10
Conclusão

Neste trabalho , investigamos o problema de análise de sentimento para notícias jornalísticas do mercado financeiro. Nosso objetivo é classificar as notícias como favoráveis ou não em relação à empresa Petrobras.

Introduzimos o PETRONEWS, um corpus anotado de notícias em português sobre a Petrobras. Nossa abordagem filtra apenas as frases sobre a petrolífera em cada notícia do corpus. Em seguida, adicionamos atributos sintáticos e estruturais e aplicamos diferentes métodos de seleção dos mesmos. Finalmente, um comitê dos melhores modelos é construído para chegarmos ao melhor resultado.

Nossos resultados indicam que filtrar as frases no tópico melhora reduz o erro em $17,7\%$. A utilização de atributos estruturais em conjunto com o dicionário de palavras não apresenta grande diferença em nossos resultados, porém eles melhoram um pouco a curácia e por isso estão presentes nos melhores comitês.

Modelos com atributos morfossintáticos e chunks tem acurácia pior. Porém, esses modelos como membros de um comitê ajudam a melhorar a acurácia do comitê. Isso indica que esses atributos são mais informativos para um grupo específico de notícias e por isso esse modelo é um especialista para esse grupo.

Entre os métodos de seleção de atributos PD e Fisher Score combinados se destacam em relação aos demais. A seleção por árvores de decisão não melhorou a qualidade dos modelos testados. Da mesma maneira, utilizar a informação dos caminhamentos na árvore como novos atributos também não favoreceu o preditor, piorando o resultado. A seleção por EWGA é satisfatória porém muito demorada e por isso não compensa utilizá-la.

A criação de um o comitê de modelos utilizando atributos estruturais e sintáticos de maneira diferenciada e aplicados a uma combinação de métodos de seleção desses atributos diminui o erro em $25,40\%$ quando comparado ao modelo clássico de saco-de-palavras. O nosso comitê final atinge uma acurácia de 87.14%. Nesse comitê estão os melhores modelos, todos eles utilizam o filtro por tópico, os atributos estruturais, 1,2,3-gramas de palavras e a seleção

de atributos PD e Fisher Score. Além dessas características, o segundo modelo utiliza bigramas da anotação morfossintáticas, enquanto que o terceiro modelo utiliza informações morfossintáticas e Chunks combinadas às palavras.

Para tentar melhorar ainda mais nosso classificador, planejamos utilizar atributos originados da informação de análise de dependência, papel semântico e entidades nomeadas. Estamos criando outros corpus sobre outras companhias com ações na Bovespa, são elas a Vale, Gerdau, Bradesco e Itau. O custo para criar preditores para outros empresas é muito pequeno pois esse trabalho não depende do domínio. A dificuldade está apenas em criar um novo corpus para cada empresa, ou seja, anotar manualmente no mínimo 1.000 notícias sobre cada uma delas. Em trabalhos futuros, planejamos ainda utilizar a informação do sentimento da notícia na construção de *traders* das ações na bolsa de valores.