Eliminação de ruído em dados coletados na rede social online Twitter

João Vitor Néias de Carvalho[1] ,Déborah Silva Alves Fernandes[2], Márcio G. C. Fernandes [3]
[1]Estudante, Instituto de Informática, agsuns@discente.ufg.br
[2]Docente, Instituto de Informática, deborah.fernandes@ufg.br
[3]Docente, Universidade Estadual de Goiás, marcio.giovane@ueg.br

Resumo

A análise de sentimentos em grandes bases de dados, como por exemplo mensagens postadas em Redes Sociais Online (RSO), tem sido explorada para a realização de predições sobre o mercado financeiro. Um dos problemas relacionados à obtenção de indicadores a partir de tais bases, é a quantidade de conteúdo ruidoso presentes nestas. O objetivo deste trabalho foi estudar e caracterizar o ruído presente em uma base de tweets de domínio de mercado financeiro e desenvolver uma ferramenta que auxilie na remoção e filtragem destes. Para tal, foi elaborado um protocolo de testes de filtragem para acompanhar a aderência dos resultados à expectativa definida. O experimento apontou que é possível obter melhores resultados de classificação automática de mensagens estudando formas de filtragem de ruídos.

Apresentação

As redes sociais online (RSO) são repletas de opiniões sobre os mais diversos assuntos, tornando pública as intenções e posturas dos usuários que as utilizam. Devido ao potencial que oferecem, as RSO, em especial o Twitter, têm sido foco de estudo de pesquisadores [3] que fazem uso da análise de sentimentos para extraírem indicadores do mercado de investimentos[4 - 7]. O Twitter, assim como outras RSO, apresenta desafios quando se trata de remoção de ruídos, especialmente por haver um número limitado de pesquisa sobre o tema [1]. Portanto, o estudo sobre dados ruidosos se mostra importante para a obtenção de

melhores resultados ao efetuar-se análises de sentimentos em grandes volumes de dados.

Neste trabalho, propõe-se realizar o estudo, implementação e análise de uma ferramenta para eliminação de ruídos em dados coletados da rede social Twitter, dando continuidade aos estudos realizados em [13]. Para tal será utilizada uma base de tweets com conteúdo sobre o mercado de bolsa de valores brasileiro, onde serão realizados diversos testes de filtragem de tweets e medições de classificações binárias.

Na Tabela 1 é apresentado um comparativo dos artigos resultantes do levantamento bibliográfico realizado para a pesquisa.

Metodologia

Fundamentação teórica

A) Ruído

Segundo o dicionário Michaelis [14], ruído é um som constituído por vibrações acústicas que não possuem harmonia entre si, ou seja, um som indesejado, algo desagradável. Analogamente, pode-se considerar que qualquer elemento que interfira na comunicação entre um emissor e um receptor durante uma transmissão é considerado ruído.

Ruído em dados de Redes Sociais Online (RSO), dentro do campo de análise de mensagens e sentimento, obtenção de indicadores, classificação de postagens e consequentemente processamento natural de linguagem, pode ser entendido como quaisquer mensagens ou itens de uma mensagem que não se aplicam ao contexto analisado.

Uma característica relevante de dados de RSO é que estes são repletos de ruídos, sejam eles provindos de erros cometidos na fase de coleta dos dados, por falha humana quanto aos dados em si, ou por diversas outras situações difíceis de lidar. Tal característica implica diretamente na qualidade da informação que deseja-se obter.

