**Ludificação: experiências para construção e marcação
de um *corpus* para Análise de Sentimentos**

**José Saias**

**Dep. Informática – ECT – Universidade de Évora**

jsaias@uevora.pt

Aferir o grau de satisfação da população relativamente a um produto, ou entidade, é uma tarefa muito valorizada no marketing, e não só, pelo valor que essa informação contém, para quem decide o rumo de uma campanha, ou para gestores de imagem/reputação online de marcas e celebridades.

A ampla visibilidade das publicações em redes sociais, ou no espaço de comentários de jornais online, torna estas plataformas num repositório de dados com elevado potencial. Para além das publicações que isoladamente interessam a uma ou outra entidade, há o potencial da informação agregada, onde é possível encontrar tendências e padrões sobre os aspetos mais valorizados ou com maior carga pejorativa.

A Análise de Sentimentos (por vezes referida como *Opinion Mining*) lida com a busca de opinião em texto, procurando automaticamente manifestações subjetivas de opinião, ou sentimento (positivo, negativo ou neutro) [1]. Estes sentimentos podem ser abstratos, ou não dirigidos, evidenciando algo apenas sobre o estado de espírito do autor (Exemplo: “E*sta situação revolta-me!*”). Podem também ser concretos, e incidirem sobre um aspeto específico de alguma entidade alvo (Exemplo: “*O restaurante X tem comida fantástica, mas preços proibitivos!*”).

Análise de Sentimentos é, portanto, um processo complexo. Tem dificuldades que decorrem da natureza ambígua da linguagem natural escrita. Há casos especialmente desafiantes, envolvendo anáforas, uso de jargão e expressões idiomáticas específicas de um tema ou grupo, ou casos onde se emprega sarcasmo, nos quais mesmo um leitor humano pode ter uma interpretação incorreta. Por outro lado, o modo de escrita em redes sociais é bastante informal. Por vezes combina frases mal estruturadas com abreviaturas e símbolos, o que torna menos eficaz o uso de ferramentas de análise linguística.

É comum o uso de técnicas de Aprendizagem Automática supervisionada para determinar a polaridade do sentimento expresso face à entidade alvo, ou relativamente a algum aspeto da mesma. Isto implica que, para treinar uma rede neuronal, ou outro classificador, existe uma coleção de instâncias previamente marcadas com a resposta pretendida, que aqui corresponde à polaridade.

A Ludificação (do Inglês *Gamification*) [2] é a adoção de técnicas de jogos e dinâmicas próprias do funcionamento de jogos, em processos que não têm essa natureza. Em geral, o objetivo é facilitar a realização de uma tarefa, ou incentivar utilizadores a experimentar um sistema, podendo incutir elementos de retribuição ou competição, como estímulo para o utilizador continuar/aprofundar a sua participação.

Este trabalho descreve o uso de ludificação num sistema para a construção de um corpus, ou coleção de textos anotados, ou marcados, em Português, e que posteriormente servirá de apoio num sistema de Análise de Sentimentos, em particular, na fase de classificação automática da polaridade do sentimento. A ideia base do sistema é combinar elementos resultantes da atividade de múltiplos utilizadores. Esta atividade tem sessões, que consistem em três passos: inserir um texto curto; efetuar a respetiva marcação de entidades, aspetos e polaridades; e validar ou corrigir as marcações do sistema para outra instância existente que é escolhida pelo sistema. Fora do âmbito deste documento, fica a descrição dos processos para reconhecimento de entidades, de aspetos e de classificação de polaridade. O sistema de Análise de Sentimentos usado para confrontação das marcações do utilizador é o sistema sentiue [3,4].

Em termos de ludificação, identificamos dois desafios, ou fatores de potencial estímulo para o utilizador. Por um lado, descobrir um texto em que o sistema falha, valorizando a superior capacidade de interpretação textual do próprio. Adicionalmente, a retribuição direta de pontos ao utilizador em função da sua atividade.

O esquema de pontuação é baseado no contributo estimado para o sistema. Por contributo, entende-se a inserção e marcação de um texto que efetivamente acrescente valor à base de conhecimento. Isto pode corresponder a uma tipologia nova para o sistema, e onde o sistema não conseguia responder acertadamente.

Um critério a observar na apreciação destes casos é a validação cruzada. Não basta um texto onde o utilizador inicial faz uma marcação diferente da atribuída automaticamente pelo sistema, para concluir que o sistema estava errado. Para despistar erros (acidentais ou premeditados), a marcação humana, manual, segue um princípio de validação cruzada, sendo considerada estável se: há pelo menos mais uma marcação igual, de outro utilizador, e nenhuma diferente; ou se há alguma diferente, mas existe uma maioria de pelo menos 75% em concordância.

O sistema encontra-se atualmente em fase de testes, tendo recebido contribuições de 3 utilizadores apenas. Em breve anunciaremos a versão oficial e temos interesse na sua colaboração, para efeitos de divulgação e principalmente para interação com o sistema e contributo para o avanço do estado da arte nesta área, onde não há muitos recursos publicamente disponíveis para o Português.

**Palavras-chave:** Informática, Processamento de Língua Natural, *Gamification*, Análise de Sentimentos

**Referências**

[1] Pang, Bo & Lee, Lillian (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, Vol. 2(1-2), pp. 1-135

[2] Kim, B. (2015). Understanding Gamification. Library Technology Reports, 51(2), 1–35

[3] José Saias (2015). Sentiue: Target and Aspect based Sentiment Analysis in SemEval-2015 Task 12. In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Denver, Colorado, USA. June 2015. p. 767-771, ACL

[4] E. Dovdon and J. Saias (2017). “ej-sa-2017 at semeval-2017 task 4: Experiments for target oriented sentiment analysis in twitter,” in Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), (Vancouver, Canada), pp. 635–638, Association for Computational Linguistics