Análise Adaptativa de Fluxo de Sentimento Baseada em Janela Deslizante Ativa

Ismael S. Silva, Glívia A. R. Barbosa, Adriano Veloso, Wagner Meira Jr. e Renato Ferreira

Universidade Federal de Minas Gerais
{ismael.silva, gliviaangelica, adrianov, meira, renato}@dcc.fmg.br

Abstract. In recent years, the task of sentiment analysis has attracted much interest from the machine learning community. Considering the benefits offered by this analysis, it is increasingly necessary to analyze feelings and opinions that are expressed continuously in sentiment streams provided by users in social media channels. Many automatic classification techniques have been used to perform this sentiment analysis, however, these techniques are not always appropriate to address the changes that occur in the sentiment stream (i.e., *sentiment drift*). This work aims to present an approach for adaptive analysis of sentiment streams that allows not only model learning, but also forgetfulness of obsolete pieces of the model during sentiment analysis. In the proposed approach, we combined a classifier based on association rules with a forgetfulness approach based on Active Sliding training Window (ASW). The use of association rules allows us to perform incremental learning while producing classification models in real-time. The processing on windows, it turn, use an active sampling technique, which keeps the most significant examples in the training set with a temporal bias. The proposed approach was empirically evaluated using Twitter. The results indicate that our approach is more efficient when compared against approaches that use all training (up to 43% gain) or a sliding window of fixed size (up to 24% gain).

Resumo. Nos últimos anos, a tarefa de análise de sentimentos tem atraído muito interesse por parte da comunidade de aprendizado de máquina. Considerando os benefícios oferecidos por esse tipo de análise, faz-se cada vez mais necessário analisar sentimentos/opiniões que são expressos continuamente em mídias sociais (i.e. análise de fluxo de sentimento). Muitas técnicas de classificação automática têm sido utilizadas para realizar tal análise, entretanto, nem sempre essas técnicas são adequadas para tratar as mudanças que ocorrem no fluxo de sentimento (i.e., o *sentiment drift*). O objetivo deste trabalho consiste em apresentar uma abordagem para análise adaptativa de fluxo de sentimentos que permite não apenas o aprendizado de modelos de classificação, mas também o esquecimento de partes obsoletas do modelo durante a análise de sentimentos. Na abordagem proposta, combinamos um classificador baseado em regras de associação com uma solução de esquecimento baseada em Janela de treino Deslizante Ativa (JDA). A utilização de regras de associação nos permite realizar um aprendizado incremental e produzir modelos de classificação em tempo real. O processamento em janelas, por sua vez, utiliza uma técnica de amostragem ativa, a qual mantém os exemplos mais significativos no conjunto de treino com um viés temporal. A abordagem proposta foi avaliada experimentalmente a partir da simulação da classificação em tempo real de mensagens enviadas através do Twitter. Os resultados indicam que nossa abordagem é mais eficiente se comparada a abordagens que utilizam todo treino (ganho de até 43%) ou uma janela deslizante de tamanho fixo (ganho de até 24%).

Categories and Subject Descriptors: H.Information Systems [**H.m. Miscellaneous**]: Databases

General Terms: Algorithms, Experimentation, Performance

Keywords: Análise de Sentimento, Fluxo de Dados, *Concept Drift*, Descoberta de Conhecimento, Twitter

1. INTRODUÇÃO

A web deixou de ser um espaço que interliga exclusivamente documentos, páginas ou recursos para tornar um ambiente de comunicação, no qual produtores e consumidores de conteúdo se misturam e interagem, estabelecendo assim novas formas de criar, organizar, compartilhar e utilizar o conhecimento [Easley and Kleinberg 2010]. Uma das fontes para geração desse conteúdo são os softwares (ou mídias) sociais (e.g., Orkut, Twitter e Facebook), onde milhões de usuários são encorajados a compartilhar e expressar suas opiniões e sentimentos. Nesse contexto, a descoberta de conhecimento a

partir do conteúdo expresso pelos usuários da web oferece oportunidades estratégicas para diferentes áreas.

Nos últimos anos a tarefa de analisar conteúdo, a partir dos sentimentos nele expressos, tem sido amplamente estudada [Bifet and Frank 2010]. Considerando os benefícios da análise de sentimento e o potencial dos canais de mídias social como fonte de conteúdo opinativo, cada vez mais se faz necessário realizar análise de sentimentos a partir dos dados gerados nos softwares sociais. Nessas aplicações é possível mensurar o sentimento expresso continuamente pela população *online* sobre diversos assuntos (e.g., epidemias [Chew and Eysenbach 2010], eleições e eventos esportivos [Silva et al. 2011]), e as métricas obtidas podem ser utilizadas por diversas áreas no processo de tomada de decisão.

