Minerando o sentimento das mídias sociais: uma análise de sentimento dos tweets a partir da coleta de suas hashtags

Proponente(s):
RA 11049114 - Richard Anemam Da Costa

Santo André, 08 de Maio de 2018

Resumo

As redes sociais tornaram-se sinônimos de big data, pois possuem uma capacidade de prover grandes quantidades de conteúdo gerado pelos seus usuários. Os conteúdos são criados a todo momento e possuem todos os formatos possíveis: texto, áudio, vídeo e foto. A partir desses dados é possível compreender opiniões, tendências e o sentimento público em geral. O conhecimento dessas informações levam as instituições, sejam elas públicas ou privadas, a melhor promover os seus produtos, possuir um melhor relacionamento com o público e engajar os seus consumidores.

O objetivo deste trabalho é fazer uma análise de sentimento simplória de uma coleção de tweets baseado em uma hashtag em comum. Os tweets são pequenas mensagens de texto carregadas de informação como data da criação, se ele foi favoritado ou retweetado, o idioma do tweet, o id do usuário, coordenadas geográficas, hashtags, menções, URLs, símbolos diversos, o texto e etc. Os tweets aqui passarão por uma série de processos, este trabalho propõe a análise apenas do texto, que são coletados a partir das hashtags contidas no tweet. Será utilizado um série de expressões regulares para remover URL`s, menções e id de usuário. As palavras serão colocadas em forma de tokens e então as stopwords são removidas. A partir de então a polaridade de cada tweet é a analisada e então o sentimento dos tweets que contêm a hashtag escolhida é mostrado em um gráfico de sentimentos que classifica os sentimentos como positivo, neutro e negativo.

Palavras-chave: tweet, hashtag, sentimento

1. Introdução

Sentimento são estados que afetam as pessoas diante de um acontecimento. Um sentimento pode influenciar uma atitude, uma emoção, ou até mesmo, a expressar uma opinião. A análise de sentimento é uma área de processamento de linguagem natural útil para identificar a emoção por trás das palavras, ou seja, é um método para que possamos compreender como as pessoas se sentem em relação a um determinado tópico.
As aplicações são inúmeras e podem ser fundamentais para empresas e instituições. As técnicas de análise de sentimento podem ser adequadas em processos de Business Intelligence, na qual as empresas buscam por reviews e opiniões sobre os seus produtos. A política também pode ser alvo da análise de sentimentos no sentido de busca por tendências, ideologias e reações. A sociologia é passível de aplicar métodos de análise sentimental no que diz respeito a busca por propagação de diferentes ideias entre a população como um todo. Portanto, as tendências, opiniões, ideias e etc., são deveras importante do ponto de vista humano. Ser capaz de detectar os sentimentos expressos por meio de análises nos permite tomar melhores decisões e distribuir melhor as informações.
Assim, verificada a importância e as aplicações da análise de sentimentos, este projeto introdutório ao tema busca compreender o sentimento dos tweets[1], em que, estes serão classificados como positivos, negativos e neutros e, então, apresentados em uma gráfico de sentimentos.

1.1. Objetivo Geral

O objetivo geral deste trabalho é fazer uma análise de sentimento simplista sobre dados assuntos discutidos nos trending topics na rede social Twitter. Os tweets contendo determinada hashtag serão tratados e então analisados e classificados como positivo, negativo e neutro. As quantidades de tweets inerentes a cada uma destas classificações nos dará uma perspectiva completa sobre o tema tratado, ou seja, seremos capazes de dizer se o sentimento da população em relação ao assunto é positivo, negativo ou neutro.

1.2. Objetivos Específicos

- Aplicação dos conceitos de processamento de linguagem natural
- Entender o que é análise de sentimento
- Extração de dados utilizando a biblioteca tweepy
- Realizar o tratamento de dados (leia-se *texto*) (remoção de stopwords, definição de tokens, utilização de expressões regulares)
- Atribuição de sentimentos por meio da biblioteca textblob

[1] Tweets são pequenas mensagens de texto de até 140 caracteres postadas na rede social Twitter.

2. Fundamentos

Alguns fundamentos são essenciais para uma melhor compreensão do projeto. Para que se tenha maior fluidez no entendimento do trabalho, é fortemente recomendado que se tenha conhecimento das técnicas de PLN[2] apresentadas abaixo e que possua uma breve compreensão do que é análise de sentimento.

