REVISTA CIENTÍFICA TRAJETÓRIA MULTICURSOS
E-ISSN 2178 - 4485

Análise de sentimentos do Twitter com Naïve Bayes e NLTK

Augusto Weiand[1]

Resumo

Este artigo propõe um algoritmo de análise de sentimentos dos tweets do microblog Twitter, utilizando o modelo probabilístico de Naïve Bayes. Foram utilizados os dados pré-analisados de (Sanders, 2011) para a construção do corpus e posterior aplicação da análise e validação cruzada. Após, demonstramos o desenvolvimento do algoritmo seguindo a metodologia estudada nos artigos relacionados, utilizando também, as bibliotecas NLTK e Scikit-Learn para o auxílio na aplicação do algoritmo com a linguagem de programação python, medidas de acurácia e validação cruzada dos dados. Organizamos este artigo em sessões que abordam os trabalhos relacionados, a metodologia utilizada, o sistema de coleta de dados, a biblioteca NLTK, o modelo probabilístico Naïve Bayes e por fim, os resultados e os trabalhos futuros, nesta ordem.

Palavras-chave: Twitter, Sentimentos, Naïve Bayes.

Analysis of Twitter feelings with Naïve Bayes and NLTK

Abstract

This article proposes a system of analysis of feelings of the Twitter microblog, using the probabilistic model of Naïve Bayes. We used the pre-analyzed data of (Sanders, 2011) for the construction of the corpus and later application of the analysis and cross-validation. Afterwards, we demonstrate the development of the algorithm following the methodology studied in the related articles, also using the NLTK and Scikit-Learn libraries to aid in the application of the algorithm, python programming language, accuracy measures as well as for the cross-validation of the data. We organize this article in sessions that deal with the related work, the methodology used, the data collection system, the NLTK library, the Naïve Bayes probabilistic model and, finally, the results and future work, in that order.

Keywords: Twitter, Feelings, Naïve Bayes.

Introdução

[1] Mestre em Ciências da Computação – PUC/RS.

Com o advento da WEB 2.0 as mídias sociais apresentaram uma nova forma de obtenção de dados e desenvolvimento de aplicações. As pessoas começaram a compartilhar suas experiências, opiniões em grande quantidade de forma online (Pereira, et al. 2007). Esta massa de dados, disponíveis para os desenvolvedores de sistemas através de Interfaces de Programação de Aplicativos[2] (API's), conjunto de dados (datasets), etc, tem despertado enorme interesse e atrai diversos estudos acerca de mineração de dados, análise automática, entre outros (Han e Kamber, 2011). Sobre este assunto, procura-se neste trabalho, apresentar um algoritmo de análise de textos, buscando entender o sentido contido neles, levando em consideração que estes textos possuem "sentimentos", e que, eles podem ser categorizados em, positivos ou negativos.

Um sentimento é muitas vezes representado em formas sutis ou complexas em um texto. Um usuário online pode usar uma grande variedade de técnicas para expressar suas emoções. Além disso, a mistura de informações objetivas e subjetivas sobre um determinado tópico também podem prejudicar esta classificação, estes casos são denominados ruídos, os quais, são comumente encontrados na maioria dos conjuntos de dados disponíveis, e variam desde simples expressões, até frases completas (palavras de parada, emojis, ironias, etc), tornando necessária a limpeza/modificação destes com técnicas específicas, (Anjaria e Guddeti, 2014). Assim, a tarefa de reconhecimento automático de sentimentos nos textos se torna mais complexa.

Com estas tecnologias emergentes e o interesse cada vez maior da indústria em ter o conhecimento destas pessoas e dos interesses delas, surge um em

[2] Interfaces de Programação de Aplicativos: conjunto de rotinas e padrões de programação para acesso a softwares.

particular em uma plataforma de micro-blogging, o Twitter. Twitter segundo Anjaria e Guddeti (2014), é uma plataforma de micro-blogging lançado em 2006, com mais de 25 milhões de visitantes únicos mensais. No Twitter, qualquer usuário pode publicar uma mensagem curta dita como tweet com um comprimento máximo de 140 caracteres, que é visível na exibição pública. Existe uma linha do tempo também pública, que transmite os tweets de todos os usuários em todo o mundo como um extenso fluxo de informações em tempo real de mais de um milhão de mensagens por hora. Especialmente durante grandes eventos.

