ISSN 0104-8112

SISTEMAS DE ANÁLISE DE SENTIMENTOS USANDO DADOS DO TWITTER

SENTIMENT ANALYSIS SYSTEMS USING TWITTER DATA

Rafael Bilibio Franco [*]
Mario Henrique Akihiko da Costa Adaniya [**]

RESUMO

A ascensão das redes sociais mudou profundamente a forma como as informações e opiniões são propagadas, disponibilizando um montante de dados sem precedentes. Isso trouxe, contudo, novos problemas. Extrair informações relevantes e assertivas de um emaranhado tão grande de dados é um dos maiores e mais recompensadores desafios contemporâneos. A análise de sentimentos se presta a resolver esse problema. Nela, são empregadas técnicas computacionais com o objetivo de construir sistemas que atuem de forma automatizada na classificação de sentimentos extraídos de documentos escritos em linguagem natural, como os provenientes de redes sociais. Neste trabalho, estão reunidos os principais conceitos, padrões e técnicas empregados na análise de sentimentos, com um foco especial sobre os trabalhos que vem sendo desenvolvidos para a predição de eleições governamentais usando dos dados do Twitter.

Palavras-chave: Análise de sentimentos. Redes sociais. Aprendizado de máquina. Processamento de linguagem natural.

ABSTRACT

The rise of social networks has profoundly changed the way information and opinions are propagated, providing an unprecedented amount of data. This has, however, brought new problems. Extracting relevant and assertive information from such a tangle of data is one of the greatest and most rewarding contemporary challenges. Sentiment analysis lends itself to solving this problem. In it, computational techniques are employed with the objective of building systems that act in an automated way in the classification of feelings extracted from documents written in natural language, such as those coming from social networks. In this work, the main concepts, standards and techniques used in sentiment analysis are gathered, with a special focus on the works that have been developed for the prediction of government elections using data from Twitter.

Keywords: Sentiment analysis. Social media. Machine learning. Natural language processing.

[*] Discente do Curso Ciência da Computação. E-mail: rbilibiofranco@gmail.com
[**] Docente do Curso Ciência da Computação. E-mail: mhadaniya@gmail.com

ISSN 0104-8112

INTRODUÇÃO

As redes sociais ganharam o mundo rapidamente desde o seu surgimento. O Twitter é uma rede social que se destaca pelas interações de caráter imediato. Nele, os usuários possuem microblogs nos quais podem escrever mensagens de até 140 caracteres, denominadas "tweets". Isolados, os tweets podem parecer aleatórios e pouco representativos da opinião geral, mas quando observados dentro de um contexto maior, os tweets permitem a análise dos assuntos mais discutidos na Internet em um dado momento, bem como a análise das opiniões sobre os mais variados temas. Este trabalho focará nesse último ponto.

À análise das opiniões emitidas sobre um determinado tema ou entidade, dáse o nome Análise de Sentimentos. O objetivo da análise de sentimentos é garimpar massas de dados, como as geradas pelo Twitter e outras redes sociais, em busca da estruturação de opiniões relevantes a respeito da entidade alvo da análise. Com isso, espera-se partir de um ponto em que há uma abundância de dados, mas uma escassez de informações assertivas, para um ponto no qual essas informações que antes estavam perdidas em meio aos dados estejam sintetizadas e passíveis de serem utilizadas para tomada de decisão ou até mesmo predição de eventos.

Neste artigo, pretende-se expor de forma sumarizada alguns dos trabalhos que vêm sendo feitos sobre o tema da análise de sentimentos, tendo como enfoque o uso do Twitter como fonte de dados para a análise preditiva de eleições governamentais. Intenciona-se mostrar que tal forma de análise pode ter uma precisão similar à das metodologias tradicionais empregadas no domínio de estudo, com um menor custo e com maior grau de automatização. Serão comentados alguns dos métodos mais utilizados na coleta e categorização de sentimentos, bem como a estrutura geral dos sistemas comumente construídos.

