Relatório da análise de sentimentos no twitter sobre as eleições 2014 utilizando os classificadores: Naive Bayes, SVM e Decision Trees

Felipe Alves Ferreira

Universidade Federal do Amazonas - UFAM
Programa de Pós-graduação em Informática - PPGI
Manaus - AM

Resumo – Este trabalho descreve os resultados obtidos na utilização de classificadores como SVM, Naive Bayes e Decision Trees para a análise de sentimentos em tweets relacionados as eleições de 2014 para presidência da República do Brasil e governo de São Paulo.

I. INTRODUÇÃO

As eleições para a escolha de presidente, governadores, senadores será em outubro de 2014. Um trabalho que ajuda a mensurar as intensões de voto da população é a pesquisa eleitoral realizada através de entrevistas. Com esta pesquisa, é possível avaliar periodicamente a aceitação dos candidatos e suas propostas para o governo durante a fase de propaganda e comícios. O Brasil tem, nos últimos anos, passado por um período de intensas manifestações sociais e isso tornou-se mais evidente com a copa do mundo. Mas um fato tem contribuído bastante para que essas manifestações sociais ocorram: o uso das redes sociais. As redes sociais não só tem aproximado as pessoas, mas também é um meio de expressar opiniões, críticas e também organizar manifestos contra o governo.

Além da pesquisa eleitoral tradicional, as redes sociais como o twitter podem ser uma fonte de evidência de sentimento da população brasileira sobre os candidatos à presidência da república. Esse trabalho pretende investigar através de técnicas de aprendizagem de máquina se um tweet sobre as eleições é positivo ou negativo.

II. BASE DE DADOS

Para a geração da base, foram coletados ao todo 237.800 mil tweets sobre as eleições presidenciais e o governo de São Paulo do dia 20 ao dia 24 de agosto de 2014. Desse total, foram aproveitados somente 911 tweets que possuíam ícones emocionais [1].

Esses 911 tweets foram distribuídos da seguinte forma: 415 tweets rotulados como negativos e 496 rotulados como tweets positivos.

II PRÉ-PROCESSAMENTO

Para a extração dos 911 tweets, foram necessários alguns pré-processamentos:

Remover tweets duplicados: Durante a coleta, foram identificados muitos tweets repetidos, principalmente os retweets. Esses tweets repetidos foram eliminados e apenas um exemplar de cada tweet foi mantido.

Identificar os tweets positivos e negativos: Os tweets foram filtrados baseados em uma lista de ícones que expressam sentimentos positivos **(":)","(:",":-)","(-:",":)", ":D","=D")** e uma outra que lista que expressam sentimentos negativos **(":(","):",":-(",")-:",") :", ": (")**. Dessa forma, os tweets foram rotulados em positivos e negativos.

Remover datas e horas: Datas e horas foram removidas.

Remover @ e URLs: O caracter "@" no início de uma palavra denota que a mesma é um "username". Este caracter foi removido e apenas o "username" foi mantido. Todas as URLs encontradas foram substituídas por "URL".

Remoção de pontuação, caracteres não alfabéticos e conversão para caixa baixa.

III EXTRAÇÃO DE CARACTERÍSTICAS

Para a extração de características, cada tweet foi modelado como bag-of-words. Antes de extrair o vocabulário da

coleção, foi aplicado stemming em todos os termos afim de reduzir a quantidade de termos considerando apenas o radical. Cada tweet é representado por um vetor p1, p2..., pn onde pi representa a presença ou ausência da característica (termo) no tweet. No total, foram encontrados 2.695 atributos iniciais.

Como a quantidade de atributos encontrada é muito maior que a quantidade de instâncias, esses vetores tornam-se esparsos e devido a isso muitos atributos foram removidos após uma análise de ganho de informação. Com base no ganho e informação, uma outra base de dados com os 900 atributos mais significativos foi criada para a realização dos experimentos.

IV. EXPERIMENTOS E AVALIAÇÃO

Os experimentos foram realizados usando três classificadores da weka: SVM(SMO), NaiveBayes, Decision Trees(J48). No caso do SVM, o valor utilizado para o parâmetro C foi 20. Valores acima de 20 produziram o mesmo resultado durante os experimentos. Em consequência disso, 20 foi o melhor valor para o parâmetro C.

A avaliação dos classificadores foi feita utilizando validação cruzada de 10 partes em cada uma das bases (2695 atributos e a outra com 900 atributos) conforme descrito nas seções anteriores. Em outras palavras, cada base foi dividida em 10 partes, das quais 9 foram usadas como conjunto de treino e a parte restante foi usada como teste. O processo foi repetido 10 vezes, utilizando cada uma das partes como teste e produzindo assim os resultados.

O desempenho de cada classificador foi avaliado utilizando cinco métricas: precisão, revocação, medida F, acurácia e RMSE.

V. RESULTADOS OBTIDOS

Primeiramente, analisando os resultados obtidos para os experimentos com a base de dados de 911 instâncias e 2695 atributos obteve-se 81.88% de acurácia para o Naive Bayes, 81.44% para o J48 e 83.09% para o SVM. Neste primeiro cenário, é interessante observar que o desempenho do J48 e NaiveBayes são comparáveis ao SVM que tem um resultado ligeiramente superior. Em relação ao RMSE, o NaiveBayes teve o melhor desempenho enquanto que o SVM foi o pior.

Fazendo uma análise sobre as outras métricas por classe, nota-se que o SVM manteve-se estável nas precisões (83.10% positivos, 83% negativos) enquanto que o Naive Bayes (79.9% e 84.9%) e o J48 (77.9% e 87.9%) tiveram variações significativas, ou seja, ambos tiveram melhores precisão para a classe negativa e piores para a positiva.