Vista a popularidade do Twitter, a análise de sentimentos em tweets tem atraído mais atenção. (Parikh e Movassate, 2009; Barbosa e Feng, 2010; Turney, 2002), seguida da abordagem de aprendizagem de máquina para análise sentimento de tweets. Turney (2002) propôs vários tipos de sentimento para classificar os tweets usando hashtags[3] e emojis como rótulos. Além disso, Barbosa e Feng (2010), propuseram uma abordagem de duas etapas para classificar os sentimentos dos tweets usando Support Vector Machine (SVM), classificadores com características abstratas.

Assim, busca-se apresentar aqui um algoritmo de análise de sentimentos de tweets coletados com auxílio da biblioteca desenvolvida por Sanders (2011), e o algoritmo de Naïve Bayes, implementando este sistema sob a linguagem de programação Python, juntamente com a utilização da biblioteca NLTK e Scikit-Learn. Descrevem- se, nos próximos capítulos mais detalhes sobre trabalhos relacionados, algoritmos e bibliotecas de referência, a metodologia utilizada, a coleta de dados, resultados obtidos e considerações finais.

Referencial teórico trabalhos relacionados

[3] Hashtags: Consiste de uma palavra-chave antecedida pelo símbolo #

Como o Twitter fornece uma API de acesso a seus dados de forma simples, ele tem se tornado um grande centro de pesquisas e desenvolvimento através do tempo. Diversas dissertações e teses são escritas utilizando seus dados, suas aplicações, novas descobertas, etc.

Turney (2002) apresenta um algoritmo de aprendizagem simples, sem supervisão para a classificação de comentários como recomendado ou não-recomendado, utilizando como revisão a orientação semântica média da frase, que contém seus adjetivos ou advérbios. A metodologia aplicada por eles alcançou índices consideráveis, chegando a 74% de acerto quando utilizado em 410 opiniões, coletadas de quatro diferentes domínios: automóveis, bancos, filmes e destinos de viajem. Ele destaca que sua metodologia alcançou um índice de acertos diferenciado para as opiniões de carros, em torno de 84% e para comentários de crítica de filmes 66%.

Anjaria e Guddeti (2014), citam uma metodologia de pesquisa para o desenvolvimento de um sistema de predição de resultados de eleições com o uso de dados do Twitter aliados a mineração de dados e algoritmos de redes neurais artificiais, Naïve Bayes, entre outros, tendo resultados significantes. Eles utilizaram para isso Análise de Componentes Principais (PCA) incorporado com SVM reduzindo as dimensões e alcançando uma melhor precisão, mas nem sempre fornecendo uma saída consistente. Além disso, foi constatado por eles que a inclusão de fatores mais influentes, com base nos dados pessoais, como idade, escolaridade, emprego, critério econômico, rural e urbano e índice de desenvolvimento social aumentaram ainda mais o processo de previsão de votação, com esta metodologia, foi possível obter uma precisão de 88% em caso de eleições presidenciais dos Estados Unidos em

2012 e para as eleições da assembleia de Karnataka em 2013, considerando o resultado obtido ao final do pleito.

Já Espinosa et al. (2013) apresentaram métodos diferentes de representação de dados realizando melhorias significativas sobre modelos de unigramas, que consistem em uma sequência contígua de n itens de uma determinada sequência de texto, com o desenvolvimento de um sistema de aprendizagem de línguas. Para isso utilizaram a rede social Facebook juntamente com o algoritmo de Naïve Bayes. A coleta de dados foi realizada de forma aleatória sem uma consulta mais específica. Com isto eles demonstraram uma precisão significativa no modelo proposto.