CONCEITOS FUNDAMENTAIS

Para dar maior claridade ao texto, convém esclarecer alguns conceitos fundamentais sobre o tema abordado neste trabalho:

ISSN 0104-8112

Sentimento: na análise de sentimentos, é uma opinião expressa por um indivíduo ou organização a respeito de uma determinada entidade. É altamente específico e particular de seu emissor. Há dois grandes tipos identificáveis de sentimentos (BALDANIA, 2017): os sentimentos comuns, e os comparativos. Sentimentos comuns podem ser diretos, quando enunciam algo a respeito de uma entidade explicitamente definida, ou indiretos, caso o sentimento só possa ser abstraído mediante uma análise do contexto. Sentimentos comparativos, por sua vez, são aqueles que tratam de similaridades ou diferenças entre duas ou mais entidades ou aspectos de uma entidade.

Documento sentimental: são artefatos que expressam objetiva ou subjetivamente um sentimento. Documentos extraídos de redes sociais tendem a ser carregados de subjetividade, característica que os torna especialmente interessantes para a análise de sentimentos, devido à riqueza e facilidade de extração desses sentimentos.

Polaridade: atributo representativo da positividade ou negatividade de um documento sentimental. Comumente definido como um resultado discreto binário ou ternário, sendo no primeiro caso positivo ou negativo, e no segundo caso positivo, neutro ou negativo. A atribuição da polaridade a documentos sentimentais é o que permite extrair informações relevantes a partir de uma grande massa de dados.

NÍVEIS DE ANÁLISE DE SENTIMENTOS

Convencionou-se na literatura a definição de três níveis de análise, indicativos do grau de granularidade sobre o qual a exploração é feita. São eles:

Análise de sentimentos a nível de documento: o nível de mais grossa granularidade. Nele, procura-se classificar todo o documento sentimental, partindo do princípio de que ele diz respeito a uma única entidade. Sua maior utilidade se dá na classificação de textos de grande especificidade, como reviews de produtos.

Análise de sentimentos a nível de sentença: neste nível, assume-se que o documento sentimental pode conter diversas opiniões que referenciam, também, diversas entidades. A partir dessa premissa, a conclui-se que o menor nível de análise aqui serão as sentenças do documento. É o nível de análise mais utilizado

ISSN 0104-8112

em investigações feitas sobre redes sociais como o Twitter, que fornecem documentos curtos mas que ainda assim podem possuir uma pluralidade de opiniões e entidades.

Análise de sentimentos a nível de aspecto: nível de mais fina granularidade. Nele, o objetivo é identificar o texto analisado como sendo uma feature, ou seja, um aspecto de uma outra entidade. Dessa forma, estes aspectos da entidade se tornam o alvo da opinião.

FLUXO GERAL DE UM SISTEMA DE ANÁLISE DE SENTIMENTOS

Os sistemas de análise de sentimentos possuem algumas etapas idiossincráticas (BALDANIA, 2017), (RAMZAN, 2017). Segue uma breve descrição delas:

- **Coleta de Dados:** consiste na obtenção dos dados a partir de sua fonte primária, como tweets, blogs, newsfeeds, etc. Neste ponto, os dados se encontram em sua forma bruta e não é possível tirar proveito deles de forma eficiente.
- **Pré-processamento:** etapa na qual são removidos os dados inconsistentes, incompletos, ou que contém ruídos. O pré-processamento dos dados tem se tornado mais sofisticado, como pode ser observado no estudo de (IBRAHIM, 2015), que obteve bons resultados ao desenvolver um método automatizado de Buzzer Detection. No trabalho, os Buzzers foram definidos como contas que só falavam sobre um dos candidatos das eleições analisadas, enquanto criavam uma imagem vilanesca de outros candidatos. Segundo os autores, isso seria um forte indicativo de que se trata de um usuário que emite opiniões artificiais, com a única intenção de direcionar o pensamento de outros usuários.
- **Seleção e extração de features:** a seleção de features é o processo no qual são identificados os aspectos mais relevantes da entidade para a construção do modelo. Já a extração de features, consiste em efetivamente destacar dos documentos as features mais relevantes

ISSN 0104-8112

para o contexto da análise. Os métodos mais utilizados para a extração de features são o bag-of-words, n-gram e híbrido, com trabalhos (ANJARIA, 2014) indicando que o método híbrido é o mais eficaz.