Artigo	Objetivos	Método de remoção de ruído	Resultados
[15]	Indicar correlação entre a queda ou o aumento de preços no mercado financeiro e o sentimento do público em relação à empresa em questão através de tweets.	1ª etapa: Tweets filtrados através de palavras-chave. 2ª etapa: Tokenização, Remoção de palavras vazias e uso de expressões regulares para a remoção de caracteres especiais.	Coletaram 250.000 tweets. Comprovaram a existência de uma comunidade financeira na rede social Twitter.
[4]	Explorar os tweets da população filipina e seus possíveis efeitos significativos no índice do mercado de ações.	1ª etapa: Tweets filtrados através de palavras-chave, hashtags e geolocalização. 2ª etapa: Tradução de todos os tweets para inglês, e tokenização dos tweets.	Coletaram 800.500 tweets. Provaram que apenas os tweets provenientes do filtro de geolocalização e veículos de comunicação proveram melhores resultados em comparação com tweets oriundos de hashtags.
[5]	Investigar se o sentimento público demonstrado em redes sociais é correlacionado ou de fato prediz valores no mercado financeiro.	Os dados foram coletados usando a API do Yahoo e também a ferramenta de busca do twitter. Utilizaram somente as mensagens dos tweets e seus respectivos timestamps para a filtragem. Realizaram tokenização, remoção de stopwords e tratamento de símbolos.	Os resultados apontaram que mudanças na opinião pública afetam o mercado financeiro.
[16]	Prever o total de vendas de automóveis em uma escala mensal levando em consideração o ajuste de variáveis sazonais.	Os dados foram obtidos do Bureau of Economic Analysis, e também do Yahoo Finances. Os tweets foram coletados através de keywords. Coletaram 6 milhões de tweets que passaram por processos como remoção de tweets idênticos, URLs e símbolos.	Concluíram que o uso de fontes híbridas de dados aumentava a acuracidade de resultados de previsão de vendas de carros, e que o uso de uma técnica mais sistematizada de seleção de palavras-chave na busca de tweets pode ajudar na obtenção de resultados ainda mais precisos.
[7]	Predizer valores para o mercado de ações baseado na detecção de opiniões no twitter.	Os dados foram coletados do Yahoo Finance e do Twitter streaming API usando hashtags que se referem a empresa estudada. Somente tweets em inglês foram utilizados, retweets foram removidos. Depois disso cada mensagem foi salva em um modelo bag-of-words.	Concluíram que existe uma correlação entre informação nas redes sociais e o mercado de ações. Provaram também que o uso de hashtags para a filtragem garante uma melhor precisão de resultados quando comparado com o uso do próprio nome da empresa.
[17]	Investigar os impactos das mídias sociais no mercado de ações.	Os tweets foram coletados apenas de contas verificadas e com mensagens com conteúdo relacionado ao mercado de ações. A rotulação foi feita por quatro pessoas diferentes o resultado se deu pela média de todos. Utilizaram tokenização e modificação de um dicionário léxico para a melhor avaliação de sentimento de acordo com o contexto do mercado financeiro.	Mostraram que existe uma causalidade entre sentimentos no Twitter e o mercado de ações e também que tweets extraídos de contas verificadas no twitter em datas específicas acompanham os picos nos preços do mercado de ações.
[18]	Medir o sentimento público refletido no twitter, e testar o seu efeito no mercado de ações.	Remoção de spam, identificação de buzzwords, remoção de stopwords, extração da raiz das palavras, conversão de palavras em maiúsculas para minúsculas e também a normalização de palavras.	Concluíram que existe uma conexão entre a opinião pública e o mercado financeiro.
[3]	Predizer valores individuais do mercado de ações usando como entrada de dados tweets relacionados com a empresa estudada.	Coletaram 1.535.167 tweets em inglês que explicitamente mencionaram as empresas estudadas. Foram removidos os nomes de usuário dos tweets, assim como URLs, links e retweets.	Demonstraram que apesar de haver uma relação entre o sentimento público e os valores de mercado, as empresas com mais menções no Twitter possuem uma relação mais fraca entre sentimento e valores.

Tabela 1- Quadro comparativo das bibliografias estudadas.

A detecção de informações ou indicadores de sentimentos provindos da linguagem natural é uma tarefa complexa que requer uma base de dados adequada, de tal forma que seja possível dizer qual é o sentimento geral sobre um determinado assunto acuradamente. Essa tarefa deve ser feita levando em consideração tudo o que acompanha a linguagem natural, como sarcasmo, expressões, gírias, metáforas e etc. Por isso a filtragem de ruídos na base de dados é uma das atividades importantes em busca da maximização dos resultados de um classificador.

Ao investigar sobre sentimentos expressados em RSOs como o Twitter, no contexto do mercado financeiro, deseja-se obter informações que podem afetar a tomada de decisão dos investidores. Tais sentimentos, a respeito de empresas que atuam ativamente no mercado de ações, podem ser encontrados em tweets que possuem algumas características, tais como: repostagem de notícias de impacto notável em uma empresa, nível de satisfação com serviços oferecidos e popularidade medida por eventos importantes, opiniões de investidores atuantes [1].

Por outro lado, várias outras características podem ser usadas para descrever uma postagem ruidosa no âmbito de investimentos, tais como postagens que utilizam o nome de alguma empresa para fazer analogias dotadas de um tom cômico, menções de presença em estruturas vinculadas à alguma empresa, e anúncios de oferta de emprego.

B) Análise de sentimentos

O objetivo da análise de sentimento é processar e analisar sentimentos sobre entidades expressos em forma de texto [1]. As entidades podem ser serviços, organizações, produtos, eventos e outros.

A análise de sentimento pode ser realizada à nível de documento, de sentença ou aspecto. No primeiro, o objetivo é analisar se o texto como um todo possui um sentimento positivo, negativo ou neutro a respeito de uma determinada entidade. Não é aplicável para documentos que tecem opiniões sobre diversas entidades, já que a análise é feita considerando todo o documento como uma única unidade. À nível de sentença considera-se cada sentença do documento. À nível de aspecto

leva em consideração o alvo da opinião ao invés de olhar somente para o documento ou sentença, diferentemente dos outros dois modos já citados.

As técnicas de análise de sentimento podem ser classificadas em duas abordagens, aprendizado de máquina e a léxica. Os baseados em aprendizado de máquina podem ser subdivididos em:

- Supervisionados: Consiste em treinar a máquina com dados rotulados, e depois submetê-la a uma sequência de testes para que obtenhamos um resultado satisfatório de acordo com os dados rotulados. A maior limitação associada com esse tipo de aprendizado é a dependência que o mesmo possui de um conjunto considerável de dados rotulados de qualidade, pois caso os dados sejam insuficientes ou de má qualidade o treinamento pode falhar. Alguns métodos importantes utilizados nesse tipo de aprendizado são *Naïve Bayes*, *Support Vector Machine* e *K-Nearest Neighbors* [2].

- Não supervisionados: O aprendizado não supervisionado é recomendado quando a rotulação dos dados é difícil ou não viável, porém a coleta dos dados em si é fácil. O objetivo da máquina é agrupar dados de acordo com padrões, semelhanças e diferenças sem nenhum treinamento prévio. Um dos lados negativos desse tipo de aprendizado é a necessidade de grandes volumes de dados para ser treinado corretamente. Métodos bem conhecidos e utilizados são *Latent Dirichlet Allocation* (LDA) e *Probabilistic Latent Semantic Analysis* (PLSA) [2].

- Semi-Supervisionados: O aprendizado semi supervisionado deriva tanto como do aprendizado supervisionado como do não supervisionado. Nesse tipo, o aprendizado pode ocorrer simultaneamente de dados rotulados e de dados não rotulados, o que remove o problema visto anteriormente com o aprendizado não supervisionado, onde a pouca quantidade de dados provavelmente acarretaria em um aprendizado falho. Os métodos mais utilizados nessa forma de aprendizado são *self-learning*, modelos geradores, *co-training*, e métodos baseados em grafos [2].

Valendo-se de método científico, no qual pode-se apoiar de temas correlatos à essa pesquisa, utilizou-se das formas apresentadas da teoria de análise de sentimento para a classificação de características de mensagens em tweets ruidosos e não ruidosos.

C) Bag-of-Words

Os dados obtidos de RSOs são dados não estruturados, portanto faz-se necessária a submissão destes à pré-processamentos para serem utilizados como entrada em algoritmos de aprendizado. *Bag-of-words* é uma abordagem de pré-processamento que transforma os dados não estruturados em dados estruturados [8]. Esse método é uma representação do texto que descreve a ocorrência de palavras dentro do documento. Isso é feito definindo um vocabulário de palavras conhecidas e também medindo a frequência das mesmas. O objetivo é tornar cada documento de texto livre em um vetor que possa ser utilizado como entrada em algum algoritmo de aprendizado. A forma mais comum de se construir o vetor (que pode ter um tamanho fixo equivalente ao vocabulário) é marcar com um valor booleano que indica a presença ou a não das palavras do vocabulário naquele documento [11].

D) Teste de aderência (Qui-Quadrado)

Com a finalidade de se medir o progresso da acurácia de classificação do modelo treinado sobre os tweets filtrados, utilizou-se o teste de aderência Qui-Quadrado. O Qui-Quadrado é um teste de hipóteses que tem como objetivo comparar duas variáveis categóricas nominais, e verificar se existe homogeneidade entre elas. É um método independente de parâmetros populacionais como média e variância. Em suma, esse teste compara proporções, objetivando a identificação de discrepâncias entre as frequências observadas e as esperadas para um determinado evento.

Para a realização do teste de Qui-quadrado, algumas condições devem ser cumpridas, como: independência dos grupos; seleção aleatória dos itens de cada; as observações devem ser frequências ou contagens; cada observação pertence a uma única categoria e o tamanho da amostra deve ser de no mínimo 5 observações em cada célula, e no caso de poucos grupos, 10. Conforme será apresentado posteriormente, as amostras de dados selecionadas para o experimento se enquadram nas condições para a realização do teste.

O cálculo do Qui-quadrado é realizado por:

$$\chi^2 = \sum_{i=1}^{n} \frac{(o_i - e_i)^2}{e_i} \qquad \text{(Eq. 1)}$$

na qual o e e são as frequências observada e esperada, ambas para a classe. Para o teste de hipótese definiu-se:

- (H_0), na hipótese nula as frequências observadas são iguais às esperadas. Não há associação entre os grupos (casualidade).
- Como (H_1), uma Hipótese alternativa, as frequências observadas ≠ frequências esperadas.

Se χ^2 calculado ≥ χ^2 tabelado: Rejeita-se H_0. Se χ^2 calculado < χ^2 tabelado: Aceita-se H_0.

O qui-Quadrado foi escolhido para uso neste experimento, pelo fato das amostras se enquadrarem nas condições para o uso do teste. Esse teste foi utilizado para ser uma medida de referência do grau de acerto do classificador frente à escolha de filtros para remoção de ruído. Portanto, como resultado espera-se verificar o quanto o resultado de classificação adere à referência, que neste experimento são os dados rotulados manualmente (Figura 1 , item 04). Não há expectativa de alcance do limiar do valor tabelado do Qui-Quadrado para o grau de liberdade 1 e significância 20% [12].

Arquitetura do experimento

A Figura 1 apresenta um desenho da arquitetura proposta para o experimento realizado. Nela observa-se os componentes utilizados para a obtenção da análise dos resultados das filtragens realizadas para o estudo da remoção do ruído. No modelo de arquitetura é apresentada a base de dados, as amostragens para preparar o treinamento do classificador e para a realização dos testes de filtragem.

A) Base de dados

A base de dados é formada por tweets, postagens públicas da rede social online Twitter coletados por [ ] com 739.609 tweets obtidos entre agosto e dezembro de 2013. Como filtro para o coletor, foram utilizados os seguintes termos: ETR4, PETR3, ITUB3, ITUB4, VALE3, VALE5, GGBR3, GGBR4, USIM3, USIM5, CSNA3, BBAS3, OGX, OGXP3, MMXM3, MPXE3, "grupo x", "GRUPO X", "grupo gerdau", "GRUPO GERDAU", petrobras, petrobrás, itau, itaú, usiminas, "companhia

siderurgica nacional", "companhia siderúrgica nacional", "vale do rio doce", "banco do brasil", bovespa, "MSCI Standard", vale5, ENEVA. Os termos se relacionam à empresas brasileiras atuantes no mercado de bolsa de valores no ano de 2013, objeto de pesquisa em [13].

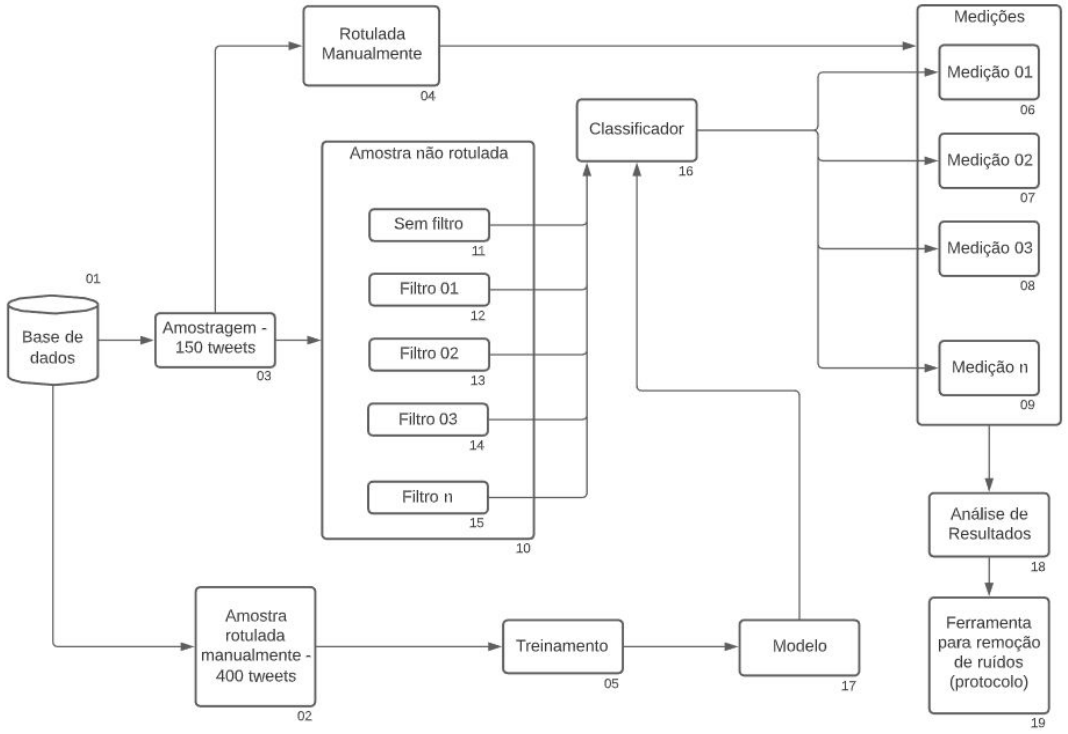

Figura 1 - Arquitetura do experimento

A.1) Amostra

Da base de dados foram selecionados aleatoriamente como amostra 400 tweets, sem presença de conteúdo em língua estrangeira, e sem referência às empresas do Grupo X (efetivas em 2013 e não mais na atualidade). Na Tabela 2 são apresentados alguns exemplos de tweets contidos na base utilizada para o experimento.

A.2) Rotulação

As mensagens foram rotuladas manualmente e separadas em dois conjuntos com 200 itens cada, o primeiro formado por tweets ruidosos e o segundo por não ruidosos. Um tweet ruidoso, no contexto desta pesquisa, foi caracterizado como um tweet escrito em língua estrangeira, que menciona o Grupo X e que não mostre qualquer relevância para investimentos no mercado de ações brasileiro. Os demais foram considerados não ruidosos.

A.3) Dados para teste

A partir da definição de ruído para o domínio de mercado financeiro, definiu-se para o experimento três tipos de tweets ruidosos: os que possuem língua estrangeira, os que possuem menções ao Grupo OGX e os que possuem conteúdo contido em uma *blacklist* desenvolvida por [13] (possui palavras e expressões recorrentes nos tweets associados ao mercado financeiro e cujo conteúdo não se relacionava ao contexto).

Tweet relevante para investimentos no mercado financeiro	"RT @RicardoBarrosao: Produção de petróleo da Petrobras no Brasil cai 4,6% em julho http://t.co/tJJT67hgAz"
Tweet não relevante para investimentos no mercado financeiro, mas que menciona empresas presentes no mercado financeiro	"RT @menino_mosso: Esses caras do banco do Brasil, vão se foderem."
Tweet em linguagem estrangeira	"$VALE5 - Brazil's Samarco Plans To Issue Overseas Bonds Worth $500 Million -Sources http://t.co/R4uGCvqt3t"
Tweet relacionado ao Grupo X	"Muita boataria do Eike... Enquanto isso quem vendeu 600 Milhoes de acoes OGX nos ultimos 5 minutos na sexta? #ogx misterio"

Tabela 2- Exemplos de tweets presentes na base de dados.

Foram selecionados 150 tweets da base de dados, de forma aleatória, e em seguida estes foram classificados manualmente. Esses tweets passaram por filtragem para pré-processamento anteriormente à classificação. Os filtros adotados são:

- Filtragem de tweets com menção ao Grupo X (grupo de empresas atuantes no mercado de ações em 2013, porém não na atualidade, por isso foram considerados ruidosos.
- Filtragem de tweets que possuíam elementos contidos na *blacklist* [13].
- Filtragem de tweets em língua estrangeira e que faziam menção ao Grupo X.
- Filtragem de tweets em língua estrangeira, que faziam menção ao Grupo X e que possuíam palavras contidas na *blacklist*.

Adotando a filtragem dos dados, foram obtidos 6 conjuntos a serem utilizados para teste, com inclusão do conjunto original sem aplicação de filtro.

B) Classificação Supervisionada

Neste experimento, o método de aprendizado adotado foi o supervisionado, à nível de documento. Com o objetivo de obter um classificador adequado para a classificação à nível de documento dos tweets da amostra, foram selecionados três modelos supervisionados: Regressão Logística, *Support Vector Machines*(SVM) e *K-nearest neighbors* (KNN).

Para a escolha do modelo mais adequado a este trabalho, analisou-se as seguintes métricas provindas do treinamento:

- AUC (*area under the curve*): Reflete a performance geral do classificador.
- Acurácia: Mede a razão entre número de instâncias classificadas corretamente e o número de instâncias avaliadas.
- F1-score: Harmonia entre Precisão e Revocação.
- Precisão: Proporção de identificações de classe corretas.
- Revocação: Mede a fração de padrões classificados corretamente [9].

A Tabela 3 apresenta os resultados obtidos pelo treinamento.

Método	**AUC**	**Acurácia**	**F1-score**	**Precisão**	**Revocação**
Regressão Logística	0,944	0,875	0,875	0,875	0,875
SVM	0,885	0,555	0,457	0,698	0,555
KNN	0,886	0,718	0,695	0,809	0,718

Tabela 3- Medições para o treinamento dos classificadores adotados .

Verificou-se que o modelo Regressão Logística apresentou melhores índices em comparação com os outros modelos, portanto, foi o modelo escolhido para o treinamento do classificador, onde utilizou-se a amostra mencionada em A.1 como entrada.

C) Arquitetura para testes

C.1) Arquitetura dos testes

Todos os conjuntos de tweets mencionados em A3 passaram por um pré-processamento, que tem como objetivo trabalhar os dados de uma maneira que facilite e torne viável a análise dos mesmos. Após a transformação os tweets passaram a conter somente caracteres em lowercase, e todos os acentos e URLs foram removidos. O conteúdo dos tweets também foram tokenizados, para que os algoritmos de análise de sentimento consigam processar as emoções humanas facilmente. Em seguida, as palavras vazias foram removidas. Por fim, os conjuntos para teste foram submetidos ao modelo *bag-of-words.*

Após o treinamento do classificador, os conjuntos foram classificados pelo modelo escolhido. A eficiência das classificações foram medidas usando o qui-quadrado.

O qui-quadrado foi usado como uma maneira de medir a divergência entre a distribuição dos dados esperados e dos dados classificados pelo modelo. Quanto menor o resultado do qui-quadrado calculado, mais aderente é a classificação obtida pelo modelo treinado com a classificação esperada.

C.2) Ferramentas

Para realizar os experimentos, a base de dados foi carregada em um sistema gerenciador de banco de dados - PostgreSQL 10.14. Para métodos de classificação (treinamento, classificação e validação), seleção de amostras, cálculo de *bag-of-words* foi utilizado o Sistema Orange 3.19.0 - Visualizador de dados e *machine learning*.

D) Protocolo - Ferramenta filtragem de ruído

A Figura 2 apresenta o protocolo obtido para experimento com a finalidade de avaliar a remoção de ruído nos dados.