2.1. Expressões Regulares

O projeto se utiliza de algumas técnicas de PLN para o tratamento do texto. A primeira que vamos tratar aqui são as expressões regulares. Estas são formas precisas de identificarmos cadeias de caracteres específicas, como uma URL ou uma menção a um usuário do Twitter, por exemplo.

2.2. Stopwords

O segundo conceito fundamental são as stopwords, que são palavras de grande frequência e que são consideradas irrelevantes em adição de significado dependendo do contexto em que se faz a pesquisa.

2.3. Tokenização

Os textos são representados por uma cadeia de palavras. A tokenização de um texto faz com que cada palavra seja transformada em um token. Por exemplo:

Meu nome é Richard.
[‘Meu’, ‘nome’, ‘é’, ‘Richard’, ‘.’]

2.5. Análise de Sentimentos

Extração de informação (sentimentos) de textos em linguagem natural.

3. Método

Dado o objetivo de realizarmos a análise de sentimento dos tweets, foram empregadas as bibliotecas Tweepy e TextBlob. A primeira nos dá autenticação para utilizar a API do Twitter, sendo a partir dela que realizamos a coleta dos dados textuais (tweets). A biblioteca TextBlob é utilizada no processamento de dados textuais. Ela provê uma API simples para

[2] Processamento de Linguagem Natural

trabalharmos com tarefas processamento de linguagem natural, tal como análise de sentimento. A biblioteca utiliza uma base de dados de reviews de filmes que foram classificadas como positivos e negativos. A base de treino da biblioteca é classificada por um classificador Naive Bayes. É válido ressaltar que o TextBlob funciona somente para a língua Inglesa. Além dessas bibliotecas, foi utilizada a biblioteca Matplotlib para a apresentação dos resultados.
O tratamento dos dados segue o seguinte fluxo: a partir da coleta dos dados, empregamos o uso expressões regulares para remover menções, URLs e espaços em branco. As hashtags são substituídas pelas palavras que elas carregam. A partir de então, o tweet passa por um processo de tokenização e as stopwords são removidas. Desse modo, é calculada a polaridade de cada tweet definindo se este é positivo, neutro ou negativo. Por fim, os resultados são apresentados. Foram coletados cem tweets em cada teste, isto é, para cada hashtag.

4. Testes Experimentais

Resultados dos testes feitos no dia 08 de maio de 2018. As hashtags utilizadas estavam nos trending topics nos seguintes país: EUA, Reino Unido, Irlanda e Austrália.

4.1 #LethalWeapon (EUA)

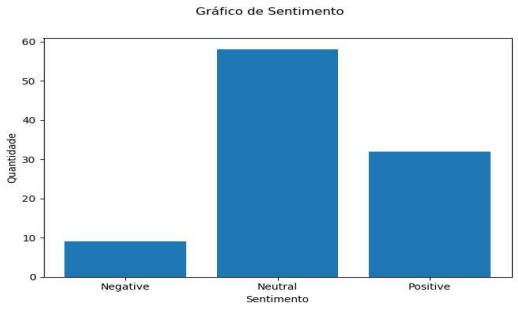

Figura 1. Gráfico de Sentimento

Positive: RT @LadyTri86: Trish is safe, thank God!!! #LethalWeapon
['trish', 'safe', 'thank', 'god', 'lethalweapon']

Neutral: CAST THIS MAN NOW. #LethalWeapon https://t.co/orVqRDJleI
['cast', 'man', 'now', 'lethalweapon']

Negative: RT @LethalWeaponFOX: Not something you hear everyday... #LethalWeapon
['something', 'hear', 'everyday', '...', 'lethalweapon']

4.2 #Newsnight (Reino Unido)

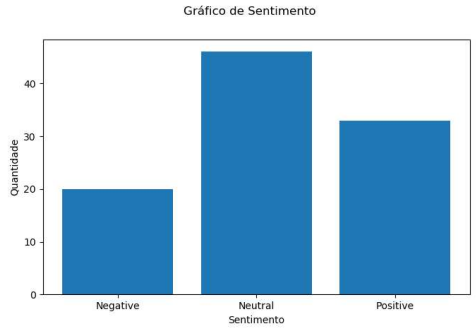

Figura 2.