A análise do conteúdo gerado a partir de mídias sociais difere das tradicionais porque segue o paradigma de fluxo de dados. Nesse modelo os dados chegam de forma contínua, em um volume quase imprevisível e sujeito a mudanças na distribuição dos dados e nos padrões relacionados às classes. Essas mudanças são conhecidas como *concept drift* e no cenário de análise de fluxo de sentimento elas serão referenciadas como *sentiment drift*. Muitas técnicas de classificação automática têm sido utilizadas para a realização da análise de sentimento, entretanto, essas técnicas normalmente não estão adequadas para realizar essa análise a partir do fluxo de dados (i.e., análise de fluxo de sentimento). Pesquisadores apontam esse tipo de análise como um dos maiores desafios enfrentados atualmente em áreas como aprendizado de máquina e mineração de dados [Bifet and Frank 2010]. Dentre esses desafios é possível destacar: (1) a necessidade dos classificadores se adaptarem às constantes mudanças no fluxo e (2) o fato dos classificadores operarem com limitações de memória, tempo de processamento e dados rotulados.

Existem muitas propostas de aprendizado incremental [Wu et al. 2006] para análise de fluxo de dados. Entretanto no contexto da análise de fluxo de sentimento, essas abordagens não são suficientes para capturar e tratar o *sentiment drift*. Nesse contexto é necessário propor soluções para atualização do modelo de classificação em tempo real, e essa atualização deve ocorrer de forma a permitir a adaptação do modelo com a inclusão e remoção (i.e., aprendizado e esquecimento) de exemplos do conjunto de treinamento. Em outras palavras, em análise de fluxo de sentimento, aprender com o decorrer do fluxo é tão relevante quanto esquecer o que já não descreve o sentimento atual.

No intuito de adequar o modelo às mudanças que podem ocorrer no fluxo de dados, muitas técnicas utilizam uma janela de treino de largura W fixa, onde os últimos W exemplos inseridos no conjunto de treino são utilizados para construção do modelo de classificação. Outros trabalhos utilizam uma função de decaimento do peso dos exemplos de treinamento. Contudo tanto a função de decaimento quanto um tamanho W ótimo são desconhecidos uma vez que, não é possível prever a taxa de mudança dos dados no cenário de análise de fluxo sentimento em tempo real.

Diante do cenário descrito, o objetivo desse trabalho consiste em apresentar uma abordagem para análise adaptativa de fluxo de sentimentos que permite não apenas o aprendizado, mas também o esquecimento durante o processo de classificação. Essa abordagem faz uso de um classificador baseado em regras de associação, o *Lazy Associative Classifier* (LAC) [Veloso and Meira Jr. 2011] e de uma proposta de esquecimento em análise de fluxo de dados denominada Janela de treino Deslizante Ativa (JDA). O LAC nos permite realizar um aprendizado incremental e produzir modelos de classificação em tempo real. A JDA é baseada na teoria do Aprendizado Ativo e seu objetivo é prover ao classificador a capacidade de manter os exemplos mais significativos no conjunto de treino com um viés temporal.

Aprendizado ativo consiste em uma técnica de amostragem de dados onde, ao invés do conjunto de treinamento ser composto de exemplos aleatórios, são selecionados exemplos que provem maior ganho de informação. Enquanto um aprendiz passivo obtém todos os dados rotulados de uma única vez, um aprendiz ativo seleciona quais exemplos ele gostaria de ver o rótulo [Zhu et al. 2007]. Esta abordagem, quando executada adequadamente, pode reduzir exponencialmente a quantidade de exemplos de treino necessária para o aprendizado [Kivinen and Mannila 1994]. Uma vez fundamentada nos princípios de

aprendizado ativo espera-se que nossa abordagem reduza o número de exemplos de treino necessários para o aprendizado, e, consequentemente, garanta que o tamanho da janela (W) seja adequado o suficiente para: (1) não sofrer os efeitos causados pelo *sentiment drift* e (2) garantir a eficácia do algoritmo.

A abordagem proposta foi avaliada utilizando o fluxo de mensagens do Twitter relacionadas a importantes eventos do ano de 2010. Os resultados revelaram que nossa solução é mais eficaz do que soluções que utilizam todo o conjunto de treino possível para classificação (ganho de até 43%) e, além disso, é melhor que as abordagens que fazem uso de janelas de tamanho fixo (ganho de até 24%).

Na próxima seção, apresentamos as estratégias existentes para esquecimento (i.e., remoção de do conjunto de treino) em fluxo de dados. Em seguida, o algoritmo proposto e a avaliação experimental realizada. Finalmente, a última seção apresenta as conclusões alcançadas.

2. TRABALHOS RELACIONADOS

Na literatura foram identificadas algumas estratégias para esquecimento no contexto de análise de fluxo de dados que fazem uso de detecção de mudanças, *ensemble* e aprendizado ativo, com intuito de contornar a necessidade de utilizar uma janela de treino de tamanho fixo. No trabalho realizado por [Bifet and Gavaldà 2007], os autores apresentam um framework para que algoritmos possam aprender com fluxo de dados de forma adaptativa. Este método é baseado no uso de detectores de mudanças. A ideia central deste algoritmo é aumentar o tamanho da janela ao valor máximo enquanto o *concept drift* não é detectado.

Em [Zhang et al. 2009] é apresentado um método baseado em *ensemble* de classificadores em fluxos de dados. [Zhu et al. 2007] propõe um framework de aprendizado ativo baseado em um conjunto de classificadores. Em [Nakayama and Yoshii 2000] é proposta uma metodologia de esquecimento ativa onde o objetivo é localizar exemplos com "má influência"para a previsão correta.

Nossa solução difere das abordagens apresentadas nessa seção por três motivos principais: (1) A teoria do aprendizado ativo suporta a possibilidade da redução exponencial do número de exemplos de treino necessários para o aprendizado. Isso faz com que o tamanho da janela seja pequeno o suficiente para não sofrer os efeitos causados pelo *sentiment drift* e grande o suficiente para garantir a eficácia do algoritmo; (2) Não é necessário pré-configurar o tamanho da janela; (3) Nossa abordagem não depende da detecção do *sentiment drift* para operar, dessa forma, ela não está sujeita ao alto custo computacional exigido pela detecção de mudanças no fluxo de sentimento e a possibilidades de erro no processo de detecção.

3. ANÁLISE ADAPTATIVA DE FLUXO DE SENTIMENTOS

Nesta seção é apresentada a abordagem proposta para análise adaptativa do fluxo de sentimento gerado pelas mensagens do Twitter. Nós modelamos o processo de aprendizado do sentimento como um problema de classificação automática, para isto, foi utilizado um classificador baseado em regras de associação. Para adaptação do modelo de classificação, os exemplos de treinamento são mantidos ou removidos de acordo com uma janela deslizante baseada na teoria do aprendizado ativo.

3.1 Aprendizado em fluxo de sentimentos

No processo de classificação automática é preciso utilizar um conjunto de treinamento previamente rotulado (referenciado nesse trabalho como $\mathcal{D}$) para que o algoritmo execute a predição. Quando essa classificação ocorre em fluxo de dados, para que os dados que chegam sejam bem representados, novos exemplos rotulados devem ser inseridos em $\mathcal{D}$ e assimilados pelo modelo de classificação com o passar do tempo. Por isso, vários algoritmos que realizam aprendizado incremental a partir desta atualização, têm sido propostos [Hulten et al. 2001].

Para realizar o processo de aprendizado em fluxo de sentimentos utilizamos o LAC [Veloso and Meira Jr. 2011]. Este classificador foi utilizado por realizar aprendizado incremental, a partir de novos exemplos que chegam rotulados no fluxo, e por demonstrar alta eficiência computacional.

Neste algoritmo, a proporção de cada sentimento contido em uma mensagem t, que chega no fluxo, é quantificada a partir de regras de associação [Agrawal et al. 1993]. Uma regra de associação é uma estrutura que pode ser representada como $\mathcal{X} \rightarrow s_i$, onde o antecedente $\mathcal{X}$ é um conjunto de termos e o consequente s_i é o sentimento previsto para as mensagens que contem todos os termos de $\mathcal{X}$. O domínio de $\mathcal{X}$ é o vocabulário de $\mathcal{D}$. A confiança da regra $\mathcal{X} \rightarrow s_i$ é denotada como $\theta(\mathcal{X} \rightarrow s_i)$, que consiste na probabilidade condicional do sentimento s_i dado os termos em $\mathcal{X}$, ou seja: $\theta(\mathcal{X} \rightarrow s_i) = \frac{\sigma(\mathcal{X} \cup s_i)}{\sigma(\mathcal{X})}$, onde $\sigma(a)$ é a probabilidade de a em $\mathcal{D}$.

No intuito de garantir que toda informação útil para classificação de t será extraída de $\mathcal{D}$, as regras são mineradas em tempo real, a partir da projeção dos dados de treinamento de acordo com os termos presentes em t [Veloso and Meira Jr. 2011]. Esta característica faz com que o aprendizado incremental seja realizado com eficiência uma vez que novas regras podem ser descobertas a cada nova mensagem que chega a partir do fluxo.

Após o modelo de classificação ser extraído de $\mathcal{D}$, as regras são utilizadas para pontuar os sentimentos presentes em t. Cada regra extraída $\mathcal{X} \rightarrow s_i$ é um voto, ponderado pela sua confiança $\theta(\mathcal{X} \rightarrow s_i)$, para o sentimento s_i. Então é calculada a média dos votos para cada sentimento s_i de acordo com sua confiança θ, gerando uma pontuação para s_i denotada como $s(t, s_i)$. Finalmente a pontuação de cada sentimento é normalizada como apresentado na Equação 1.

$$\hat{p}(s_i|t) = \frac{s(t, s_i)}{\sum_{j=0}^{k} s(t, s_j)} \tag{1}$$

3.2 Janela treino Deslizante Ativa

No contexto de análise de fluxo de sentimento, o *sentiment drift* consiste em um dos maiores problemas a ser tratado pelo classificador. Isso porque, uma vez que o fluxo representa a opinião das pessoas *online* sobre um determinado assunto, essa opinião pode mudar de maneira inesperada em função de acontecimentos imprevisíveis. Para que o conjunto de treino possa representar o fluxo de sentimento atual de forma significativa é preciso considerar os efeitos do *sentiment drift*, não apenas assimilando novos exemplos, mas também identificando e efetivamente removendo exemplos que já não descrevem o fluxo. Em outras palavras, em análise de fluxo de sentimento, aprender sobre o fluxo é tão relevante quanto esquecer exemplos que já não o descrevem.

A Janela de treino Deslizante Ativa (JDA) consiste em uma solução, fundamentada na teoria do aprendizado ativo, cujo objetivo é permitir ao classificador esquecer exemplos que não representam o fluxo atual. Dessa forma, com essa abordagem, espera-se tratar os efeitos do *sentiment drift* no fluxo de sentimento, não apenas com o aumento do treino, mas também com a remoção de exemplos afetados pelo *sentiment drift* e a partir deste processo aumentar a eficácia do classificador.

Como estratégia de esquecimento, a JDA mantém na janela de treino exemplos que provêem maior ganho de informação para que o classificador aprenda sobre o fluxo. Em outras palavras, ela mantém um conjunto de treino que provê maior ganho de informação com um viés temporal objetivando descrever melhor o fluxo de sentimento atual. Através dessa estratégia, espera-se alcançar uma janela de tamanho ótimo (i.e., uma janela cujo tamanho seja pequeno suficiente para não sofrer com os efeitos do *sentiment drift* e grande suficiente para que o algoritmo possa aprender com eficácia). Tal resultado pode ser alcançado, uma vez que a teoria do aprendizado ativo sustenta a possibilidade de diminuir exponencialmente o conjunto de treino necessário para que o classificador aprenda [Kivinen

and Mannila 1994].

No intuito de alcançarmos o objetivo anteriormente descrito utilizamos uma função de *rank* que elege potencias candidatos a serem removidos da janela atual, a cada novo exemplo rotulado (t) inserido em $\mathcal{D}$, esta função é descrita na Equação 2. Para que a operação seja realizada com eficiência computacional o treino é projetado, de acordo com os termos em t [Veloso and Meira Jr. 2011].

$$r(j) = \frac{s(t,j) + (1 - \frac{m(j)}{m(t)})}{2} \tag{2}$$

Na função de *rank* (Equação 2) t é o exemplo a ser inserido em $\mathcal{D}$, j é um exemplo $\in \mathcal{D}$ e $j \neq t$, m denota o momento que o exemplo foi inserido no treino (uma vez que os exemplos são inseridos sequencialmente), e finalmente, s indica a similaridade entre dois exemplos. Nesse caso, quanto maior a similaridade entre os exemplos e a idade do exemplo do treino, maior é o *rank*. A função de similaridade utilizada no trabalho consiste no coeficiente de *Jaccard*, esta que é apresentada na Equação 3.

$$s(i,j) = \frac{|i \cap j|}{|i \cup j|} \tag{3}$$

Calculado o *rank* para os exemplos em $\mathcal{D}$, aquele com maior pontuação é removido da janela de treino. Para definir se a pontuação do exemplo selecionado é suficiente, utilizamos um limiar que deve ser arbitrariamente pequeno. Se a pontuação for menor ou igual ao limiar o exemplo será mantido no conjunto de treino. O limiar foi definido experimentalmente como 0.1. Dessa forma o limiar garante que exemplos significativos sejam mantidos no treino na inserção de exemplos com informações novas.

Através dessa solução pretende-se manter na janela os exemplos que fornecem maior ganho de informação, uma vez que removemos o exemplo com maior similaridade ao que está sendo inserido no treino. Além disto, espera-se manter uma amostra de treino com um viés temporal, uma vez que os exemplos mais recentes terão menor probabilidade de serem removidos da janela de treino. Tal estratégia pode impactar significativamente na redução do tamanho da janela e minimizar a chance da mesma contemplar exemplos afetados pelo *sentiment drift*.

4. AVALIAÇÃO EXPERIMENTAL

Assim como no trabalho realizado por [Bifet et al. 2009], o método utilizado para avaliar nossos resultados consistiu na abordagem denominada *Interleaved Test-Then-Train*. Através dessa abordagem o exemplo é utilizado pelo classificador para teste e logo após como treino. Desta forma é possível verificar o comportamento do modelo que está sendo testado em exemplos que ele ainda desconhece. A vantagem desse método consiste no fato dele não exigir um conjunto de validação para o teste, em outras palavras, ele utiliza ao máximo os dados disponíveis [Bifet et al. 2009]. A seguir apresentamos as coleções de dados utilizadas no experimento.

4.1 Coleção de Dados

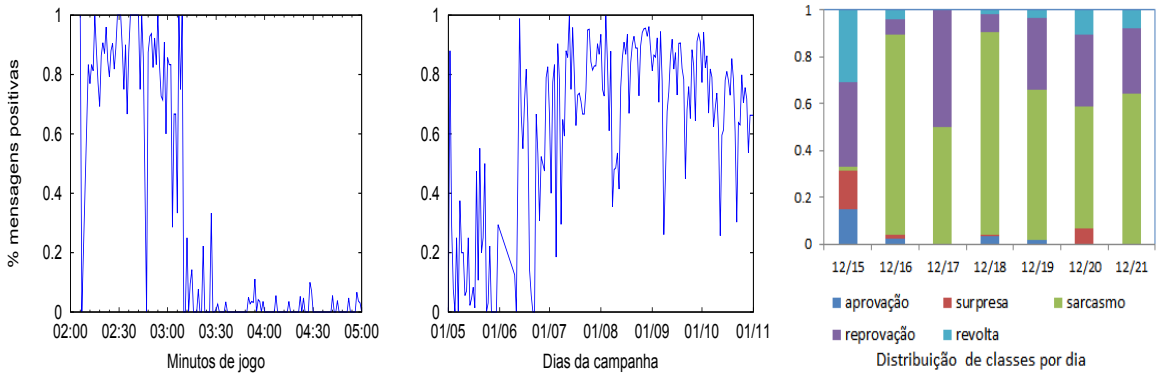
Para a avaliação foi simulada a classificação em tempo real de mensagens do Twitter relacionadas a importantes eventos de 2010. As coleções de dados utilizadas foram apresentadas em [Silva et al. 2011] e suas principais características são apresentadas na Tabela I. A partir destas coleções é possível avaliar a solução proposta em diferentes idiomas, tipos de sentimentos rastreados e *sentiment drift* (i.e., em cenários que os sentimentos mudam de diferentes maneiras no decorrer do tempo).