Ferramenta para filtragem de ruído
(1) Definir o que é ruído no contexto. (2) Preparar classificador supervisionado (treinamento / validação). (3) Selecionar amostra de dados da base. (4) Rotular manualmente a amostra (3) em dados ruidosos ou não ruidosos . (5) Definir filtros a serem aplicados nos dados que possivelmente podem “melhorar” a classificação dos dados (ruidosos e não ruidosos). (6) Executar o classificador na amostra (3) com a aplicação dos filtros do passo 5. (7) Realizar o teste de Qui-Quadrado para medir o quanto os filtros definidos em (5) influenciaram no grau de aderência dos resultados classificados e os dados rotulados.

Figura 2 - Protocolo para filtragem de ruído em tweets.

Resultados e Discussão

Visto que os ruídos retirados da base de treinamento eram conhecidos, surgiu-se a hipótese de que ao aplicar o mesmo filtro, utilizado na limpeza da base, na amostra de treinamento, obteria-se um melhor resultado no teste de aderência.

A hipótese foi verificada através do seguinte experimento:

No item 5 - figura 1, tem-se a certeza que os ruídos (filtros do teste 01 e teste 02, Tabela 4), utilizados como filtro de uma forma controlada no item 3, não estavam

presentes. Desta forma, o modelo gerado no item 17 não está treinado para reconhecer esses ruídos.

Teste	ROLUTADO (referência)		Classificados (expectativa)		Qui-quadrado calculado
	Ruidosos	Não ruidosos	Ruidosos	Não ruidosos	
Teste 01 - Nenhum filtro foi aplicado	105	45	89	61	8,126
Teste 02 - Remoção de tweets escritos em língua estrangeira	74	45	61	58	6,039
Teste 03 - Remoção de tweets que fazem menção ao Grupo X	92	45	81	56	4,004
Teste 04 - Combinação dos testes 02 e 03	67	45	59	53	2,377
Teste 05 - Remoção de tweets que contêm palavras presentes na *blacklist*	57	44	42	59	9,061
Teste 06 - Combinação dos testes 04 e 05	35	44	27	52	3,283

Tabela 4 - Resultados dos testes realizados para qui-quadrado.

Ao aplicar-se os filtros 01 e 02 da Tabela 4, houve uma expectativa que os dados classificados no item 16 - figura 1 demonstrariam uma aderência significativa com os dados rotulados no item 4 - figura 1, já que um processo análogo foi realizado mas com os dados do item 11 - figura 1.

Analisando-se a Tabela 4, testes 2 e 3 demonstraram uma aderência significativa, principalmente quando combinados com os filtros do teste 4.

Na aplicação do filtro 5, que elimina ruído não controlado proveniente da *blacklist* obtida em [13], que considerou tal filtragem importante para a remoção de tweets que continham palavras e expressões fora do domínio do mercado financeiro, o experimento permite identificar uma piora na aderência com os dados rotulados. Manipulações aleatórias com o conteúdo da *blacklist* demonstraram variação para aumento e diminuição do grau de aderência. Após uma reflexão heurística percebeu-se que que características contextuais justificavam a variação obtida. Investigar as causas dessa variação fogem do escopo proposto ao experimento.

Dadas essas observações, pode-se concluir que o experimento alcançou seu objetivo, pois permitiu observar o desempenho de um classificador a partir da filtragem de dados hipoteticamente considerados ruidosos, além de gerar um protocolo de limpeza em dados ruidosos.

Conclusão de Considerações Finais

Neste trabalho, mostramos como a filtragem de ruídos pode afetar significativamente o resultado de um classificador. A remoção de língua estrangeira e remoção de tweets relacionados a empresas obsoletas mostraram-se satisfatórias na classificação, pois permitiu determinar de maneira mais eficiente quais tweets se configuram como ruidosos e quais não. O experimento apontou também que é necessário um estudo mais aprofundado sobre o uso e características da *blacklist*[13] para filtragem, pois apesar de ajudar a remover tweets ruidosos da base, não apresentou efetividade no uso do classificador, conforme resultados expostos na Tabela Y, o que torna evidente o motivo da disparidade da aderência da classificação esperada e classificação obtida em comparação com os outros testes.

