Proceedings of Symposium in Information and Human Language Technology. Natal, RN, Brazil, November 4–7, 2015.

7x1-PT: um *Corpus* extraído do Twitter para Análise de Sentimentos em Língua Portuguesa

Silvia M. W. Moraes, Isabel H. Manssour, Milene S. Silveira

Faculdade de Informática
Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS)
Caixa Postal 1429– 90619-900 – Porto Alegre – RS – Brazil

{silvia.moraes,isabel.manssour,milene.silveira}@pucrs.br

Abstract. *This paper describes the 7x1-PT corpus that contains a set of tweets, in Portuguese, posted during the match Germany vs Brazil at the FIFA World Cup 2014. We describe data collection, cleaning and organization, and also the current stage of the linguistic annotation of this corpus.*

Resumo. *Este artigo descreve o corpus 7x1-PT que contém um conjunto de tweets, em português, postados ao longo da partida da Alemanha com o Brasil durante a Copa do Mundo de 2014 da FIFA. Nós descrevemos como foi realizada a coleta, a limpeza e a organização, bem como comentamos o estágio atual de anotação linguística desse corpus.*

1. Introdução

A disponibilidade de recursos linguísticos e ferramentas computacionais consolidadas são fundamentais para realização de pesquisas que envolvam processamento da língua natural (PLN). Apesar dos esforços contínuos dos pesquisadores, a Língua Portuguesa ainda não dispõe de uma gama ampla e variada de tais recursos e ferramentas. Embora tenham ocorrido avanços nesse sentido, comparado a outras línguas, ainda há muito por fazer. Indo ao encontro dessa necessidade, construímos o *corpus* 7x1-PT, o qual poderá ser usado em pesquisas na área de Análise de Sentimentos[1] [Liu 2012]. O *corpus* 7x1-PT é um subconjunto da base de *tweets* WorldCupBrazil2014. Esta base é formada por 851.292 *tweets* coletados em vários idiomas (Português, Inglês e Espanhol) durante a copa do mundo de 2014 que ocorreu no Brasil. O *corpus* 7x1-PT contém apenas os *tweets* em português que foram postados durante a partida em que o Brasil perdeu de 7 a 1 para a Alemanha. Escolhemos formar este *corpus* especificamente em razão da grande repercussão gerada pelo resultado desfavorável para o Brasil. Este artigo está organizado em 5 seções. A Seção 2 descreve como os dados do *corpus* foram coletados e organizados. A Seção 3 menciona brevemente a limpeza realizada nos *tweets* do *corpus*. A Seção 4 descreve as anotações de natureza linguística que foram realizadas e as que estão em andamento. E a Seção 5 apresenta as considerações finais deste trabalho.

2. Coleta dos Dados e Anotação Estrutural

Os *tweets* da base WorldCupBrazil2014 foram coletados usando a API Twitter4J, a qual é baseada na API Twitter Rest. O processo de captura dos *tweets* ocorreu entre 30 de

1 A Análise de Sentimentos tem como objetivo determinar as opiniões das pessoas, seus sentimentos, avaliações, apreciações, atitudes e emoções quanto a produtos, serviços, organizações, pessoas, problemas, fatos, eventos e seus atributos [Liu 2012].

maio e 13 de Julho de 2014 e foi baseado em palavras-chave. Foram usadas palavras-chave como "copa", "vencedor", "turistas", "hexa", entre outras. A base WorldCupBrazil2014 foi estruturada em um banco de dados MySQL. Esta base contém tanto informações externas como os horários em que os *tweets* foram coletados, quanto informações internas tais como as *hashtags* que foram encontradas nos textos das mensagens. De cada *tweet*, a base mantém as seguintes informações: *tweet_id* (número de identificação do *tweet*), *message* (texto do *tweet*), *keyword* (palavra-chave usada durante a coleta dos dados para capturar o *tweet*), *timestamp* (horário local, em BRST), *user_id* (identificação do usuário que postou a mensagem), *hashtags* (*hashtags* existentes na mensagem), *links* (URLs presentes no corpo do *tweet*) e *location* (local de origem da postagem da mensagem, quando disponível). O *corpus* 7x1-PT está no formato *csv* e possui ainda mais duas anotações de natureza estrutural [Almeida e Correia, 2008]. São elas: *preprocessed message* (texto do *tweets* limpo e anotado) e *polarity* (polaridade[2] atribuída ao *tweet*). O *corpus* 7x1-PT, atualmente, contém 2.728 *tweets* em Língua Portuguesa, totalizando 35.024 *tokens* e 4.925 *types.* Na Figura 1, são apresentados alguns exemplos de *tweets* do *corpus* 7x1-PT.

"Começou!"
"Eu nao consigo torcer pela seleçao canarinho"
"A a a a a a a a coração !!!!! Vamos q vamos Brasil #BrasilCampeao #vaitercopasim #rumoaohexa #eToiss #BRAvsALE #BRAvsGER que venha a #ARG"
"Não adianta vir de rubro negro #GER"
"O brasil tinha que diminuir a vergonha com 3 gols"
"Era melhor ter ido ver o filme do Pelé."
"Cade Nazareth pra roubar a taça pra gente? #copadomundo #WorldCup #BRAvsGER"
"A BOLA É NOSSA O ESTÁDIO É NOSSO O BRASIL É NOSSO E NÓS SIMPLESMENTE PODEMOS CANCELAR TUDO"

Figura 1. Exemplos de *tweets* existentes no *corpus* 7x1-PT.

3. Limpeza e Normalização

Inicialmente, nós preprocessamos o texto das mensagens para facilitar a anotação de polaridade. Em razão da origem (*web*) e natureza das mensagens (*tweets* sobre futebol), enfrentamos problemas bem conhecidos e amplamente descritos na literatura da área [Duran et al, 2014; Xue et. Al, 2011]. Os *tweets*, em geral, são mensagens curtas, informais, que têm duração limitada e podem conter erros gramaticais (ortografia, pontuação, ...), gírias, clichês, abreviaturas usuais em *chats* ("internetês"), acrônimos, repetições de vogais e *emoticons*[3]. Todos esses elementos são considerados um desafio para abordagens automáticas visto que as ferramentas computacionais disponíveis para o processamento linguístico foram projetadas para textos bem escritos[4] . Inicialmente, as *hashtags*[5] e os *links* existentes no corpo das mensagens foram removidos. Essa limpeza, no entanto, não foi muito simples e teve que ser realizada de forma semiautomática. Um dos principais problemas foi a remoção das *hashtags.* Algumas delas faziam parte das sentenças, desempenhando algum papel sintático. Elas apareciam frequentemente como sujeitos ou objetos, tal como em "... que venha a #ARG". O que fizemos, nesse caso, foi

2 Polaridade: determinação dos pólos de um texto através de suas características (sentimentos em palavras) positivas ou negativas [Liu, 2012].
3 *Emoticon*: Junção dos termos em inglês: *emotion*(emoção) +*icon*(ícone). É uma sequência de caracteres tipográficos, tais como: :), :(ou ^-^ que expressa um estado emotivo.
4 O termo "bem escrito" foi usado no sentido de "bem formado", que segue a gramática da língua.
5 As *hashtags* foram removidas do corpo das mensagens, mas permanecem nas informações externas.

simplesmente remover o caracter #. Assim, para a frase-exemplo, obtivemos: ".. que venha a ARG". Nos casos em que as *hashtags* não faziam parte do texto das sentenças, nós as removemos completamente. Para realização dessa etapa, nós implementamos alguns padrões (expressões regulares) de limpeza e revisamos manualmente o resultado obtido. Outra dificuldade foram os acrônimos e as abreviações usadas pelos internautas, o chamado "internetês". Para resolver esses problemas, tal como em [Agarwal et al, 2011] , criamos um léxico, que mapeava tais abreviações aos termos correspondentes. O que permitiu que, por exemplo, "ARG" fosse transformada em "Argentina" e as ocorrências "q", substituídas pelo termo "que". Essa etapa de transformação também teve que ser revisada manualmente, pois encontramos casos para os quais o léxico não foi suficiente. A ausência de um delimitador (espaço em branco) entre os termos das frases impediu a substituições das abreviações em alguns casos. Isso aconteceu com o termo "oq", o qual deveria ter sido transformado em "o que". Outro elemento que dificultou o processamento automático, foi a falta de pontuação. Em geral, os internautas em mensagens curtas costumam não observar a pontuação, até mesmo no final das frases.

