [1] um estudo de curto prazo revela que no Twitter 40% das mensagens postadas são inúteis, porém, os outros 60% merecem nossa atenção e é o que tomamos como base para este trabalho.

O Twitter [3] é um micro blog que oferece uma base de dados, na forma de *tweets*, os quais são atualizações de status ou reflexões sobre notícias de destaque, cultura popular contendo no máximo 140 caracteres.

Na literatura poucas iniciativas utilizando o Twitter como Corpus da *Web* são encontradas, motivo: ser recente e apresentar pouca importância se comparado com outras fontes de dados.

No presente trabalho temos como objetivo geral analisar as mensagens enviadas sob o ponto de vista da frequência e do tempo de permanência de um determinado assunto.

Experimentos

■ Captura de dados da *Web* oriundos do Twitter, *tweets* da região de Porto Alegre (raio de 15Km), sobre o assunto Copa do Mundo, através da biblioteca Twitter4J [2].

Exemplos:

```
Sun Jun 20 15:30:34 GMT-03:00 2010 - vamooo Brasil!!!
Sun Jun 20 15:31:11 GMT-03:00 2010 - Costa do Marfim acaba de entrar na
briga com Paraguai e Eslovênia por uniforme mais bonito da copa.
Sun Jun 20 15:35:06 GMT-03:00 2010 - Jogo de Copa em pleno domingo não faz
sentido. Perdemos o Faustão e a folga durante a semana.
Sun Jun 20 16:56:33 GMT-03:00 2010 - Ai, será que o Elano quebrou a perna?
Tomara que não!
```

■ Limpeza de dados coletados, remoção de caracteres especiais e de *stopwords* [4], como por exemplo: @,#,?,!, de, em, que, a, o.

Exemplos:

```
Sun Jun 20 15:51:08 GMT-03:00 2010 - vamo Brasil
Sun Jun 20 16:56:33 GMT-03:00 2010 - Ai, Elano quebrou perna Tomara
```

■ Construção de bases de treino (2426 *tweets*) e de teste (4333 *tweets*).
■ Utilização do algoritmo de aprendizado de máquina Naïve Bayes.
■ Realização de uma análise linguística (semântica) [5] dos *tweets* verdadeiros positivos.

Exemplos:

```
'sun jun 20 15:33:10 gmt-03:00 2010 - promoção bolão do michel na copa meu
palpite é brasil 4 x 1 costa do marfim',true
'sun jun 20 15:38:15 gmt-03:00 2010 – a mãe tá lá na copa fazendo pipoca
pro jogo',true
```

■ Apresentação do gráfico de freqüência de *tweets* versus tempo (intervalo de 5 em 5 minutos) sobre o assunto Copa do Mundo.

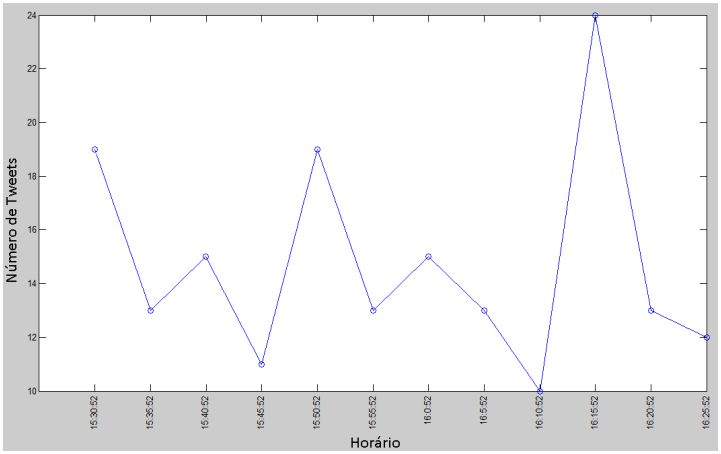

Figura 1: *Tweets* verdadeiros positivos.

A **Figura 1** apresenta os 178 *tweets* verdadeiros positivos, os quais foram coletados no dia 20 de junho de 2010, durante o jogo do Brasil contra a Costa do Marfim na Copa do Mundo.

É possível visualizar picos em determinados horários, como 15:30:52h e 16:15:52h que correspondem respectivamente ao início da partida e ao intervalo do jogo, o que já era esperado.

Com a análise lingüística (semântica) foi possível constatar que o classificador se comportou bem. Observamos também que alguns *tweets* estavam fora de contexto e que outros continham palavras ambíguas como copa (torneio de competição para disputa de uma taça) e copa (divisão adjacente à cozinha que pode ser usada para refeições) .

Considerações Finais

Em suma, através deste trabalho foi possível observar que inúmeras iniciativas podem ser realizadas utilizando o Twitter como Corpus.

Dentre os problemas enfrentados podemos elencar principalmente: a coleta dos dados, a construção da base de treinamento e da base de teste para ser utilizada como entrada no algoritmo de aprendizado de máquina Naïve Bayes.

Além disso, outro fator que merece ser mencionado é que os *tweets* são informais, inconsistentes em termos de linguagem e curtos, portanto, dificultando o processo de classificação.

Os resultados podem ser melhorados utilizando outras técnicas, em especial, na fase de pré-processamento, como remoção ou substituição de termos do "internetês" (vc por você, blz por beleza).

Pretendemos como trabalho futuro utilizar outros algoritmos de aprendizado de máquina como Máxima Entropia, Máquina de Vetor Suporte, Árvore de Decisão e abordar outras categorias como Saúde Pública, Governo Eletrônico e Educação.

Referências

[1] Twitter tweets are 40% 'babble'. Disponível em: http://news.bbc.co.uk/2/hi/technology/8204842.stm

[2] Twitter4J. Disponível em: http://twitter4j.org/en/index.html

[3] Twitter. Disponível em: http://twitter.com

[4] Lista de stopwords. Disponível em: http://mininigtext.blogspot.com/2008/11/listas-de-stopwords-stoplist-portugues.html

[5] CEGALLA, D.P.. Novíssima Gramática da Língua Portuguesa. 46ª. Edição, Companhia Editora Nacional, 2005, 693 p.