Table I. Coleções de Dados

Coleção	**Idioma**	**# mensagens**	**Sentimentos Rastreados**	**Velocidade do Fluxo**
Eleições Presidenciais no Brasil	Português	66.643	Positivo e negativo	0.02 mensagens/seg.
Personalidade do Ano	Inglês	5.616	Aprovação, surpresa, sarcasmo, reprovação e revolta	0.2 mensagens/seg.
Copa do Mundo de Futebol	Português	3.214	Positivo e negativo	1.12 mensagens /seg.
Copa do Mundo de Futebol	Inglês	1.432	Positivo e negativo	

Nessas coleções as mensagens sobre Eleições Presidenciais no Brasil foram classificadas como negativas ou positivas em função da candidata Dilma Rousseff. A base formada por mensagens relacionadas ao evento Personalidade do Ano eleita pela revista Times foram classificadas em relação ao sentimento das pessoas diante da escolha de Mark Zuckerberg como merecedor do prêmio ao invés de Julian Assange. Finalmente, a coleção de dados com mensagens sobre a copa do mundo aborda a opinião das pessoas em relação ao jogador Felipe Melo no último jogo da Seleção Brasileira pela Copa de 2010. Os gráficos da Figura 1 evidenciam a ocorrência do *sentiment drift* presente nessas coleções descrevendo mudanças na distribuição dos sentimentos. Além disto, nos permite ilustrar a correlação entre as mensagens do Twitter e os sentimentos no mundo real.

Na Figura 1(a), por exemplo, é possível verificar que a opinião dos usuários do Twitter em relação ao jogador Felipe Melo mudou durante a última partida do Brasil pela Copa em função da participação do jogador nesse jogo. No primeiro tempo do jogo é possível observar uma maior porcentagem de mensagens positivas em relação ao jogador uma vez que ele foi o responsável pelo passe que finalizou em um gol a favor da Seleção Brasileira. Contudo, no segundo tempo, houve uma mudança inesperada de opinião - aumento de mensagens negativas. Essa mudança reflete no sentimento das pessoas a partir do momento em que o jogador faz um gol contra e posteriormente é expulso. O sentimento negativo tende a se intensificar no final da partida, isso porque o Brasil foi eliminado da Copa nessa ocasião e a responsabilidade da derrota foi atribuída ao jogador.



(a) Copa do Mundo (Português) (b) Eleições Presidenciais no Brasil (c) Personalidade do Ano

Fig. 1. Distribuição de classes ao longo do tempo

Nas Figuras 1(b) e (c) é possível verificar, respectivamente, a positividade do sentimento da população em relação à candidata Dilma e a distribuição dos sentimentos, ao longo do tempo, em relação à eleição da personalidade do ano de 2010. Em ambos os casos, os sentimentos expressos no refletiam aos acontecimentos da vida real, sejam eles relacionados à campanha a presidência da república, ou relacionados ao período em que se discutiu sobre o prêmio de personalidade do ano.

Através desses gráficos é possível verificar a alta correlação entre o fluxo de sentimento expresso nas mensagens do Twitter e a opinião das pessoas no mundo real sobre o assunto em análise. Tal constatação serve como motivação para analisar o fluxo de sentimento gerado a partir dessa aplicação.

4.2 Resultados Obtidos e Discussão dos Resultados

Nesta seção apresentamos e discutimos os resultados obtidos com a solução proposta contrastados com outras abordagens. Em nossa avaliação utilizamos como métrica o *Mean Squared Error* (MSE) - para essa métrica, quanto menor o valor alcançado, melhor o resultado da abordagem. O MSE foi utilizado para verificar a capacidade do algoritmo de quantificar a proporção de cada sentimento presente na mensagem analisada, uma vez que esse é o objetivo da análise de sentimento e, além disto, uma mensagem pode conter mais de um sentimento.

No intuito de contrastarmos os resultados obtidos através da JDA com outras abordagens, utilizamos como linha de base duas diferentes abordagens que também foram aplicadas ao LAC: (1) uma abordagem que utiliza janela de treino de tamanho fixo (W), para isto nós executamos experimentos com diferentes tamanhos de W, e (2) uma abordagem que utiliza todo o conjunto de treino possível. No caso em que a janela de treino era de tamanho fixo, apresentamos os melhores resultados na variação do tamanho (W) da janela. A Tabela II sumariza os resultados obtidos.

Através da análise dos resultados descritos na Tabela II é possível constatar que, para a maioria das coleções de dados, a nossa abordagem (JDA) alcançou melhores resultados. No que se refere à coleção de mensagens sobre as Eleições Presidenciais, o ganho da JDA em relação à solução que utiliza janela de tamanho fixo foi de 24%, em relação a abordagem que utiliza todo treino o ganho foi de 43%. Para a base que contempla mensagens sobre a Personalidade do Ano a JDA apresentou um ganho de 11% sobre as demais abordagens. Já em relação à coleção da Copa do Mundo (mensagens sobre Felipe Melo) em Português o ganho da JDA foi de 1% se comparada à janela de tamanho fixo e 18% se comparada à abordagem que utiliza todo treino.