4. Anotação Linguística

A anotação de polaridade, no estágio atual, foi baseada unicamente no sentimento que as pessoas expressavam em relação à seleção brasileira. Nós anotamos manualmente cada *tweet* como negativo, neutro[6] ou positivo. Nós consideramos como positivos os *tweets* que elogiavam ou encorajavam a seleção brasileira. Nós anotamos como negativos aqueles que criticavam ou expressavam sentimentos pessimistas quanto ao desempenho do time brasileiro. As demais mensagens foram classificadas como neutras. A Tabela 1 mostra a distribuição atual de polaridade dos *tweets.*

Tabela 1 . Distribuição de Polaridade do *corpus* 7x1

Polaridade	**# *Tweets* (%)**
Negativo	800 (29 %)
Neutro	1,771 (65%)
Positivo	157 (06%)

A anotação das mensagens é uma tarefa subjetiva e foi realizada por dois anotadores humanos. O índice inicial de concordância observada [Artstein e Poesio, 2008] ficou em 53%. Cabe mencionar que o segundo anotador teve como principal função discutir e revisar a polaridade quando o primeiro anotador tivesse dúvidas. E essas dúvidas ocorreram em vários momentos. Por exemplo, o *tweet* "A Copa das copas" no inicio do jogo era postado como positivo. A partir do quinto gol da Alemanha, no entanto, passou a ser postado de forma irônica, tendo claramente um sentimento negativo. A cada gol realizado pela Alemanha, ironias ("*Esse ... joga muito... vou ate comprar as chuteiras dele*")[7] e sátiras ("*Tira o Hulk e chama os Vingadores*") passaram a acontecer com mais regularidade. Tais construções assim como *tweets* de cunho político ("*Hoje tem manisfestação*"), informativo ("*Cancelamento online de serviços de telefonia passa a valer a partir de hoje*") ou publicitária ("*Expo-noivas...*") foram anotadas como neutras.

6 No estágio corrente de anotação, a polaridade "neutra" possui um sentido mais amplo que o usual. Ela inclui mensagens sem polaridade definida como "Começou!" e também mensagens que não pertencem ao domínio Futebol.

7 O nome foi omitido, mas se referia a um jogador que não estava apresentando um bom desempenho em campo.

É importante mencionar que protelamos a anotação de ironias e sátiras, nesse estágio, pois a abordagem automática de classificação desses *tweets*, que também está em desenvolvimento, não trata textos dessa natureza.

Nós também organizamos o *corpus* conforme a ocorrência dos gols. No gráfico da Figura 2, podemos observar o sentimento dos torcedores ao longo do jogo. Houve mais *tweets* negativos quando o Brasil sofreu o primeiro gol (1T_1x0), no primeiro tempo, e ainda mais quando ele sofreu o quinto gol no segundo tempo de jogo (2T_5x0). A partir do segundo gol da Alemanha (1T_2x0), os *tweets* positivos praticamente deixaram de existir.

