DINCON 2017
CONFERÊNCIA BRASILEIRA DE DINÂMICA, CONTROLE E APLICAÇÕES
30 de outubro a 01 de novembro de 2017 – São José do Rio Preto/SP

Método de mineração de textos com foco no português do Brasil com redes bayesianas

Wilker José Caminha dos Santos[1]
Centro de Ciências Naturais e Tecnologia, UEPA, Redenção, PA
Douglas Mariano Tavares[2]
Centro de Ciências Naturais e Tecnologia, UEPA, Redenção, PA
Thiago Nicolau Magalhães de Souza Conte[3]
Centro de Ciências Naturais e Tecnologia, UEPA, Castanhal, PA
Antônio Marcos Cardoso Silva[4]
Centro de Ciências Naturais e Tecnologia, UEPA, Castanhal, PA
Ítalo Flexa Di Paolo[5]
Centro de Ciências Naturais e Tecnologia, UEPA, Belém, PA

Resumo. As redes sociais digitais são cada vez mais populares na sociedade contemporânea, e por isso muita informação útil sobre produtos e serviços podem ser extraídos a partir de opiniões manifestadas nestes ambientes virtuais. Neste trabalho foi analisado o contexto do evento Olimpíadas 2016 na rede social Twitter, com foco de obter os tweets em português brasileiro. Assim, esse trabalho tem como objetivo o desenvolvimento de um método que consegue minerar, analisar e gerar resultados decorrente da classificação com base em perspectivas de aprendizado de máquina e utilizando o classificador Naive Bayes com perspectivas em redes bayesianas escrito em linguagem de programação Python. Para a mineração de textos foi utilizada a API Streaming e base de dados fornecida pelo Twitter. O resultado da análise de sentimentos se deu de forma satisfatória, pois o algoritmo desenvolvido gerou a classificação dos tweets como positivo ou negativo.

Palavras-chave. Aprendizagem de máquina, Mineração de textos, Análise de sentimentos, Olimpíadas

[1] wilkercaminha@uepa.br
[2] douggie.mariano@gmail.com
[3] thiagonconte@uepa.br
[4] amarcos@uepa.br
[5] itflexa@uepa.br

1 Introdução

Com a mineração de textos e análise de sentimentos, buscamos através deste trabalho, desenvolver um método de mineração de textos utilizando a base de dados do Twitter para extração de informações possibilitando à análise de sentimentos com foco no português brasileiro tornando-o capaz de avaliar e analisar o comportamento, interação, opiniões e influências sobre determinado produto, marca, empresa, eventos, como também outras pessoas, não obstante sua opinião ser positiva ou negativa. O método desenvolvido tem como objetivos: a extração de tweets utilizando aprendizagem de máquina; selecionar e filtrar os dados reais a fim de construir um modelo de análise de textos minerados do Twitter; definir, elaborar e validar computacionalmente uma solução baseada em textos extraídos do Twitter.

Pode-se mencionar, por exemplo, que através do feedback que o usuário emitiu sobre um produto em sua rede social online sendo a compra satisfatória, resultar na influência em outros consumidores na compra daquele produto, no qual muitas empresas através de suas aplicações de monitoramento, buscam conquistar mercados e aproximar seus produtos aos seus consumidores através de redes sociais [4].

Para o desenvolvimento do método de mineração e análise de texto é necessário a utilização de um classificador. Utilizaremos o Naive Bayes com perspectivas em redes bayesianas escrito em linguagem de programação Python para executar as tarefas de manipulação e classificação. Durante os procedimentos de extração utilizamos um termo para eventuais testes que envolva minerar, analisar e gerar resultados, sendo escolhido o termo “rio2016”. O resultado da análise após a mineração será visualizado em forma gráfica decorrentes da utilização da biblioteca Matplotlib, também pertencente a linguagem Python.

2 Propósito

O método de mineração e análise de texto, desenvolvido e aqui abordado, reflete características peculiares que tornam sua fundamentação e sua aplicação certamente eficaz, por se tratar de um algoritmo que age em etapas e sub-etapas durante todo o seu processo de execução, dentre elas a capacidade de auto-ajustar a precisão das análises de sentimento extraídos da rede social Twitter, o que torna o método ainda mais relevante no quesito análise de aspectos positivos e negativos, sendo incluído os fatores sentimentais representados com caracteres específicos expressando sentimentos por usuários da rede. Pode ser visto como exemplo o caractere especial que entendido como o alfabeto de emoções pelos usuários da base de dados utilizada. A autenticação da aplicação é executada através de uma chave de segurança cedida pela base de dados do Twitter para que exista a indexação entre o algoritmo e a base de dados contendo as informações dos usuários a qual se destina a extração e análise dos sentimentos.

3 Métodos

O Python é uma poderosa e robusta linguagem de programação de alto nível dotada de uma sintaxe que possibilita a criação de mecanismos e princípios complexos que podem ser aplicadas à inteligência computacional e, sendo assim, enfatizando a extrema importância com base no esforço a qual o programador exerce sobre o esforço computacional. O método

desenvolvido, utilizou dos recursos da linguagem Python para criar um algoritmo capaz de executar uma sintaxe de abstração de dados capaz de comunicar-se com a base de dados do espaço amostral e extrair informações para possíveis análises sentimentais [6].

De fato, a base de dados do Twitter possui uma API (Application Programming Interface) que possibilita a interação entre desenvolvedores e o conteúdo contido em sua base de dados em tempo real, possibilitando a criação de aplicações, onde a mesma, por sua vez, possa utilizar das informações contidas e disponíveis dentre os diversos usuários portadores de uma conta ativa na rede social. A requisição pode ser configurada e executada em tempos determinados para experimentos de extração.

Todo o esquema de desenvolvimento, configuração e adaptação do algoritmo para iniciar o processo de extração consiste na modificação e inserção de permissões para acesso a base de dados do Twitter. Para que haja tal comunicação entre o algoritmo e a base de dados, algumas autenticações são necessárias, tais como Consumer key (ckey), Consumer Secret (csecret), Access Token (atoken) e Access Token Secret (asecret).

É possível escolher o conteúdo para qual o algoritmo fica encarregado de minerar caracteres ou textos referentes ao assunto escolhido. A filtragem de dados do Twitter no evento Olimpíadas 2016 foi realizado no período de 05 a 21 de agosto, fazendo o uso do termo “rio2016”, dessa forma o algoritmo minera apenas caracteres ou informações relacionada ao assunto escolhido.

A pesquisa a qual foi desenvolvida se enquadra como qualitativa-analítica, em virtude de buscar trabalhar no aprofundamento dos dados que estão sendo extraídos do Twitter, sendo que esta coleta será determinante para o estudo e observação dos tweets objetivando a compreensão do contexto de determinado evento ocorrido e qual foi o impacto gerado a partir deste, observando que o evento analisado possa atingir dimensões com um grande público ou grupo específico para avaliar quais as necessidades e discussões de caráter positivo ou negativo envoltas destes [5].

Além disso, a pesquisa se define com traços exploratórios sendo que o pesquisador possuirá uma abordagem de conhecer o tema, enfatizando compreender os acontecimentos vindos da problemática e buscando sua solução [3]. Ainda assim, será necessário a obtenção de conhecimento através de materiais disponibilizados que complementem para o estudo a qual está sendo tratado a linha de pesquisa e que sejam utilizados para fundamentar a teoria com uma pesquisa bibliográfica [3, 5]. Diante da metodologia deste trabalho a ser utilizada e os procedimentos de extração, citados anteriormente, segundo [1, 2] os processos de análise se deram com o uso do classificador Naive Bayes que tem a finalidade de ser um conjunto de algoritmos de aprendizado de máquina supervisionado, pois este tipo de aprendizado segundo [1] se baseia em “utilizar uma série de exemplos (chamados de instâncias), já classificados, para induzir um modelo que seja capaz de classificar novas instâncias de forma precisa, com base no aprendizado obtido com o conjunto de dados de treinamento”, sendo com base no Teorema de Bayes utilizando métodos de estatística e probabilidade para se chegar aos resultados. Ainda convém lembrar que ao utilizar esta classificação, leva-se em conta que esta faz-se o uso do Teorema de Bayes (teoria da probabilidade), tendo como criador Thomas Bayes (1702-1761).

Vale ressaltar que este algoritmo de classificação consiste que os predicativos desse modelo possam a ser utilizados em grandes conjuntos de dados, sendo observado que empregamos como referência a rede social Twitter. Diante isso, incide o uso do método de extração de tweets, e

posteriormente ocorre a aplicação deste poderoso algoritmo que auxiliará na classificação de textos extraídos nos tweets e será utilizado em um módulo específico para tal e assim conter um conjunto de métodos para análise de sentimentos. Por ventura, este módulo será responsável em distinguir quais tweets serão caracterizados como positivo ou negativo a partir do treinamento do classificador, e ao fim apresentar um gráfico com os resultados sentimentais diante do termo que foi solicitado para extração.

