2
Pesquisas Anteriores

Na área de análise de sentimento existem muitas pesquisas voltadas para classificar emoções e opiniões. Em 2002, o artigo de Pang e Lee (12) introduz um dataset de críticas de cinema, que é um benchmark da área. Pang e Lee (13) apresentam os primeiros experimentos nesse corpus utilizando Naive Bayes e SVM (19). Seus resultados na classificação de sentimento de textos apenas como positivos ou negativos são bastante satisfatórios para a época. Em 2004 (14), elas apresentam resultados adicionais em torno desse corpus, porém desta vez utilizando uma abordagem de classificação por sentença. Em 2008, Abbasi, Chen e Salem (01) apresentam o melhor resultado até então publicado sobre esse corpus. Este trabalho está centrado em torno de seleção de atributos. Para isto, esses autores introduzem um algoritmo genético ponderado por entropia, o *Entropy Weighted Genetic Algorithm* (EWGA). Esse algoritmo utiliza o ganho de informação para ponderar cada atributo, sendo esses pesos incorporados tanto na população inicial, como no momento do crossover e no da mutação.

Muitos estudos foram feitos para examinar a relação entre notícias do mercado financeiro e as ações na bolsa, porém essa relação é muito complexa. Desde 1998, diversos pesquisadores tentam prever o movimento da bolsa fazendo uma análise automática do noticiário. Wüthrich (22) apresenta predições diárias considerando apenas o léxico e usando Naive Bayes, Nearest Neighbor e Redes Neurais como algoritmos de Aprendizado de Máquina. Lavrenko (09) melhora as predições intra-dia a partir de notícias do *YAHOO!Finance* utilizando TF-IDF como seletor de atributos.

Em 2001, Gidófalvi (08) apresenta fortes evidências indicando que as notícias influenciam o mercado apenas dentro de uma janela de 20 minutos. Para fazer suas predições ele utiliza Mutual Information (MI) como seletor de atributos.

Em 2008, Généreux (07) apresenta os melhores resultados para essa tarefa, alcançando uma acurácia de 75%. Ele classifica automaticamente as notícias cujos ativos variavam em mais de 6% de um dia para o outro. Ele utiliza diferentes tipos de atributos: unigramas de substantivos, verbos, adjetivos e advérbios ; lista de termos financeiros; lista de agentes metamorfos; e orientação

semântica de Osgood. Esses atributos são então selecionadas por Ganho de Informação (IG) e representados por TF no lugar do clássico modelo binário de presença.

Tetlock (18) mede a interação entre a mídia e o mercado, usando notícias do *Wall Street Journal*. Seus resultados mostram que o sentimento negativo ajuda a prever quedas no valor das ações em até 4 semanas posteriores a liberação da notícia.

Devitt e Ahmad (04) exploram a classificação do sentimento de notícias sobre duas empresas aéreas e mostram que a classificação é consistente para as pessoas que classificaram o dataset. Ou seja, a opinião pessoal e profissional sobre a empresa influencia no momento da classificação do texto, assim como o fato da pessoa trabalhar ou não no mercado de compra e venda de ações. Eles utilizam o *SentiNetWord* e não obtém melhoras com a inclusão de informações morfossintáticas. Seus resultados se apresentam melhores para classificação de notícias positivas, porém não ultrapassam 80% de F-Score.

Schumaker (16) apresenta um estudo para investigar se notícias realmente influenciam o movimento do preço de ações. Para notícias subjetivas, a inclusão da informação do sentimento melhora em 17%, chegando a uma acurácia de 59% na predição da direção. Seu estudo também indica que prever a direção é mais fácil para notícias negativas, sendo obtida 1% de melhora quando se conhece o sentimento. Estes resultados indicam que o mercado reage mais a notícias negativas do que a notícias positivas. Eles utilizam uma representação por Substantivo Próprio no lugar do clássico saco-de-palavras e *Support Vector Regression*(SVR) como algoritmo de aprendizado.