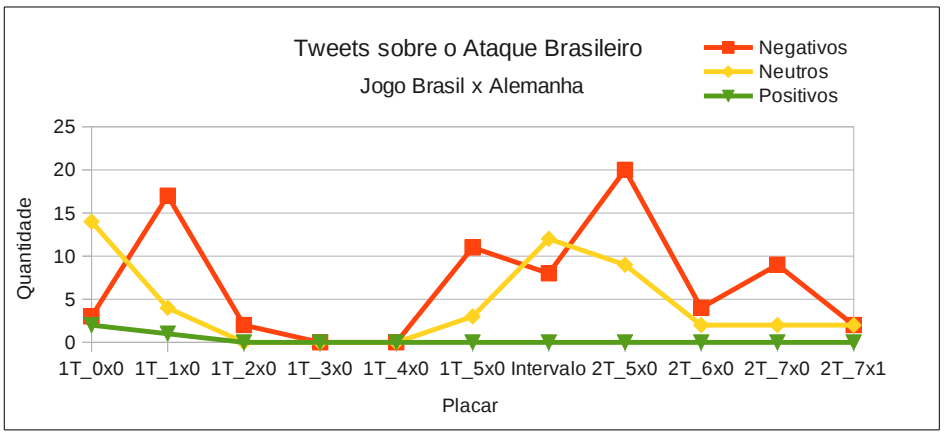

Figura 2. Polaridade dos *tweets* ao longo do jogo.

Atualmente, estamos revisando a anotação de polaridade. Para isso, contamos com dois novos anotadores humanos. O objetivo é dar mais confiabilidade à anotação. Além disso, vamos estender a anotação, definindo polaridade também dos *tweets* de cunho político. Esses *tweets* são uma parte significativa das mensagens que no momento estão anotadas como neutras. Em paralelo, estamos também incluindo etiquetas morfossintáticas, por meio de etiquetadores, no texto dos *tweets*. Desta forma, poderemos incluir análises quanto à distribuição dos termos do *corpus* em categorias gramaticais tal como feito em [Pak e Paroubek, 2010], bem como viabilizaremos estudos em abordagens nas quais a análise de sentimentos é baseada em léxicos.

5. Considerações Finais

A identificação automática dos sentimentos expressos em um texto é um desafio. No caso do Twitter, o desafio é ainda maior em razão da natureza desse serviço (*microblogging*). A postagem de mensagens curtas em tempo real estimula uma escrita diferenciada. Tais diferenças vão desde abreviações a erros de ortografia e sintaxe. Apesar disso, parte do preprocessamento requerido para preparar o *corpus* para análise de sentimentos já foi realizado. Atualmente, estamos revisando e ampliando a anotação de polaridade, bem como incluindo informações linguísticas às mensagens. Nosso próximo passo será incluir a anotação de ironia, visto que uma quantidade significativa do *corpus* é composta por tais mensagens. Hoje, essas mensagens fazem parte das mensagens anotadas como neutras. Uma outra motivação para esta anotação é o fato de existirem poucos *corpora* em Português com essa anotação. Pretendemos também realizar experimentos em Análise de Sentimentos com este *corpus*.

Referências

Agarwal, A.; Xie, B; Vovsha, I.; Rambow, O. e Passonneau, R (2011). "Sentiment analysis of Twitter data". In Proceedings of the Workshop on Languages in Social Media (LSM '11). Association for Computational Linguistics, Stroudsburg, PA, USA, 30-38.

Almeida, G. M. B. e Correia, M. (2008), "Terminologia e corpus: relações, métodos e recursos." In: Avanços de Linguistica de Corpus no Brasil", São Paulo, Humanitas, p.67-94.

Artstein, R. e Poesio, M. (2008), "Inter-coder agreement for computational linguistics". Computational Linguistics, 34, 4, p. 555-596.

Duran, M. S.; Avanço, L. V.; Aluísio, S. M.; Pardo, T. A. S.; Nunes, M. G. V. (2014), "Some issues on the normalization of a corpus of products reviews in Portuguese". In: 9th Web as Corpus Workshop (WAC-9), 2014, Gothenburg, Sweden. 14th Conference of the European Chapter of the Association for Computational Linguistics – EACL, p. 1-7.

Liu, B. (2012), Sentiment Analysis and Opinion Minin, Morgan & Claypool Publishers.

Pak, A. e Paroubek, P. (2010), "Twitter as a Corpus for Sentiment Analysis and Opinion Mining". In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Malta, ELRA.

Xue, Z.; Yin, D. e Davison, B. D. (2011), "Normalizing Microtext". In: Proceedings of the AAAI-11 Workshop on Analyzing Microtext , San Francisco,p. 74-79 .