4 Resultados

A seguir são apresentados os resultados do método levando-se em consideração diversos fatores para sua análise, alguns questionamentos podem contribuir para melhor percepção das suas fundamentações. Em vista que: Ele conseguiu resolver nosso problema? Quanto tempo o algoritmo consome para processar uma entrada de tamanho n? Qual seria o consumo e tempo em sua pior performance? No quesito onde citamos “Ele conseguiu resolver nosso problema? ”, podemos observar que assim como outros algoritmos de mineração ele alcançou o sucesso esperado, o algoritmo consegue minerar, analisar e gerar resultados decorrente da classificação com base em perspectivas de aprendizado de máquina (AM). Com base no tempo de execução do método de mineração, “Quanto tempo o algoritmo consome para processar uma entrada de tamanho n? ”, o tempo é decorrente da quantidade de dados que forem extraídos e o nível de complexibilidade do caractere solicitado. Ou seja, o tamanho de caractere que forma uma palavra e/ou simbologia a ser minerada interfere no volume de tweets e no tempo de processamento.

Com base nos testes realizados, quanto maior a velocidade do link utilizado melhor será o retorno da análise dos dados minerados. Os testes com links de conexão acima de 2MB possuem suficiência se levarmos em consideração tempo de execução. Testes realizados com 2MB de conexão mostram que o algoritmo consegue minerar 3000 tweets em um período de 3 horas de execução contínua, já com 16MB a capacidade aumenta para 12000 tweets. Algumas conexões de fato não são relevantes para a execução do algoritmo, por não está dentro dos requisitos mínimos para um cenário em tempo real como é o caso de uma simples conexão de 1MB onde a quantidade de dados minerados em um longo período de tempo não se torna relevante como espaço amostral. Esse fator se enquadra no quesito “Qual seria o consumo e tempo em sua pior performance? ” pois é neste ponto em que o algoritmo apresenta o momento de sua pior performance bem como o tempo de execução não é relevante para a quantidade de dados retornados para análise. É notável que durante o processo de desenvolvimento, a etapa que consiste na coleta de dados teve como margem de 2.883.514 tweets armazenados localmente, no qual posteriormente realizamos a etapa de pré-processamento tendo como relevância a retirada de informações que não eram úteis para o desenvolvimento deste trabalho, visto que priorizamos o texto do tweet relacionado ao evento Olimpíadas 2016.

Vale ressaltar que o classificador Naive Bayes garante melhor eficácia e rapidez na previsão de conjunto de dados sendo que através do aprendizado realizado no algoritmo, é possível atingir porcentagem de acurácia satisfatória para se definir a qualidade dos valores classificados. Como é demonstrado na Figura 1, a precisão do classificador Naive Bayes se deu em aproximadamente 71,98% sendo que está utilizando 7000 tweets como conjunto de treinamento. Destaca-se também a utilização de modelos de estimativa de evento que fazem parte do classificador Naive Bayes sendo eles: Multinomial Naive Bayes e Bernoulli Naive

Bayes.

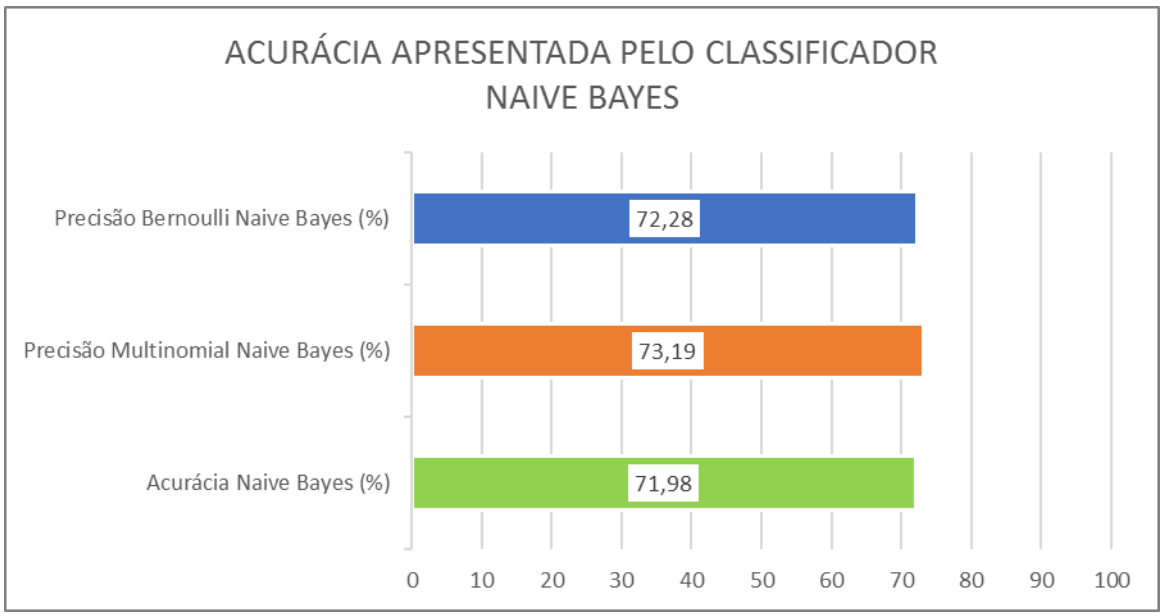

Figura 1: Porcentagem de acurácia apresentada pelo classificador Naive Bayes.

Vale salientar que após conseguir a acurácia do classificador, obtemos junto a ele as palavras que possuem característica de se apresentar como positiva ou negativa. Sendo assim é demonstrado a quantidade de vezes que aquela palavra pode aparecer como valor sentimental positivo em relação ao negativo, ou vice-versa, como é expressa na Figura 2.

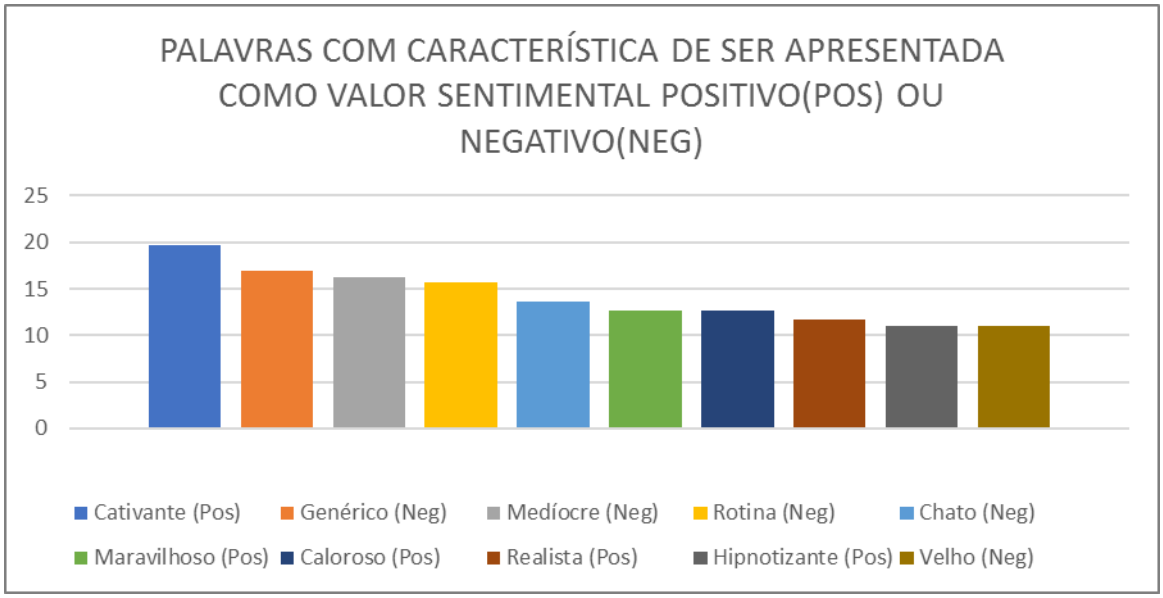

Figura 2: Palavras com característica de ser apresentada como valor sentimental positivo ou negativo.