Referências

[1] LIU, Bing. **Sentiment analysis: Mining opinions, sentiments, and emotions**. Cambridge university press, 2015.

[2] MADHOUSHI, Zohreh; HAMDAN, Abdul; ZAINUDIM, Suhaila. **Sentiment analysis techniques in recent works**. 2015 Science and Information Conference (SAI), IEEE p. 288-291, 2015.

[3] BERNARDO, Ivo; HENRIQUES, Roberto; LOBO, Victor. **Social Market: Stock market and twitter correlation.** International Conference on Intelligent Decision Technologies. Springer, Cham, p. 341-356, 2017.

[4] CALIÑGO, Anthony; SISON, Ariel; TANGUILIG III, Bartolome. **Prediction Model of the Stock Market Index Using Twitter Sentiment Analysis**. International

Journal of Information Technology and Computer Science(IJITCS), Vol.8, No.10, pp.11-21, 2016.

[5] KORDONIS, John; SYMEONIDIS, Symeon; ARAMPATZIS, Avi. **Stock Price Forecasting via Sentiment Analysis on Twitter**. PCI '16: Proceedings of the 20th Pan-Hellenic Conference on InformaticsNovember 2016 Article No.: 36 Pages 1–6, 2016.

[6] YANG, Steve; MO, Sheung; LIU, Anqi. **Twitter financial community sentiment and its predictive relationship to stock market movement**. Quantitative Finance, v. 15, n. 10, p. 1637-1656, 2015.

[7] ROMANOWSKI, Andrzej; SKUZA, Michał. **Towards predicting stock price moves with aid of sentiment analysis of Twitter social network data and big data processing environment.** Advances in Business ICT: New Ideas from Ongoing Research, Springer, Cham, pp.105-123, 2017.

[8] MATSUBARA, Edson; MARTINS, Claudia; MONARD, Maria. **Pretext: Uma ferramenta para pré-processamento de textos utilizando a abordagem bag-of-words.** Technical Report, v. 209, n. 4, 2003.

[9] HOSSIN, Mohammad; SULAIMAN, M. N. **A review on evaluation metrics for data classification evaluations**. International Journal of Data Mining & Knowledge Management Process, v. 5, n. 2, p. 1, 2015.

[10] MAGALHÃES, N. M. L. A. C. P. **Noções de Probabilidade e Estatística**. Edusp, São Paulo, Brasil, 2005.

[11] BROWNLEE, Jason. A Gentle Introduction to the Bag-of-Words Model. Machine Learning Mastery, Vermont, 09 de Ago. de 2017. Disponível em: <https://machinelearningmastery.com/gentle-introduction-bag-words-model/>. Acesso em: 30 de Ago. de 2020.

[12] CORREA, Ana; QUEIROZ, Eder; TREVISAN, Newton. Teste do Qui-Quadrado. UFPR. Disponível em: <http://www.leg.ufpr.br/lib/exe/fetch.php/disciplinas:ce001:teste_do_qui-quadrado.pdf>. Acesso em 30 de Ago. de 2020.

[13] ALVES, Deborah. **Uso de técnicas de Computação Social para tomada de decisão de compra e venda de ações no mercado brasileiro de bolsa de valores.** 2015**.**

[14] RUÍDO. In: Michaelis. Disponível em: <https://www.dicio.com.br/risco/>. Acesso em: 30/09/2020.

[15] PAGOLU, Venkata Sasank et al. **Sentiment analysis of Twitter data for predicting stock market movements**. 2016 international conference on signal processing, communication, power and embedded system (SCOPES). IEEE. p. 1345-1350, 2016.

[16] PAI, Ping-Feng; LIU, Chia-Hsin. **Predicting vehicle sales by sentiment analysis of Twitter data and stock market values**. IEEE Access, v. 6, p. 57655-57662, 2018.

[17] TABARI, Narges et al. **Causality analysis of twitter sentiments and stock market returns**. Proceedings of the First Workshop on Economics and Natural Language Processing, p. 11-19, 2018.

[18] SHEHADEH, Mohammad; KHRITANTSEV, Maksim. **How 140 Characters can be related to the Stock Market Movements: Sentiment Analysis of Twitter**. 2018.