Positive: RT @damocrat: Ken Clarke on Customs Union: This will decide how successful the British economy is for the next generation or two. Young p
['ken', 'clarke', 'customs', 'union', 'will', 'decide', 'successful', 'british', 'economy', 'next', 'generation', 'two', 'young', 'p']

Neutral: RT @TravisRuger: We have money for healthcare, education and infrastructure. Both corporate bought warmongering parties keep using it to p
['money', 'healthcare', 'education', 'infrastructure', 'corporate', 'bought', 'warmongering', 'parties', 'keep', 'using', 'p',]

Negative: RT @Mr_JDTraynor: Gorka on the TV is not "balance" #Newsnight. It is just idiotic. He is a racist, a thug, a liar, a professional conman.
['gorka', 'tv', 'balance', 'newsnight', 'just', 'idiotic', 'racist', 'thug', 'liar', 'professional', 'conman']

4.3 #Eurovision (Irlanda)

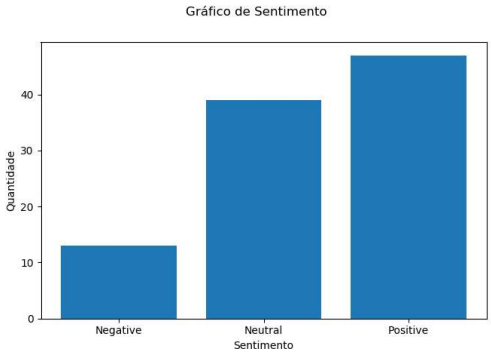

Figura 3.

Positive: RT @EquinoxBulgaria: Thank you all for you support! #EUROVISION is a dream come true for all of us and we so proud to perform for #Bulgaria
['thank', 'support', 'eurovision', 'dream', 'come', 'true', 'us', 'proud', 'perform', 'bulgaria']

Neutral: Having @VanessaVanjie nominated at the #Eurovision is a gift.
['nominated', 'eurovision', 'gift']

Negative: @TuckerCarlson #IranDeal || #Eurovision 10 antiterrorism personnel are killed during the ongoing confrontations with https://t.co/UzmmkryoMN
['irandeal', 'eurovision', '10', 'antiterrorism', 'personnel', 'killed', 'ongoing', 'confrontations', 'wit']

4.4 #changetherules (Austrália)

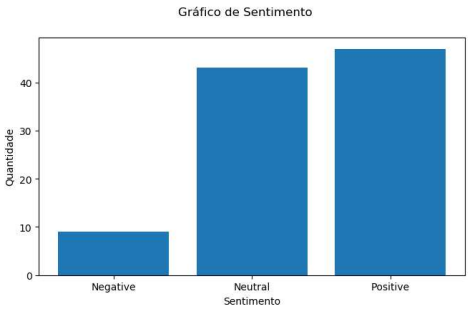

Figura 4.

Positive: RT @P_Gooding: Pretty extraordinary scenes in Melbourne. #ChangeTheRules https://t.co/PgpVKhLRJw
['pretty', 'extraordinary', 'scenes', 'melbourne', 'changetherules']

Neutral: RT @WeFightForFair: We are here to #ChangeTheRules https://t.co/IIefM2Ma3P
['changetherules']

Negative: RT @lhilakari: Melbourne is a union town. This is what 120,000 people on the streets look like. The system is broken. Workers are standing
['melbourne', 'union', 'town', '120,000', 'people', 'streets', 'look', 'like', 'system', 'broken', 'workers', 'standing']

5. Considerações Finais

O trabalho buscou evidenciar a relevância das práticas de PLN, sobretudo, empregado técnicas para fazer a análise de sentimento dos trending topics em alguns países de língua Inglesa. Os resultados foram satisfatórios, mas não perfeitos de um modo geral, dado que a

língua é flexível e abstrata, possui expressões inerentes a cada povo o que torna a classificação um tanto difícil.

6. Bibliografia

TWITTER SENTIMENT ANALYSIS USING PYTHON. GEEKSFORGEEKS. Disponível em: <https://www.geeksforgeeks.org/twitter-sentiment-analysis-using-python/> . Acesso em 28 Mar. 2018.

MINING TWITTER DATA WITH PYTHON. MARCO BONZANINI. Disponível em: <https://marcobonzanini.com/2015/03/02/mining-twitter-data-with-python-part-1/>. Acesso em 28 Mar. 2018.

STEP-BY-STEP TWITTER SENTIMENT ANALYSIS: VISUALIZING UNITED AIRLINES' PR CRISIS. VICKY QIAN. Disponível em: <http://ipullrank.com/step-step-twitter-sentiment-analysis-visualizing-united-airlines-pr-crisis/>. Acesso em 28 Mar. 2018.

Turney, P.D. 2002. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews