A comparison of parameter selection measures for sensor learning from financial news events

Alex S. Farias[1], **Solange O. Rezende**[2], **Ricardo M. Marcacini**[1]

[1]Laboratory of Scientific Computing (LivES)
Federal University of Mato Grosso do Sul (UFMS)
Três Lagoas – MS – Brazil
http://lives.ufms.br/
astfarias@gmail.com, ricardo.marcacini@ufms.br

[2]Laboratory of Computational Intelligence (Labic)
Institute of Mathematics and Computer Science (ICMC)
University of São Paulo (USP)
http://labic.icmc.usp.br/
São Carlos – SP – Brazil
solange@icmc.usp.br

Resumo. *The popularization of web platforms promoted a significant increase in the publication of financial news and reports in digital media. In this sense, a multidisciplinary research area called "learning to sense" (or sensor learning) has received attention recently. Unlike traditional machine learning methods, in sensor learning there is an interest in obtaining a time series that indicates the activity of a particular topic over time. A sensor is represented by a set of parameters learned from a historical news events dataset. The sensor generates time series as news events are processed and these time series are used in decision support systems. This paper presents an overview of sensor learning for financial news. We compared six parameter selection measures for sensor learning, with the differential of considering an unsupervised scenario. The general idea is to use the concept of k-recurrent events, i.e, news events that are similar and occur together in different periods of up-trends and down-trends of a financial time series. Thus, if a specific event (extracted from news) occurred at least k times in the past always associated with up-trends, then such news is labeled as positive news. Analogously, it can be labeled as negative. The experimental results from real data provided evidence that the approach investigated in this work is a promising alternative for sensor learning from financial news events, especially in contexts where there are no domain experts or external information to label a training set.*

1. Introdução

Diversos estudos na área de ciências econômicas apontam que o desenvolvimento das sociedades modernas é dependente de suas economias de mercado [Schumpeter 2017]. Nesse sentido, pesquisas multidisciplinares envolvendo economia e ciência de computação tem estudado o comportamento dos mercados financeiros, seus equilíbrios de oferta e demanda, bem como a capacidade de identificar variáveis para prever ou entender este comportamento [Einav and Levin 2014]. No entanto, é reconhecido que a

natureza dos mercados é extremamente difícil de prever devido ao seu comportamento dinâmico e caótico [Nassirtoussi et al. 2014].

O sensoriamento do mercado financeiro e análises preditivas para ativos financeiros são geralmente divididas em análise técnica ou análise fundamentalista [Chandra 2017]. Estas análises são diferenciadas principalmente pelos tipos de dados utilizados. A análise técnica costuma utilizar dados históricos de mercado, como séries temporais com a cotação dos ativos financeiros, enquanto a análise fundamentalista utiliza qualquer outro tipo de informação ou notícias sobre o país, sociedade, empresa e o ativo financeiro. A maior parte dos estudos existentes foca em abordagens de análise técnica, principalmente devido à disponibilidade de dados históricos do mercado e pela maior facilidade em aplicar métodos computacionais [Taylor et al. 2014]. Já a análise fundamentalista é mais complexa de ser automatizada, uma vez que depende de análise de informação externa e dados textuais, como indicadores macroeconômicos, relatórios financeiros, leis governamentais, boletins e notícias. Os dados disponíveis no formato textual representam o aspecto de pesquisa mais desafiador [Mitra and Mitra 2011]. Se por um lado representam um conhecimento valioso, por outro lado estão em um formato não estruturado e que exige uma etapa de pré-processamento desses dados para permitir a extração de conhecimento.

Recentemente, a popularização de plataformas *web* promoveu um crescimento significativo da publicação de notícias e relatórios financeiros em meio digital. Esse cenário favorece a análise fundamentalista, pois cientistas de dados podem analisar notícias e acompanhar eventos de interesse, sendo uma fonte de informação para apoiar a tomada de decisão. Nesse contexto, as áreas de aprendizado de máquina e *big data* têm se tornado populares para apoiar esta análise. A ideia geral é que notícias e eventos representam em um mundo virtual os acontecimentos que ocorrem em nosso mundo físico e, dessa forma, os métodos de aprendizado de máquina têm sido propostos encontrar mapeamentos entre esses dois mundos [Radinsky and Horvitz 2013, Ackland 2013]. Uma das estratégias atuais é sensoriamento de notícias utilizando métodos de aprendizado de máquina, que neste contexto são conhecidos como "*learning to sense*" ou aprendizado de sensores [Radinsky and Horvitz 2013, Radinsky et al. 2013, Marcacini et al. 2017]. Ao contrário dos métodos tradicionais de aprendizado de máquina que produzem como saída classificadores, agrupamentos e regras, no aprendizado de sensores há o interesse de obter uma série temporal que indica a atividade de um determinado tema ao longo do tempo.

Um dos principais desafios para o aprendizado de sensores é definir um conjunto de parâmetros que representam o conteúdo a ser monitorado pelo sensor [Marcacini et al. 2017]. No contexto de notícias financeiras, esses parâmetros são atributos textuais, como palavras-chave e sua importância a respeito de um determinado tema de interesse. Assim, dado um ativo financeiro, o objetivo do sensoriamento é identificar notícias financeiras relacionadas às possíveis altas e baixas do valor da cotação deste ativo. Uma limitação das abordagens existentes nessa linha é a exigência de um conjunto de notícias rotuladas por especialistas da área, indicando quando uma determinada notícia afeta positivamente o valor da cotação (aumentando o valor) ou negativamente (reduzindo o valor). Com a existência desse conjunto de dados rotulados, diversas medidas estatísticas de seleção de atributos podem ser empregadas para extração dos parâmetros do sensor. No entanto, muitos estudos afirmam que tal tarefa de rotulação

manual pode ser ineficaz na prática, uma vez que os temas que afetam o valor da cotação mudam de forma constante, o que exigiria uma frequente rotulação dos dados [Chan and Chong 2017, Florence et al. 2017]. Além disso, muitas vezes não há um entendimento conclusivo a respeito dos temas que afetam um ativo financeiro.

Neste trabalho é realizada uma comparação de medidas estatísticas para seleção de parâmetros visando o aprendizado de sensores em notícias financeiras. Ao contrário das abordagens existentes, neste trabalho é considerado um cenário de rotulação automática de notícias a partir da série temporal do ativo financeiro. A ideia geral é utilizar o conceito de eventos k-recorrentes, ou seja, eventos de notícias que são similares e apresentam co-ocorrência em diferentes períodos de altas ou baixas. Assim, se um determinado evento (extraído de uma notícia) ocorreu pelo menos k vezes no passado sempre associado a valores de altas da cotação, então tal notícia é rotulada como *positiva*. De forma análoga, é rotulada como *negativa* se o evento ocorreu pelo menos k vezes no passado associado a períodos de baixa da cotação. Dada esta estratégia de rotulação, neste trabalho foram comparadas experimentalmente seis medidas para seleção de parâmetros na construção de sensores em um conjunto de notícias da Petrobrás, considerando a série temporal do seu respectivo ativo financeiro PETR4. Os resultados experimentais obtidos forneceram evidências de que a seleção de parâmetros de sensores para notícias financeiras pode ser realizada de forma não supervisionada por meio de critérios estatísticos, sendo potencialmente útil em contextos em que não há especialistas de domínio ou informação externa para apoiar tal tarefa. Assim, é possível apoiar cientistas de dados em análises econômicas fundamentalistas, além de considerar também as informações úteis provenientes da análise técnica.

O restante deste trabalho está organizado da seguinte maneira. Na Seção 2 são apresentados fundamentos básicos utilizados neste trabalho, tanto em relação aos conceitos econômicos que norteiam tarefas de sensoriamento do mercado financeiro, quanto em relação ao uso de notícias para apoiar esta atividade. Na Seção 3 é apresentada a abordagem proposta neste trabalho para comparação experimental de medidas de seleção de parâmetros no sensoriamento de notícias financeiras. A avaliação experimental e seus resultados são discutidos na Seção 4. Por fim, as considerações finais e direções para trabalhos futuros são apresentados na Seção 5.

2. Sensoriamento de Dados para o Mercado Financeiro

A produção de sensores para mercados financeiros é uma atividade multidisciplinar. Um recente trabalho que discute conexões entre aprendizado de máquina para dados textuais e mercado financeiro é apresentado por [Nassirtoussi et al. 2014], que afirma que são necessários pelo menos três áreas de conhecimento para fundamentar a pesquisa: Linguística (para compreender a natureza da linguagem), Aprendizado de Máquina (para possibilitar a modelagem computacional de reconhecimento de padrões) e Economia Comportamental (influência da psicologia humana nas decisões econômicas).

Nas próximas seções são discutidos os pontos principais descritos em [Nassirtoussi et al. 2014] denominados Hipótese do Mercado Eficiente, Economia Comportamental, Hipótese dos Mercados Adaptáveis, Previsibilidade dos Mercados, Análise Técnica *versus* Análise Fundamentalista e Análise de Sentimentos.

2.1. Hipótese do Mercado Eficiente e Hipótese dos Mercados Adaptáveis

A Hipótese do Mercado Eficiente (HME) propõe que os mercados são completamente aleatórios e não são previsíveis [Fama 1995], o que inviabilizaria seu monitoramento para apoiar investimentos. O autor desta hipótese afirma que os mercados financeiros são informacionalmente eficientes e que não é possível obter, de forma consistente, retornos acima dos retornos médios do mercado. Essa afirmação se dá pela existência de um ajuste do risco, dada a informação disponível no momento em que o investimento é feito. Esta hipótese não é completamente precisa e o próprio autor a revisou para incluir três níveis de eficiência como forte, semiforte e fraco [Fama 1970]. Isto indica que existem muitos mercados onde a previsibilidade é viável, que são denominados "fracamente eficientes". Foi demonstrado que a eficiência do mercado está correlacionada com a disponibilidade de informações e um mercado é apenas "fortemente eficiente" quando toda a informação está disponível, o que raramente acontece. [Fama 1970] admitiu que sua teoria é mais forte em certos mercados onde a informação é amplamente divulgada e instantaneamente disponível para todos, e fica mais fraca em mercados onde esta situação não ocorre.

Já na Hipótese dos Mercados Adaptáveis (HMA), a premissa geral é que os mercados tendem a serem eficientes e racionais até o surgimento de alterações (ou rupturas), ponto em que as previsões se tornam irracionais ou aleatórias [Majumder 2013]. Nesses períodos, os agentes econômicos que melhor se adaptarem à nova realidade são os que irão obter lucros. [Urquhart and Hudson 2013] conduziram uma investigação empírica usando dados de longo prazo sobre a Hipótese dos Mercados Adaptáveis (HMA) em três dos mercados de ações mais estabelecidos do mundo: os mercados dos EUA, Reino Unido e Japão. A pesquisa forneceu evidências de que a HMA fornece uma melhor descrição do comportamento dos retornos das ações do que o HME.

Tanto a Hipótese do Mercado Eficiente (HME) quanto a Hipótese dos Mercados Adaptáveis (HMA) são conceitos relacionados à hipótese da Previsibilidade dos Mercados, que considera que quando os mercados estão fracamente eficientes, deve ser possível determinar critérios com impacto preditivo sobre eles, o que motiva estudos para sensoriamento do mercado.

2.2. Análise Técnica e Análise Fundamentalista

A hipótese que argumenta que os movimentos históricos do mercado se repetem de tempos em tempos é tratada como Análise Técnica. A base deste estudo são padrões visuais em gráficos de mercado. Com base nessa crença, muitos dos movimentos gráficos recebem identificadores, o que forma a base da Análise Técnica. Em um nível mais alto, os analistas técnicos tentam detectar tais modelos matemáticos pelo uso de técnicas computacionais de reconhecimento de padrões. Embora as técnicas de Análise Técnica sejam mais difundidas, há pouco esforço em investigar e interpretar a existência de padrões. Algumas das técnicas comuns em análise técnica são as regras de média móvel, regras de força relativa, regras de filtro e as regras de quebra tendência [Nassirtoussi et al. 2014].

Na Análise Fundamentalista, os analistas observam dados que estão disponíveis em diferentes fontes, como relatórios financeiros e notícias, e fazem suposições com base nisso [Nassirtoussi et al. 2014]. Entretanto, utilizar tais informações para projetar de um ativo apresenta muita incerteza devido à subjetividade envolvida [Kaltwasser 2010]. Uma abordagem utilizada para facilitar o uso programático de dados não estruturados é rea-

lizar o pré-processamento para extrair dados estruturados e, então empregar métodos de aprendizado de máquina como algoritmos de classificação [Lupiani-Ruiz et al. 2011].

2.3. Análise de Sentimentos

Um tema popular para apoiar a previsão do mercado financeiro a partir de notícias é a análise de sentimentos. O objetivo é detectar a polaridade (positivas, neutra e negativa) existente no texto através de análise semântica especializada para uma variedade de propósitos [Liu 2012], por exemplo, para avaliar a qualidade da recepção do mercado para um novo produto e o *feedback* geral dos clientes, ou para estimar a popularidade de um produto ou marca entre as pessoas [Ghiassi et al. 2013, Mostafa 2013]. Há diversas pesquisas focadas na análise de sentimentos ou na chamada "mineração de opinião" [Nassirtoussi et al. 2014]. Esta metodologia se baseia principalmente na identificação de palavras positivas e negativas e processamento de texto com o objetivo de classificar sua postura emocional como positiva ou negativa.

Análise de Sentimentos em notícias também pode ser utilizada para a predição do mercado. [Schumaker et al. 2012] avaliaram o sentimento em artigos de notícias financeiras em relação ao mercado de ações em sua pesquisa, mas reportaram dificuldades relacionadas à obtenção de um conjunto rotulado de treinamento. [Yu et al. 2013] apresentaram uma abordagem para identificar um conjunto de palavras de emoção semelhantes e suas correspondentes intensidades das notícias do mercado de ações online. Isto foi realizado extraindo a importância de uma palavra dada um conjunto de palavras emocionais de referência. A medida estatística utilizada foi baseada em entropia, similar a um critério supervisionado para seleção de atributos. Em seguida, as palavras selecionadas são utilizadas para determinar a polaridade de cada notícia. Seus resultados experimentais mostram que o uso Análise de Sentimentos aumentou o desempenho de classificação das notícias e identificação de eventos importantes para o ativo financeiro. Também é interessante notar que a análise de sentimento não precisa ser meramente baseada na positividade-negatividade e pode ser feita em outras dimensões ou em multi-dimensões [Ortigosa-Hernández et al. 2012].

Estratégias de seleção de parâmetros para sensores de notícias financeiras investigadas neste trabalho têm relação com trabalhos de análise sentimento d. No entanto, os trabalhos existentes geralmente assumem a existência de uma lista de referência com palavras de emoção ou rotulação do sentimento das notícias por meio de especialistas de domínio. Ao contrário dessas abordagens, aqui é proposto uma estratégia automática que combina padrões de análise técnica e análise fundamentalista para identificar notícias de interesse, sendo aplicado em cenários com ausência ou pouca informação de domínio.

3. Comparação de Medidas de Seleção de Parâmetros para Aprendizado de Sensores

Com base nos trabalhos da literatura, é possível concluir que há um espaço significativo para uso de métodos computacionais para apoiar o sensoriamento de dados em mercados financeiros. No entanto, o maior desafio é realizar esse processo de forma não supervisionada, uma vez que as variáveis envolvidas na previsibilidade de ativos financeiros podem ser alteradas constantemente. Nesse contexto, é proposta uma abordagem visando tanto a rotulação automática de notícias, quando a extração de parâmetros de um sensor (palavras-chave e seus pesos) para monitoramento de notícias de interesse.

A metodologia proposta é baseada em quatro etapas: (1) coleta de notícias, (2) pré-processamento, (3) rotulação de notícias usando eventos k-recorrentes e (4) seleção de atributos textuais, conforme detalhado a seguir.

1. **Coleta de Notícias e Eventos:** Para este projeto foi utilizada a base de notícias disponível na plataforma Websensors[1]. Essa plataforma disponibiliza uma base de conhecimento que possui cerca de 18 anos de notícias das principais fontes de informação do Brasil. As notícias podem ser coletadas com base em uma expressão de busca, local de ocorrência e período.
2. **Pré-processamento:** Nesta etapa, os textos não estruturados são representados em um formato estruturado conhecido como *bag-of-words*. O modelo clássico *bag-of-words* para representação estruturada de dados textuais é baseado no modelo espaço-vetorial, no qual cada documento é um vetor em um espaço multidimensional, e cada dimensão é um termo da coleção [Feldman and Sanger 2006, Aggarwal 2018]. Os termos são considerados independentes, formando um conjunto desordenado de palavras. Na Tabela 1 é ilustrado um esquema da representação no modelo espaço-vetorial, em que d_i corresponde ao i-ésimo documento (notícia), t_j representa o *j*-ésimo termo e a_{ij} é um valor que relaciona o *i*-ésimo documento com o *j*-ésimo termo. Desta forma, cada documento pode ser representado como um vetor $\vec{d_i} = (a_{i1}, a_{i2}, \ldots, a_{im})$.

	t_1	t_2	$\ldots$	t_m
d_1	a_{11}	a_{12}	$\ldots$	a_{1m}
d_2	a_{21}	a_{22}	$\ldots$	a_{2m}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
d_n	a_{n1}	a_{n2}	$\ldots$	a_{nm}

Tabela 1. Modelo espaço-vetorial para representação de dados textuais.

   Em relação ao valor da medida a_{ij}, é utilizado um valor que indica a importância ou distribuição do termo ao longo da coleção de notícias, no caso, o valor de TF (*Term Frequency*); que representa a frequência do termo em cada notícia. É importante ressaltar que são removidas da *bag-of-words* todos os termos que são *stopwords* (preposição, artigos, pronomes, etc) e também é utilizada apenas o radical de cada termo (técnica conhecida como *stemming*).
3. **Rotulação usando eventos k-recorrentes:** Para a etapa de rotulação, as notícias são rotuladas em positiva ou negativa conforme a série temporal do ativo financeiro. Assim, considere que uma determinada notícia foi publicada na data D, na qual o valor da cotação foi fechado em $C(D)$. A notícia será rotulada como positiva se, e somente se, $C(D+l) < C(D+l+1)$, para $1 \leq l \leq m$, em que m representa o número de observações da cotação. Assim, se $m = 4$, então significa um filtro que seleciona notícias que, após sua publicação, ocorrem 5 altas seguidas (e.g. dias) na cotação do ativo financeiro. Além disso, a mesma notícia (dado um nível de similaridade) deve ocorrer em pelo menos outros k diferentes períodos da série histórica. Ao respeitar tais critérios, então a notícia é rotulada com polaridade positiva. De forma análoga, são rotuladas as notícias com polaridade negativa. É

[1]https://websensors.net.br/

importante observar que este é um critério conservador, sendo que quanto maior o valor de m e k, menor o número de notícias rotuladas.

4. **Seleção de Parâmetros:** Dado um conjunto de notícias rotuladas nas classes positiva ou negativa por meio da abordagem de k-recorrência, a seleção de parâmetros pode ser definida como determinar um conjunto de palavras-chave e seus respectivos pesos que são mais representativos no contexto de alta ou no contexto de baixa do ativo financeiro. Uma vez definido esse conjunto de termos, uma notícia pode ser recomendada para o usuário interessado no ativo considerando a probabilidade dessa notícia afetar o valor das cotações nos próximos dias. No entanto, existem várias técnicas que podem ser aplicadas para seleção dos parâmetros. Em particular, nesse trabalho são exploradas medidas de seleção de atributos baseadas em ranking.

   As técnicas para seleção de parâmetros utilizam uma estrutura denomina Tabela de Contingência. Para cada termo $t \in T$, realiza-se uma expressão de busca $Q(t)$ sobre todas notícias, recuperando-se um subconjunto de notícias que contêm o termo t. Com o conjunto de notícias recuperados $Q(t)$ e o conjunto de notícias de uma classe G, é construído uma tabela de contingência do termo t para identificar quando documentos recuperados são relevantes (ou não) para uma determinada classe, conforme ilustrado na Figura 1):

Q(t) \ G	**Relevante**	**Não Relevante**
Relevante	*acertos*	*ruído*
Não Relevante	*perda*	*rejeitos*

Figura 1. Tabela de contingência com os possíveis resultados de recuperação por meio da expressão de busca Q(t).

Os itens da tabela de contingência são calculados da seguinte forma [Chu 2003]:

- *acertos (tp)*: número de notícias recuperadas por $Q(t)$ que pertencem a G;
- *perda (fn)*: número de notícias em G que não foram recuperadas por $Q(t)$;
- *ruído (fp)*: número de notícias recuperadas por $Q(t)$ que não pertencem a G; e
- *rejeitos (tn)*: número de notícias que não pertencem a G e que também não foram recuperadas por $Q(t)$.

Com base na tabela de contingência, são comparadas seis medidas comumente utilizadas nessas tarefas [Forman 2003]. A medida $F1$ (Eq. 1) procura obter uma média harmônica entre a $Precision$ (Eq. 2) e $Recall$ de um termo. A $Accuracy$ (Eq. 3) é uma medida para identificar o quanto um termo consegue recuperar apenas notícias de uma mesma classe. A χ^2 (Eq. 4) representa a probabilidade esperada de um termo ser de uma classe. A medida $InfoGain$ (Eq. 5) determina a organização obtida pelo termo para uma classe utilizando o conceito de entropia. A $OddsRatio$ (Eq. 6) faz uma razão entre acertos e erros para cada uma das classes.

Também é utilizado uma medida denominada $Random$, que simplesmente seleciona atributos de forma aleatória, sem considerar as informações de polaridade. Assim, é possível verificar o ganho potencial obtido por cada medida.

$$F1 = \frac{2 \times \frac{tp}{(tp+fp)} \times \frac{tp}{(tp+fn)}}{\frac{tp}{(tp+fp)} + \frac{tp}{(tp+fn)}} \tag{1}$$

$$Precision = \frac{tp}{tp + fp} \tag{2}$$

$$Accuracy = \frac{(tp + tn)}{tp + fp + tn + fn} \tag{3}$$

$$\begin{aligned} \chi^2 = t\left(tp, (tp+fp)\left(\frac{(tp+fn)}{(tp+fp+fn+tn)}\right)\right) + \\ t\left(fn, (fn+tn)\left(\frac{(tp+fn)}{(tp+fp+fn+tn)}\right)\right) + \\ t\left(fp, (tp+fp)\left(\frac{(fp+tn)}{(tp+fp+fn+tn)}\right)\right) + \\ t\left(tn, (fn+tn)\left(\frac{(fp+tn)}{(tp+fp+fn+tn)}\right)\right) \\ onde\ \ t(a,b) = \frac{(a-b)^2}{b} \end{aligned} \tag{4}$$

$$\begin{aligned} InformationGain = e(tp, fn) - \\ \left(\frac{(tp+fp)}{(tp+fp+fn+tn)} e(tp, fp) + \left(1 - \frac{(tp+fp)}{(tp+fp+fn+tn)}\right) e(fn, tn)\right) \\ onde\ \ e(x,y) = \left(\frac{(x)}{(x+y)}) * log_2\left(\frac{(x)}{(x+y)}\right)\right) - \left(\frac{(y)}{(x+y)}) * log_2\left(\frac{(y)}{(x+y)}\right)\right) \end{aligned} \tag{5}$$

$$OddsRatio = \frac{(tp * tn)}{(fp * fn)} \tag{6}$$

4. Avaliação Experimental

Para avaliar efeito de diferentes medidas de seleção de parâmetros para o contexto de sensoriamento de notícias financeiras, foi utilizado o algoritmo Centroid-Based Classifier (CBC) [Pang and Jiang 2013] para aprendizado de sensores. Esse algoritmo constrói um representante para cada classe considerando um vetor média das notícias, bem como os parâmetros selecionados. O CBC foi selecionado para esta avaliação por não utilizar nenhum método implícito de seleção de atributos durante o aprendizado do modelo (ao

contrário de métodos como Redes Neurais, SVM e Árvores de Decisão), além de permitir aprendizado incremental e em tempo linear. Cada sensor foi treinado utilizando os 30 melhores parâmetros extraídos com base nas medidas descritas anteriores (baseado na ranking de termos obtidos por cada medida). Por fim, também foi construído um modelo que utiliza todas as palavras de cada classe, ou seja, não há nenhum processo de seleção de parâmetros do sensor (modelo Baseline).

Foi utilizada a expressão de busca "petrobrás" no título da notícia para coleta da base de notícias. Foram coletadas 27.570 notícias no período de 2002 à 2014. Foi utilizada a cotação financeiro $PETR4$, coletada do índice da BM&FBOVESPA no mesmo período das notícias para a série temporal financeira. Foi realizada uma rotulação automática por k-recorrência, com $k = 3$ e $m = 4$ (para definir períodos de altas e baixas do ativo financeiro), permitindo rotular 1097 notícias, com 481 de polaridade negativa e 616 com polaridade negativa. Na Tabela 4 são apresentados os resultados experimentais considerando a taxa de acerto para notícias financeiras sobre o ativo $PETR4$. Para calcular a taxa de acerto, foi utilizada a técnica de validação cruzada com 10 pastas. Se durante o teste a polaridade de determinada classe for pelo menos duas vezes maior que a outra classe, então esta notícia é selecionada para notificar o usuário, uma vez que tal notícia pode ter um efeito (de alta ou baixa) no ativo financeiro com alta confiança.

Tabela 2. Taxa de acerto obtida pelo sensor considerada as diferentes medidas de seleção de parâmetros.

Técnica	Taxa de Acerto (%)
F1	68
Accuracy	71
Precision	69
χ^2	74.8
InformationGain	75
OddsRation	72.3
Random	59
Baseline	65.1

Na Figura 2 é apresentada uma comparação gráfica entre a taxa de acerto do sensor construído com cada medida de seleção de parâmetros. O sensor *Baseline* (sem seleção de atributos) é apresentado como referência. É possível notar que a seleção de parâmetros por meio das técnicas de $InformationGain$ e χ^2 obtiveram os melhores resultados. Este resultado fornece evidências de que técnicas baseadas na entropia e distribuição estatística dos termos são mais apropriadas para seleção de parâmetros. Esse resultado é similar aos trabalhos da literatura que exploram tais medidas utilizando listas de referência, ou seja, com alguma supervisão humana.

Outra análise interessante é que todas as medidas de seleção de parâmetros permitiram aumentar a taxa de acerto quando comparado ao método *Baseline*. Isso indica que a seleção de parâmetros é relevante para identificar termos relacionadas à alta e baixa de um ativo financeiro.

Como complemento à avaliação experimental, foi desenvolvida pelos autores deste trabalho uma ferramenta web para aprendizado de sensores com base nas medidas

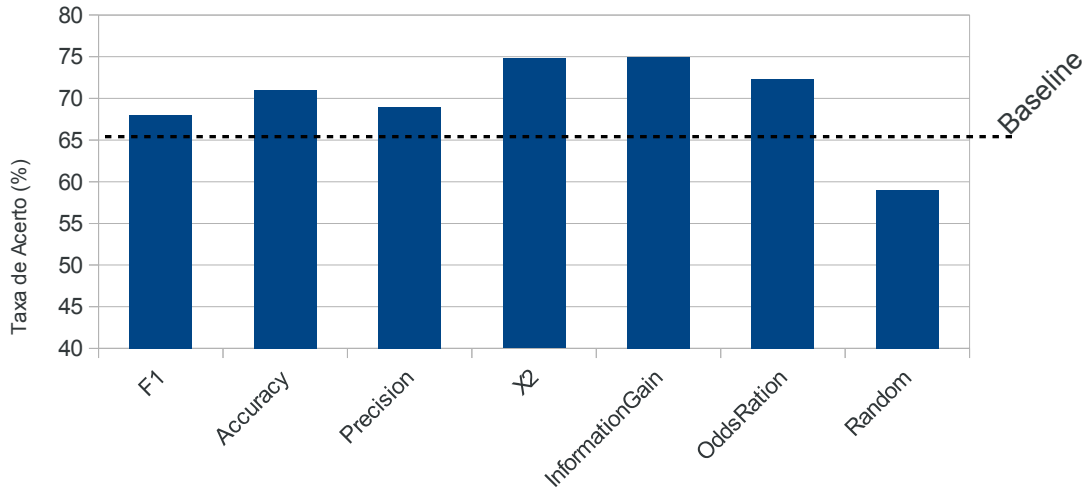

Figura 2. Comparação gráfica entre as medidas de seleção de parâmetros para aprendizado de sensores.

aqui investigadas, o que permite uma inspeção visual e subjetiva das notícias classificadas por cada medida, disponível em `https://websensors.net.br/eniac2018/`. Também estão disponíveis o código-fonte e documentação, bem como os dados utilizados nos experimentos.

5. Considerações Finais

Neste trabalho foi apresentada uma comparação experimental das medidas para seleção de parâmetros visando o aprendizado de sensores de notícias financeiras. Ao contrário das abordagens existentes, que dependem de especialistas de domínio para rotulação de notícias de interesse, nesse trabalho há o diferencial de considerar um cenário não supervisionado, por meio de uma estratégia proposta para rotulação automática de notícias, que explora os pontos de alta e baixa de uma série temporal financeira pra identificar possíveis notícias de interesse.

Foram comparadas seis diferentes medidas de seleção de parâmetros. Os resultados obtidos para notícias envolvendo a Petrobrás e o ativo financeiro PETR4 forneceram evidências de que é possível identificar notícias relevantes para pontos de altas de baixa da série temporal. Dessa forma, cientistas de dados podem explorar esse resultado para apoiar a tomada de decisão, especialmente em situações que exigem uma análise fundamentalista.

As direções para trabalhos futuros envolvem uma avaliação experimental com mais séries e ativos financeiros, bem como empregar outros algoritmos para aprendizado dos sensores. Além disso, os autores também planejam incluir aprendizado ativo durante o aprendizado dos sensores. Assim, quando for possível consultar especialistas de domínio, o método pode solicitar *feedback* apenas em situações que aparentam ser mais promissoras para melhoria do sensor, reduzindo o esforço humano.

Agradecimentos

Este trabalho contou com o apoio das seguintes agências de fomento: FAPESP (Processo 2017/08804-2), Fundect-MS (Processo 14/08996-0), CAPES, CNPq e FINEP. Os autores agradecem a NVIDIA pela doação de GPUs (*GPU Grant Program*). Os autores também agradecem ao Prof. Rafael Geraldeli Rossi (UFMS/CPTL) pelo apoio no desenvolvimento deste trabalho.

