Resumo 16 – PLN

Continuando com as aulas de aplicações de Processamento de Linguagem Natural, neste décimo sexto encontro os dois assuntos tratados são o de análise de sentimento e de similaridade entre ementas nos cursos da UFABC.

Introduzindo os conceitos, o professor Jesús mostrou alguns exemplos de avaliações de um livro de Elon Musk. Aparentemente, as críticas são positivas e, para nós humanos, é extremamente fácil identificar o que leva uma análise a ser considerada positiva. Mesmo levando em conta a subjetividade dos critérios, termos como “alta excelência”, “ótima tradução”, “muito bom”, entre outros, funcionam como agentes positivos dentro do contexto.

Por outro lado, também existem termos que podem ser considerados negativos, como “decepcionante”, “abaixo das expectativas”, “muito ruim”. Além disso, podem existir análises que possuem indícios de termos positivos e também negativos com, inclusive, alguns termos e gírias características do país ou região.

Análise de sentimento é uma área extremamente popular no campo de PLN, obtendo um crescente número de publicações, principalmente nos últimos 3 anos. Um excelente exemplo mostrado pelo professor envolve a análise de sentimentos para o mercado de ações, onde algoritmos são aplicados para auxiliar na tomada de decisão de investidores, levando em consideração o humor e o sentimento de pessoas importantes para o mercado e seus efeitos na bolsa de valores. Comentários em redes sociais são analisados com o objetivo de tentar prever os impactos financeiros e as mudanças nos rumos do mercado.

Outro exemplo mostrado é adotado pelo Google Product Search e traz análise de sentimentos gerada automaticamente para indicar atributos/aspectos e o respectivo sentimento coletado para cada um dos produtos. Entre os sinônimos, um termo importante que é bastante difundido é a “Mineração de Opinião”. Um site muito interessante recomendado pelo professor chama-se carrot2 que, ao receber um texto como parâmetro, gera um grafo com todos os termos similares onde a área indicam a quantidade de termos presentes em publicações. É provável que tenha uma API para consumir essa análise. O Google Trends também é uma ótima ferramenta para análise.

São muitas as aplicações envolvendo Análise de Sentimentos. Entre elas, opinião de filmes, avaliação de produtos e serviços, opiniões em redes sociais, etc. Um outro exemplo muito interessante mostrado pelo professor refere-se a uma análise (utilizando o Twitter), o quão feliz era o público de acordo com dias da semana e datas comemorativas. Exemplos de “picos de felicidade” foram encontrados em datas como Dia dos Namorados, Natal, Ano Novo, entre outros. “Dias tristes” tiveram ápice em eventos como terremoto no Japão, 11 de setembro, etc. Fonte: Temporal patterns of happiness and infrmation ina global social network: Hedonometrics and Twitter de DODDS, Peter Sheridan et al. Publicado em 2011.

São subtarefas em Análise de Sentimentos: Classificação de sentimentos (utilizando o clássico algoritmo de Naive Bayes), Geração de léxicon (associar a palavra a um sentimento positivo, negativo ou neutro), quantificação de sentimento (estima a prevalência de sentimentos) e extração de opinião (dada uma opinião em uma sentença, identifica o titular da opinião, seu objetivo, sua polaridade, a força dessa polaridade e o tipo de opinião: A frase é de quem, falando sobre quem e em qual grau).

Utilizando um exemplo onde há duas possíveis classificações de sentimento, sendo (1) positivo e (0) negativo, o professor trouxe uma base de dados do UCI contendo 3000 linhas associadas a avaliações do imdb, Amazon e yelp, sendo 500 análises positivas e 500 análises negativas de cada um deles. Nome do dataset: Sentiment Labelled.

O professor então demonstrou uma implementação prática de um classificador de Bayes em duas classes (positiva e negativa) e mostrou técnicas para avaliação do desempenho em problemas de classificação (Acurácia, Precision, Recall ou Sensibilidade).

Um dos grandes problemas ao falar de análise de sentimentos é o da negação. Comparando as frases “eu não gostei muito desse filme, mas...” e “eu realmente gostei muito desse filme”. Quando é feita a análise em n-gramas, as palavras são separadas e, embora a palavra “gostei” esteja ao lado da palavra “não”, estas são tratadas de forma separadas. O grande desafio é replicar a palavra “não” para todas as palavras da frase, atrelando esta ao significado geral da frase. Em um artigo de 2001 de CHEN, a proposta mostra a adição do termo “não\_” a todas as palavras da frase até encontrar um sinal de pontuação. Somente com isso já é possível melhorar a performance do modelo. Também é possível expandir essa análise para outros termos como “not”, “didn’t”, “don’t”, “no”, entre outros.

Na segunda etapa da aula, foram analisados casos de similaridade entre ementas da UFABC como forma de aprendizado.