- **Classificação de sentimentos:** etapa em que um algoritmo de classificação é aplicado para categorizar os dados até então processados em sentimentos bem definidos. Alguns algoritmos amplamente utilizados nesta etapa são o Naïve Bayes, Support Vector Machine e Maximum Entropy.
- **Polarização de sentimentos:** fase final em que o sentimento expresso no documento é posicionado de acordo com sua polaridade. É a partir da agregação dos sentimentos polarizados que uma análise mais assertiva da informação passa a ser viável.

ABORDAGENS E TÉCNICAS

Os trabalhos feitos no campo da análise de sentimento costumam se valer de pelo menos uma das três abordagens principais para classificação e polarização dos sentimentos. São elas a abordagem baseada em aprendizado de máquina, a baseada na análise léxica e a híbrida.

Na abordagem baseada em aprendizado de máquina, são aplicados algoritmos supervisionados, não-supervisionados, ou semi-supervisionados para a classificação dos sentimentos. A classe de algoritmos mais utilizada é a dos supervisionados, uma vez que ainda há uma grande dificuldade na utilização eficaz de técnicas não-supervisionadas neste campo. É bastante desejável, porém, que esta dificuldade seja revertida, uma vez que os algoritmos supervisionados demandam um nível considerável de tempo daqueles que treinam o modelo utilizado pelo sistema. Sobre os algoritmos de aprendizado de máquina mais utilizados, há um grande número de estudos utilizando os algorítmos Naïve Bayes, Support Vector Machine e Maximum Entropy com altos níveis de acurácia, como visto em (ANJARIA, 2014), (YANG, 2017).

ISSN 0104-8112

TRABALHOS RELACIONADOS

Já na abordagem baseada na análise léxica, classifica-se os documentos a partir de dicionários que contém palavras associadas aos seus respectivos valores sentimentais. Aqui, pode haver uma dificuldade caso a linguagem escolhida não possua tais dicionários prontos, ou possua dicionários limitados. Como era de se esperar, a língua inglesa é a que possui o maior número de dicionários disponíveis. Felizmente, apesar de ser sim um inconveniente, isso não chega a ser um impeditivo para o desenvolvimento de trabalhos focados em outras línguas que não a inglesa.

A abordagem híbrida, por fim, se vale de uma mistura das duas abordagens anteriores. Há pesquisas que indicam que a abordagem híbrida é a que traz um nível mais alto de acurácia na classificação, a exemplo de (ANJARIA, 2014).

Baldania (2017) fez uma revisão dos conceitos e metodologias empregadas na Análise de Sentimentos voltada para a análise de resenhas de filmes. Embora o domínio analisado no estudo em questão não seja o mesmo que o colocado em foco neste trabalho, é importante salientar que as técnicas e metodologias empregadas na análise de documentos sentimentais são essencialmente as mesmas, especialmente quando essa análise é feita ao nível de sentença.

Yang (2017) focou na revisão dos métodos de aprendizado de máquina mais utilizados na análise de sentimentos. Foram comparados os algoritmos *Support Vector Machine* (SVM), *Naive Bayes* (NB), *Maximum Entropy* (ME) e *Artificial Neural Network*. Na comparação, foram definidos as acurácias teóricas desses algorítmos: SVM - baixa, NB - alta, ME - moderada e ANN - alta. Aliado a isso, foram examinadas as velocidades de treinamento de cada técnica, no aspecto teórico, apontando para a seguinte relação: SVM - alta, NB - alta, ME - moderada e ANN - baixa. É interessante observar, portanto, que na comparação feita o algoritmo *Naive Bayes* destaca-se por conciliar uma alta precisão a uma rápida velocidade de treinamento, aspectos muito interessantes para a avaliação desse tipo de técnica.

A análise feita em Anjaria (2014) examinou duas eleições distintas, uma sendo a de 2012 dos Estados Unidos e outra a de 2013 da Índia. Um dos objetivos do trabalho foi comparar os diferentes métodos de *feature extraction* comumente utilizados, sendo eles o *Unigram*, *Bigram*, *Bag of Words* e híbrido, com a conclusão de que o método híbrido é o mais eficaz, seguido na maioria dos casos pelo *Bag of Words*. Um fênomeno peculiar observado no experimento foi a diferença nas taxas de acerto das análises entre as duas eleições analisadas. Verificou-se uma taxa de acerto muito maior na análise das eleições dos Estados Unidos, o que poderia ser explicado, segundo os autores, pela diferença no percentual de usuarios do Twitter perante a população em geral, que é muito mais elevado nos Estados Unidos do que na Índia.

O trabalho feito por Joyce (2017) utilizou na análise léxica o *OpinionFinder Lexicon* para a classificação dos tweets em positivos ou negativos. Esse sistema foi desenvolvido pelas universidades de Pittsburgh, Cornell e Utah, e é capaz de identificar sentenças subjetivas e agentes que são fontes de opinião de forma automática. Ele serve, também, como um dicionário de palavras positivas e negativas pré-classificadas. Dessa forma, a metodologia aplicada contava o número de palavras positivas e negativas em cada tweet, classificando-o como positivo caso a contagem de positivos fosse superior, e negativo caso a contagem de negativos fosse inferior. Caso a contagem fosse igual, o tweet era classificado como neutro.

Ramzam (2017) seguiu uma abordagem ligeiramente diferente ao analisar 10.000 tweets a respeito das eleições de 2017 da Índia, armazenando os tweets já acompanhados de suas classificações em um banco de dados NoSQL, o MongoDB, a fim de analisar posteriormente esses dados.

CONCLUSÃO

A análise de sentimentos é um campo de grande interesse tanto para a computação quanto para as ciências sociais, uma vez que permite extrair informações relevantes sobre os sentimentos de uma quantidade massiva de indivíduos, o que é impossível utilizando métodos tradicionais de pesquisa e consulta popular.

Assim como outras aplicações da Inteligência Artificial, há uma constante e rápida evolução nos métodos empregados, que atualmente já permitem um grau de acerto comparável ao de métodos de análise estatística convencionais, como os utilizados em pesquisas eleitorais. Espera-se, portanto, que esta técnica seja cada vez mais utilizada em cenários do tipo, trazendo uma visão mais dinâmica da opinião dos indivíduos.

Este trabalho expôs os principais conceitos e técnicas utilizadas em sistemas de análise de sentimento, e espera-se que tenha ficado evidente que, embora muito já tenha sido feito, ainda há grande possibilidade de avanço na produção de conhecimento sobre as questões aqui levantadas.

REFERÊNCIAS

BALDANIA, R. Sentiment analysis approaches for movie reviews forecasting: A survey. In: 2017 INTERNATIONAL CONFERENCE ON INNOVATIONS IN INFORMATION, EMBEDDED AND COMMUNICATION SYSTEMS (ICIIECS), 2017, Coimbatore, India**. Proceedings…** Coimbatore, India: IEEE, 2017, p. 1-6.

RAMZAN, M.; MEHTA, S.; ANNAPOORNA, E. Are tweets the real estimators of election results? In: 2017 TENTH INTERNATIONAL CONFERENCE ON CONTEMPORARY COMPUTING (IC3), 2017, Noida, India. **Proceedings…** Noida, India: IEEE, 2017, p. 1-4.

IBRAHIM, M. et al. Buzzer Detection and Sentiment Analysis for Predicting Presidential Election Results in a Twitter Nation. IN: 2015 IEEE INTERNATIONAL CONFERENCE ON DATA MINING WORKSHOP (ICDMW), 2015, Atlantic City. **Proceedings…** Atlantic City: IEEE, 2015, p.1348-1353.

ANJARIA, M. et al. Influence factor based opinion mining of Twitter data using supervised learning. In: 2014 SIXTH INTERNATIONAL CONFERENCE ON COMMUNICATION SYSTEMS AND NETWORKS (COMSNETS), 2014, Bangalore. **Proceedings…** Bangalore: IEEE, 2014, p. 1-8.

YANG, P. et al. A survey on sentiment analysis by using machine learning methods. In: 2017 IEEE 2ND INFORMATION TECHNOLOGY, NETWORKING, ELECTRONIC AND AUTOMATION CONTROL CONFERENCE (ITNEC), 2017, Chengdu. **Proceedings…** Chengdu: IEEE, 2017, p. 117-121.

JOYCE, B. et al. Sentiment analysis of tweets for the 2016 US presidential election. In: 2017 IEEE MIT Undergraduate Research Technology Conference (URTC), 2017, Cambridge. **Proceedings…** Cambridge: IEEE, 2017, p. 1-4.