Parikh e Movassate (2009) descrevem a análise de sentimentos de atualizações do Twitter, comparando esta tarefa com a de classificação de revisões de filmes e produtos. Para isto eles implementaram duas técnicas de classificação utilizando unigramas de Naïve Bayes, uma com modelo de bigramas e a outra com Entropia Máxima. Os resultados obtidos descreveram uma melhor performance utilizando a classificação desenvolvida por eles com o Naïve Bayes sem a utilização de Entropia Máxima, porém, os autores descrevem que necessitam desenvolver melhor o corpus utilizado aumentando o número de dados para validar de forma mais precisa as etapas de pré-processamento, treinamento e outras, garantindo assim uma melhor fidedignidade da metodologia.

Pode-se observar em todos os estudos realizados nestes e em outros textos lidos, que existem grandes dificultadores para os processos de análise de textos e extração de conhecimento destes, sendo eles de análise de sentimentos ou outros. Um dos mais citados pelos autores é a dificuldade em tratar os ruídos que aparecem nos textos, como as palavras de parada, as

frases irônicas, etc. Outro problema citado para realizar este tipo de avaliação em diversos idiomas é o fato de existirem poucas bibliotecas prontas que dispõem de uma função de redução de radicais das palavras, onde, algoritmos como o Naïve Bayes, podem acabar gerando classificações diferentes para uma mesma palavra, pois ela pode se encontrar conjugada de uma forma diferente nas frases ocasionando problemas de categorização, caso não seja efetuada esta redução.

Segundo Manning et al. (2009) por razões gramaticais, os documentos vão usar diferentes formas de uma palavra, como, organizar, organiza, e organizadora. Além disso, existem famílias de palavras derivacionalmente relacionadas com significados semelhantes, tais como, democracia, democrática e democratização. Em muitas situações, é útil para sistemas de busca, categorização e outros que de uma dessas palavras possam ser retornados documentos que contêm uma outra palavra no conjunto.

Algoritmos e bibliotecas relacionados

Análise de Componentes Principais – PCA

Do inglês, Principal Component Analysis, é um dos métodos estatísticos de múltiplas variáveis mais simples. Ela é considerada a Transformação Linear Ótima, dentre as transformadas de imagens, sendo muito utilizada pela comunidade de reconhecimento de padrões.

A PCA matematicamente definido (JOLLIFFE, 2002) como uma transformação linear ortogonal que transforma os dados para um novo sistema de coordenadas de forma que a maior variância por qualquer projeção dos dados

fica ao longo da primeira coordenada, a segunda maior variância fica ao longo da segunda coordenada, e assim por diante.

Máquinas de Vetores de Suporte – SVM

As Support Vector Machines, constituem em uma técnica de aprendizado que vem recebendo crescente atenção da comunidade de Aprendizado de Máquina nos últimos anos (MITCHELL, 1997). A aplicação desta técnica, demonstra por vezes, resultados comparáveis e em alguns casos, consideravelmente superiores aos

obtidos por outros algoritmos de aprendizado, como por exemplo as Redes Neurais Artificiais (BRAGA et al., 2000). Exemplos de aplicações de sucesso podem ser encontrados em diversos domínios, como na categorização de textos (JOACHIMS, 2002), na análise de imagens (KIM, 2002) e em Bioinformática (NOBLE, 2004).

Naïve Bayes

É um modelo probabilístico simples com base na regra de Bayes com seleção de recursos independentes, foi implementado no sistema e gerou bons resultados na categorização de texto. Ele possibilita a não restrição do número de classes ou atributos, assim como é um dos algoritmos de aprendizado mais rápido para a fase de treinamento de um sistema deste tipo.

O algoritmo é baseado em torno da regra de Bayes, uma maneira de olhar para as probabilidades condicionais que permite que você alterne em torno da condição de uma forma conveniente. A condicional é um evento que provavelmente irá ocorrer X, dada a evidência Y. Isso é normalmente escrito P

(X | Y). A regra de Bayes nos permite determinar essa probabilidade, quando tudo o que temos é a probabilidade de o resultado oposto e dos dois componentes individualmente:

$$P(X|Y) = \frac{P(X)P(Y|X)}{P(Y)}$$

Essa correção pode ser muito útil quando estamos tentando estimar a probabilidade de algo baseado em exemplos de sua ocorrência (Espinosa et al., 2013).