Considerando também as métricas recall e F1, nota-se que o SVM teve um desempenho mais equilibrado em relação as classes do que o NaiveBayes e J48.

Table 1 – Cenário 1 - Acurácia e RMSE

	Accuracy	RMSE
Naïve Bayes	81.88%	0.3744
Decision Trees (J48)	81.44%	0.3982
SVM (SMO)(c=20)	83.09%	0.4112

Table 2 – Cenário 1 - Precisão, Revocação e F1

	Class	NB	J48	SMO
Precision	pos	79.90%	77.90%	83.10%
	neg	84.90%	87.70%	83.00%
Recall	pos	89.10%	91.90%	86.50%
	neg	73.30%	68.90%	79.00%
F1	pos	84.30%	84.40%	84.80%
	neg	81.70%	77.20%	81.00%

```
=== Confusion Matrix ===

   a   b   <-- classified as
 442  54 |   a = positive
 111 304 |   b = negative
```

Figure 1 - Cenário 1 - Matriz confusão Naive Bayes

```
=== Confusion Matrix ===

   a   b   <-- classified as
 456  40 |   a = positive
 129 286 |   b = negative
```

Figure 2 - Cenário 1 - Matriz confusão J48

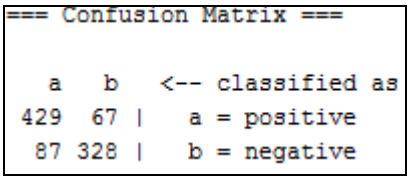

Figure 3 – Cenário 1 - Matriz confusão SVM

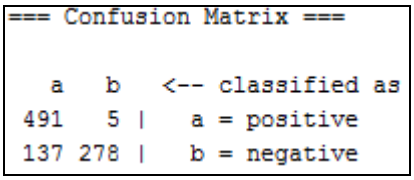

Figure 4 – Cenário 2 - Matriz confusão Naive Bayes

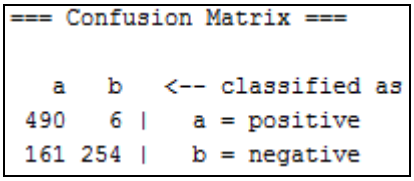

Figure 5 – Cenário 2 - Matriz confusão J48

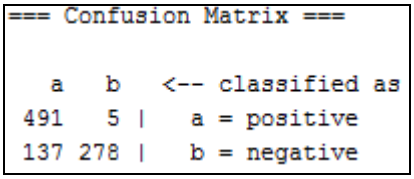

Figure 6 – Cenário 2 - Matriz confusão SVM

Analisando os resultados obtidos para os experimentos com base de dados de 911 instâncias 2 900 atributos, obteve-se 77.16% de acurácia para o Naive Bayes, 81.66% para o J48, e 84.41% para o SVM. Neste segundo cenário, nota-se que enquanto o desempenho do SVM teve uma ligeira melhora, o Naive Bayes piorou e o J48 praticamente se manteve inalterado. Em relação ao RMSE, o J48 foi um pouco melhor que o SVM e o Naive Bayes foi o pior.

Fazendo uma análise sobre as outras métricas por classe, nota-se que todos os classificadores tiveram uma precisão melhor para a classe negativa (acima de 98%) enquanto que para a classe positiva eles obtiveram percentuais abaixo de 79 %.

Considerando também as métricas recall e F1, nota-se que o SVM teve um melhor desempenho em relação as classes do que o NaiveBayes e J48.

Table 3 - Cenário 2 - Acurácia E Rmse

	Accuracy	RMSE
NaiveBayes	77.16%	0.425
Decision Trees (J48)	81.66%	0.3804
SVM (SMO)(c=20)	84.41%	0.3948

Table 4 - Cenário 2- Precisão, Revocação E F1

	Class	NB	J48	SMO
Precision	pos	70.50%	75.30%	78.20%
	neg	99.50%	97.70%	98.20%
Recall	pos	99.80%	98.80%	99.00%
	neg	50.10%	61.20%	67.00%
F1	pos	82.60%	85.40%	87.40%
	neg	66.70%	75.30%	79.70%

VI. Conclusão

Analisando os resultados obtidos, observa-se que o SVM teve um desempenho um pouco melhor que os demais. O Naive Bayes mostrou-se o mais sensível a redução de características tanto que teve uma queda significativa de desempenho. Uma das dificuldades encontradas neste tipo de problema é encontrar o conjunto de atributos relevantes para a geração do modelo. A técnica utilizada foi bag-of-words onde os termos dos tweets viraram as características. Com isso, cada tweet foi representado por um vetor de n posições onde n representa o total de termos do vocabulário dado um conjunto de tweets. Consequentemente, os vetores tornam-se esparsos. Além disso, outros problemas oriundos do twitter não foram tratados neste trabalho como por exemplo as gírias, erros de escrita, ironia e etc.

A rotulagem inicial dos tweets baseada em ícones emocionais mostrou-se válida também para os tweets sobre as eleições de 2014 onde foi possível obter uma boa massa inicial rotulada de tweets para este primeiro experimento.

Como trabalhos futuros, pretende-se aprofundar mais as pesquisas em análise de sentimentos em tweets na língua portuguesa bem como investigar técnicas de machine

learning / data mining com o objetivo de extrair informações relevantes sobre o domínio de tweets relacionados as eleições de 2014.

Também, pretende-se disponibilizar uma base de tweets relacionados as eleições de 2014.

REFERÊNCIAS

[1] M. Felipe, at al. “Polarity Analysis of Micro Reviews in Foursquare”.

[2] http://www.cs.waikato.ac.nz/ml/weka/ Em 26/09/2014.

[3] A. Matheus, at al. “Métodos para Análise de Sentimentos no Twitter”. Webmedia 2013.