Para exemplificar a Figura 2, é de se notar que a palavra “cativante” tem característica sentimental de 19,7 vezes de aparecer no conjunto de dados como positivo do que negativo. Por outro lado, a palavra “chato” apresenta valor de 13,6 vezes de se apresentar como negativa em relação ao positivo. Além disso, deu-se ênfase na seleção dos dados de forma random sendo escolhidos 10000 tweets para desempenhar a criação da classificação. Sendo assim, com o conjunto responsável pelo treinamento devidamente classificado, foi separado posteriormente uma porcentagem de 70% dos tweets (7000) para o treino do classificador sendo que 30% dos tweets (3000) para ser utilizado no conjunto de teste do modelo com base em testes realizados entre o hardware disponível e o algoritmo desenvolvido.

Logo após realizado o treinamento e aplicado o conjunto de teste para se obter a porcentagem de acurácia do algoritmo, é realizado posteriormente a etapa de análise de sentimentos utilizando os tweets provenientes da rede social Twitter. Além disso, é importante frisar que o método aplicado tem proporção de atingir textos curtos pois o tweet tem característica de limitação de 140 caracteres por cada mensagem.

Figura 3: Representação gráfica de tweets com valores sentimentais positivos e negativos em tempo real.

A Figura 3 exemplifica o resultado alcançado após o processo de análise de sentimentos, no qual aponta nesta imagem o percentual de tweets classificados como positivos e negativos durante o evento, no qual mostra de forma objetiva os tweets positivos em linha crescente e tweets negativos em direção decrescente. Sendo assim, tweets utilizando o termo “rio2016” como “Vem pro Rio celebrar com a gente que a coisa aqui tá animada #Rio2016!; É ouro!!! Brasil é campeão olímpico de futebol masculino #Rio2016; Parabéns ao futebol Olímpico

masculino Brasileiro #Rio2016" apresenta valor sentimental de caráter positivo. Em contrapartida, mensagens que possuem aspecto de negação, como exemplo o seguinte tweet, "Destaques negativos da Rio 2016: a imprensa paulista, a imprensa dos EUA, a picaretagem dos nadadores norte-americanos e o francês chorão #Rio2016" gera outro significado mesmo contendo o termo utilizado, pois as palavras "negativos, picaretagem e chorão" atrelado ao termo gera um sentimento negativo no texto. É importante salientar que o algoritmo obteve comportamento satisfatório em sua execução e tarefas propostas tendo apresentado classificação de tweets eficiente sendo realizado diariamente durante o evento Olimpíadas 2016.

5 Conclusão

O algoritmo desenvolvido abrange os objetivos propostos deste trabalho, visto que realizou a coleta de tweets sendo realizado a filtragem de dados pertinentes ao evento com o termo "rio2016". As fases de pré-processamento garantiram o conjunto de dados utilizados bem definidos e tratados para o aprendizado e como forma de tornar os resultados autênticos. Diante isso, o método apresentou resultados eficientes provenientes da utilização de memória RAM a partir de 8GB aliado com um processador na arquitetura 64 bits com base nos testes realizados.

Vale ressaltar que em um ambiente de conexão utilizando link com velocidade de 16MB foi possível obter o melhor resultado na mineração de tweets no qual conseguimos 12000 tweets durante o período de 3 horas de execução contínua. Com o uso do conjunto de treinamento e teste apresentados anteriormente, obtivemos a acurácia no classificador Naive Bayes em 71,98% que resultou em melhor eficácia e aprendizado na análise de sentimentos. Com o classificador devidamente treinado, conseguimos obter a classificação dos tweets com valor sentimental positivo ou negativo.

Referências

[1] H. M. Carvalho, Aprendizado de Máquina voltado para Mineração de Dados: Árvores de Decisão, Monografia de Graduação em Engenharia de Software, UNB, (2014).

[2] L. G. Castanheira, Aplicação de Técnicas de Mineração de Dados em Problemas de Classificação de Padrões, Dissertação de Mestrado em Engenharia Elétrica, UFMG, (2008).

[3] A. C. Gil, Como Elaborar Projetos de Pesquisa, Atlas, vol. 5, (2010).

[4] J. Han, M. Kamber and J. Pei, Data Mining Concepts and Techniques, Morgan Kaufmann Publishers - Elsevier, vol. 3, (2012).

[5] E. M. Lakatos e M. de A. Marconi, Fundamentos da metodologia científica, Atlas, vol. 7, (2010).

[6] A. Pak and P. Paroubek, Twitter as a Corpus for Sentiment Analysis and Opinion Mining, Proceedings of the Seventh conference on International Language Resources and Evaluation, 1320-1326, (2010).