Neste caso, estamos tentando estimar a probabilidade de que um texto é positivo ou negativo, dado o seu conteúdo. Podemos reafirmar que, em termos de probabilidade de que o texto se tiver sido previamente determinado que é positivo ou negativo. Isso é importante, porque temos exemplos de opiniões positivas e negativas dos dados referidos.

Natural Language Toolkit

NLTK é uma biblioteca escrita para a construção de softwares que visam trabalhar com linguagem humana. Ela é desenvolvida na linguagem de programação Python,

sendo hoje uma das ferramentas para este fim mais utilizada, principalmente nos meios acadêmicos.

Dentro de seus recursos podemos citar, interfaces simples para diversos recursos lexicais, como WordNet, juntamente com um conjunto de bibliotecas de processamento de texto para a classificação, tokenização, decorrentencia, marcação, análise e raciocínio semântico (JOACHIMS, 2002).

A utilização desta biblioteca teve como principal objetivo o auxílio ao desenvolvimento e implementação do algoritmo Naïve Bayes, mas, também foram utilizados outros recursos disponíveis por ela, como stemming3 e treino de palavras. Porém, como esta pesquisa também propunha a aplicação de um algoritmo de validação cruzada e a biblioteca NLTK não possuí este algoritmo, foi necessária a utilização de outra biblioteca, a Scikit-Learn para esta etapa.

A validação cruzada, consiste em avaliar a capacidade de generalização de um modelo a partir de um conjunto de dados (KOHAVI, 1995). A técnica consiste em particionar o conjunto de dados em subconjuntos mutualmente exclusivos, e posteriormente, utilizar alguns destes subconjuntos para a estimação dos parâmetros do modelo, como dados de treinamento, sendo o restante dos subconjuntos, dados de validação ou de teste, empregados na validação do modelo, obtendo-se um índice de acurácia.

Scikit-Learn

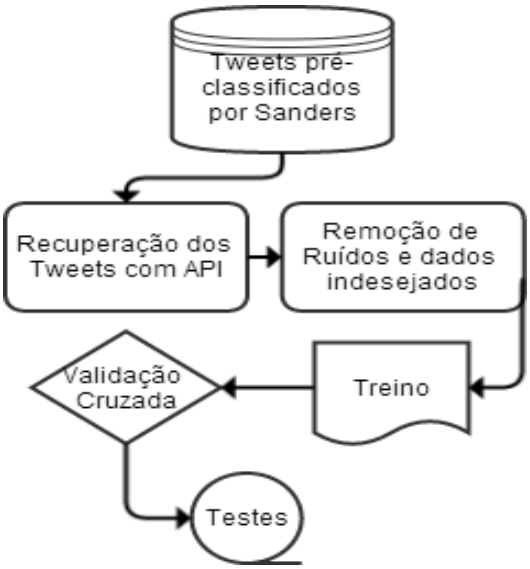
A biblioteca Scikit-Learn implementa diversos algoritmos de aprendizado de máquina, dentre eles podemos citar, algoritmos de classificação, regressão, clusterização, redução de dimensionalidade, seleção de modelo e pré-processamento. Para os fins desta pesquisa utilizou-se o algoritmo que fornece estimativas de performance do modelo preditivo, e de validação cruzada.

Metodologia

Para o desenvolvimento desta pesquisa foi necessário construir um corpus com as classificações dos tweets em positivos e negativos, para que desta forma fossem realizadas as inferências necessárias ao algoritmo de treinamento e

após realizar a aplicação destes nas etapas de cross validation e testes. Assim como a utilização de um algoritmo que forneça uma função de stemming das palavras, a validação cruzada, e a classificação segundo a metodologia proposta, a Naïve Bayes. Seguiu- se para tanto as etapas demonstradas na imagem abaixo de modo que se tornasse possível a obtenção do corpus de forma íntegra.

Figura 1 - Fluxo de análise



Fonte: Autoria Própria, 2017.