Referências

Ackland, R. (2013). *Web social science: Concepts, data and tools for social scientists in the digital age*. Sage.

Aggarwal, C. C. (2018). *Machine learning for text*. Springer.

Chan, S. W. and Chong, M. W. (2017). Sentiment analysis in financial texts. *Decision Support Systems*, 94:53–64.

Chandra, P. (2017). *Investment analysis and portfolio management*. McGraw-Hill Education.

Chu, H. (2003). *Information representation and retrieval in the digital age*. Information Today, Inc.

Einav, L. and Levin, J. (2014). Economics in the age of big data. *Science*, 346(6210):715–721.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The journal of Finance*, 25(2):383–417.

Fama, E. F. (1995). Random walks in stock market prices. *Financial analysts journal*, 51(1):75–80.

Feldman, R. and Sanger, J. (2006). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press.

Florence, R., Nogueira, B., and Marcacini, R. (2017). Constrained hierarchical clustering for news events. In *Proceedings of the 21st International Database Engineering & Applications Symposium*, pages 49–56. ACM.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of machine learning research*, 3(Mar):1289–1305.

Ghiassi, M., Skinner, J., and Zimbra, D. (2013). Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. *Expert Systems with applications*, 40(16):6266–6282.

Kaltwasser, P. R. (2010). Uncertainty about fundamentals and herding behavior in the forex market. *Physica A: Statistical Mechanics and its Applications*, 389(6):1215–1222.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Lupiani-Ruiz, E., GarcíA-Manotas, I., Valencia-GarcíA, R., GarcíA-SáNchez, F., Castellanos-Nieves, D., FernáNdez-Breis, J. T., and CamóN-Herrero, J. B. (2011). Financial news semantic search engine. *Expert systems with applications*, 38(12):15565–15572.

Majumder, D. (2013). Towards an efficient stock market: Empirical evidence from the indian market. *Journal of Policy Modeling*, 35(4):572–587.

Marcacini, R. M., Rossi, R. G., Nogueira, B. M., Martins, L. V., Cherman, E. A., and Rezende, S. O. (2017). Websensors analytics: Learning to sense the real world using web news events. In *Proceedings of the Workshops of the 23rd Brazillian Symposium on Multimedia and the Web*, pages 1–4.

Mitra, G. and Mitra, L. (2011). *The handbook of news analytics in finance*, volume 596. John Wiley & Sons.

Mostafa, M. M. (2013). More than words: Social networks' text mining for consumer brand sentiments. *Expert Systems with Applications*, 40(10):4241–4251.

Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., and Ngo, D. C. L. (2014). Text mining for market prediction: A systematic review. *Expert Systems with Applications*, 41(16):7653–7670.

Ortigosa-Hernández, J., Rodríguez, J. D., Alzate, L., Lucania, M., Inza, I., and Lozano, J. A. (2012). Approaching sentiment analysis by using semi-supervised learning of multi-dimensional classifiers. *Neurocomputing*, 92:98–115.

Pang, G. and Jiang, S. (2013). A generalized cluster centroid based classifier for text categorization. *Information Processing & Management*, 49(2):576–586.

Radinsky, K. and Horvitz, E. (2013). Mining the web to predict future events. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 255–264. ACM.

Radinsky, K., Svore, K. M., Dumais, S. T., Shokouhi, M., Teevan, J., Bocharov, A., and Horvitz, E. (2013). Behavioral dynamics on the web: Learning, modeling, and prediction. *ACM Transactions on Information Systems (TOIS)*, 31(3):16.

Schumaker, R. P., Zhang, Y., Huang, C.-N., and Chen, H. (2012). Evaluating sentiment in financial news articles. *Decision Support Systems*, 53(3):458–464.

Schumpeter, J. A. (2017). *Theory of economic development*. Routledge.

Taylor, L., Schroeder, R., and Meyer, E. (2014). Emerging practices and perspectives on big data analysis in economics: Bigger and better or more of the same? *Big Data & Society*, 1(2):2053951714536877.

Urquhart, A. and Hudson, R. (2013). Efficient or adaptive markets? evidence from major stock markets using very long run historic data. *International Review of Financial Analysis*, 28:130–142.

Yu, L.-C., Wu, J.-L., Chang, P.-C., and Chu, H.-S. (2013). Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news. *Knowledge-Based Systems*, 41:89–97.