A etapa de obtenção dos dados iniciais foi feita com a utilização do material desenvolvido por Sanders (2011), o qual é descrito no capítulo seguinte de Coleta de Dados. Após estas etapas, foram utilizadas as bibliotecas NLTK para a aplicação do algoritmo de Naïve Bayes, a redução dos radicais das palavras, o treino e também para a medição de acurácia das aplicações. A Scikit-Learn foi utilizada para a aplicação da validação cruzada após a etapa de treino.

Estas bibliotecas foram utilizadas em conjunto para obter um melhor desempenho do sistema, o qual foi desenvolvido, assim como as bibliotecas em Python, fornecendo o suporte aos requisitos necessários.

Os dados coletados e utilizados nos procedimentos acima foram distribuídos segundo a seguinte ordem: 90% para treino e 10% para validação cruzada. Com isto foi possível validar a ferramenta de forma a dar continuidade ao seu desenvolvimento.

Coleta de dados

Para a coleta de dados foi utilizada a classe desenvolvida por Sanders (2011), que contém um arquivo em formato “csv”, no qual se obtém números de identificação únicos de tweets juntamente com dois tipos de classificação de sentimentos diferentes, assim como uma classe em Python para que seja possível obter os textos destes tweets diretamente do Twitter, os quais não são distribuídos com a classe pois, segundo o autor, há direitos autorais que não permitem disponibilizar estes de forma pública, já categorizados.

Foram necessárias algumas modificações na biblioteca utilizada pois a mesma estava utilizando a versão 1.0 da RestAPI[4] de acesso aos dados do Twitter, a qual já não está mais disponível, necessitando a atualização da classe em Python para a versão mais atual da RestAPI.

Desta forma foram coletados 5513 tweets pré-classificados advindos deste script, porém diversos destes estavam em idiomas que não o inglês, constavam mais de duas categorias elencadas pelo autor, as quais não seriam de interesse nesta pesquisa, assim como muitos não foram baixados por algum

[4] Twitter RestAPI: https://dev.twitter.com/rest/public

problema com o número de identificação do tweet, portanto, foram eliminados do dataset final os registros que continham estes padrões fora do esperado, resultando então em 494 registros.

Resultados e discussão

Com a utilização da biblioteca de Sanders (2011), foi possível trabalhar em um corpus de 5513 tweets pré-processados, porém, necessitando a recuperação dos mesmos através da RestAPI do Twitter. A partir das correções efetuadas no script e do download destes dados com sistema resultante, foi possível obter um corpus de 494 tweets, o que corresponde a 8,9% do dataset original, a falta de dados se decorreu de tweets que não foram possíveis de se recuperar pela RestAPI do Twitter, assim como também devido as classificações que estes dados continham, onde nesta pesquisa foram utilizadas somente duas das quatro classificações fornecidas por Sanders além, é claro, do idioma, nem todos os dados estavam em inglês.

Estes dados foram separados para o desenvolvimento do treino e da validação cruzada, assim como para posteriores testes, com isto o dataset foi separado de modo aleatório em dois grupos, um com 80% dos dados, os quais foram utilizados para o treino do algoritmo, 10% dos dados para a validação cruzada e os 10% restantes para o teste. Para a aplicação destas etapas foram utilizadas as bibliotecas NLTK e Scikit-Learn, sendo que os resultados obtidos trouxeram um número de acerto relativamente alto, 0,91 nos conjuntos mencionados.

Após o treino, foram aplicados os testes para constituição da validação cruzada, resultando nos dados descritos na tabela abaixo.

Tabela 1 - Validação Cruzada

	Pré-Classificação	**Classificação**
Positivo	19	15
Negativo	30	34
Validação Cruzada	49	49

Fonte: Autoria Própria, 2017.

Ressalta-se a diferença entre a pré-classificação e a classificação, que em ambos os casos foi de 4 itens, para o cálculo de validação cruzada foi utilizada a implementação da própria biblioteca Scikit-Learn.

Considerações finais

Espera-se que com o desenvolvimento deste algoritmo, se torne possível uni-lo a trabalhos relacionados com a área de computação gráfica, tornando possível um sistema de visualizações mais completo, contemplando mais sentimentos, mais idiomas, seleções de filtros de assuntos ou palavras, etc. A integração com a pesquisa no microblog Twitter poderia ser expandida, ou tornar o sistema disponível como uma interface que possa ser integrada como uma aplicação para diversos outros tipos de blogs, redes sociais, entre outros.

Referências

ANJARIA, Malhar; GUDDETI, Ram Mahana Reddy. **Influence factor based opinion mining of Twitter data using supervised learning**. Communication Systems and Networks (COMSNETS), Sixth International Conference on. Disponível em: http://ieeexplore.ieee.org/xpls/icp.jsp?arnumber=6734907 acesso em: 10/06/14. Acesso em: 2 mar. 2017.

BARBOSA, Luciano; FENG, Junlan. **Robust sentiment detection on Twitter from biased and noisy data**. Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Pages 36-44 Association for Computational Linguistics Stroudsburg, PA, USA.

BRAGA, Antônio De Pádua; CARVALHO, André Ponce De Leon E De; LUDERMIR, Teresa Bernarda. **Redes Neurais Artificiais: Teoria e Aplicações**. Editora LTC, 2000.

ESPINOSA, Kurt Junshean; LLAGUNO, Kevin; CARO, Jaime. **Sentiment analysis of Facebook statuses using Naive Bayes classifier for language learning.** Information, Intelligence, Systems and Applications (IISA), Fourth International Conference on. . Disponível em: http://ieeexplore.ieee.org/xpls/icp.jsp? arnumber=6623713. Acesso em: 2 mar. 2017.

HAN, Jiawei; KAMBER, Micheline. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers Inc. San Francisco, 2011.

JOACHIMS, T. **Learning to classify texts using support vector machines: methods, theory and algorithms**. Kluwer Academic Publishers, 2002.

JOLLIFFE I.T. **Principal Component Analysis**. Springer Series in Statistics, Springer, NY, 2002.

KIM, Kwang In; JUNG, Keechul; PARK, Se Hyun; KIM, Hang Joon. **Support vector machines for texture classification**. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002.

MANNING, Christopher D.; RAGHAVAN, Prabhakar; SCHÜTZE, Hinrich. **An Introduction to Information Retrieval**. Cambridge, England. Cambridge University Press. Disponível em: http://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf. Acesso em: 2 mar. 2017.

MITCHELL, T. **Machine Learning**. McGraw Hill, 1997.

NLTK. Natural **Language Processing with Python**. Disponível em: http://www.nltk.org/book/. Acesso em: 2 mar. 2017.

NOBLE, William Stafford. **Support vector machine applications in computational biology**. In: B. Schölkopf, K. Tsuda, and J.-P. Vert, editors, Kernel Methods in computational biology, pages 71–92. MIT Press, 2004.

PARIKH, Ravi; MOVASSATE, Matin. **Sentiment Analysis of User-Generated Twitter Updates using Various Classification Techniques**, Stanford University.

PEREIRA, Alice Theresinha Cybis; SCHMITT, Valdenise; DIAS, Maria Regina Alvares. **Virtual learning environments. Virtual learning environments in different contexts**, Ciência Moderna, Rio de Janeiro, 2007.

KOHAVI, Ron. **A study of cross-validation and bootstrap for accuracy estimation and model selection**. In Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2 (IJCAI'95), Vol. 2. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1995, 1137-1143.

SANDERS, Niek. **Twitter Sentiment Corpus**. Disponível em: http://sandersanalytics.com/lab/twitter-sentiment/. Acesso em: 2 mar. 2017.

SCIKIT-LEARN. **Machine Learning in Python**. Disponível em: http://scikit- learn.org/stable/index.html. Acesso em: 2 mar. 2017.

TURNEY, Peter D.. **Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews**. Proceedings of the 40th Annual Meeting on

Association for Computational Linguistics, Pages 417-424. Association for Computational Linguistics (ACL), USA DOI 10.3115/1073083.1073153.