Somente para a coleção que contempla mensagens em inglês relacionadas ao jogador Felipe Melo, o resultado alcançado com a JDA é equivalente ao obtido com a solução que faz uso de todo o treino. Esse resultado pode ser devido a uma característica particular dessa coleção. Nessa coleção existe uma modificação brusca nos sentimento fazendo com que mais de 92% da coleção seja formada por mensagens negativas.

Table II. Resultados comparando Janela Deslizante Ativa, Janela de Tamanho Fixo, Todo Treino Disponível. Métrica MSE (melhor resultado marcado com ▲).

	Janela de Tamanho Fixo			**Todo Treino**	**Janela Deslizante Ativa**
W	***1000***	***2000***	***3000***	**-**	***526(médio)***
Eleições Presidenciais no Brasil	0.1074	0.1227	0.1337	0.1445	0.0820 ▲
W	***700***	***2100***	***2800***	**-**	***104(médio)***
Personalidade do Ano	0.3042	0.3050	0.3026	0.3006	0.2690 ▲
W	***500***	***1000***	***1500***	**-**	***73(médio)***
Copa do Mundo (Português)	0.0539	0.0594	0.0618	0.0649	0.0531 ▲
W	***300***	***600***	***900***	**-**	***69(médio)***
Copa do Mundo (Inglês)	0.2238	0.2061	0.2028	0.2015 ▲	0.2030

A partir desses resultados é possível verificar que a abordagem utilizando JDA pode ser considerada eficaz se comparada a soluções que utilizam todo o conjunto de treino e, além disso, ela é melhor que a abordagem que utiliza de janela de tamanho fixo. Nesse último caso, a JDA se destaca com a vantagem de não exigir a pré-configuração do tamanho da janela de treino.

5. CONCLUSÃO

Nesse artigo apresentamos uma solução para análise adaptativa de fluxo de sentimento utilizando um algoritmo baseado em regras de associação e uma nova abordagem para esquecimento denominada de Janela de treino Deslizante Ativa (JDA). Isso porque para que o conjunto de treino possa representar o fluxo de sentimento atual de forma significativa é preciso considerar os efeitos do *sentiment drift*, não apenas aumentando o conjunto de treino, mas também identificando e efetivamente removendo exemplos que já não descrevem o fluxo. Nossa solução é baseada na teoria do aprendizado ativo e provê ao classificador automático um conjunto de treino altamente significativo com viés temporal que melhor descreve o fluxo de sentimento atual.

A abordagem utilizando JDA foi avaliada durante a classificação de mensagens do Twitter relacionadas a importantes eventos ocorridos em 2010 e contrastada com outras abordagens que fazem uso de todo conjunto de treino ou de uma janela de treino fixa para classificação. A JDA se mostrou mais eficiente se comparada às outras abordagens citadas, com ganhos de até 43%.

Esses resultados refletem aspectos da teoria de aprendizado ativo, que sustenta a hipótese da JDA de reduzir número de exemplos de treino necessários para o aprendizado, de tal forma que o tamanho da janela seja pequeno o suficiente para não sofrer os efeitos do *sentiment drift* e grande o suficiente para garantir a eficácia do algoritmo. Além disso, uma vez que a JDA não depende da detecção do *sentiment drift* para operar, ela não está sujeita ao alto custo computacional dessa detecção. Como trabalho futuro pretendemos evoluir essa abordagem para integrá-la efetivamente a solução de auto aumento de treino, pelo classificador, apresentada em [Silva et al. 2011].

6. AGRADECIMENTOS

O presente trabalho foi realizado com o apoio do UOL (www.uol.com.br) através do Programa UOL Bolsa Pesquisa processo número 20110215215201, CNPq, Capes, Fapemig e Inweb.

REFERENCES

Agrawal, R., Imielinski, T., and Swami, A. N. Mining association rules between sets of items in large databases. In *Proc. of the ACM SIGMOD Conf.* pp. 207–216, 1993.

Bifet, A. and Frank, E. Sentiment knowledge discovery in twitter streaming data. In *Discovery Science*. pp. 1–15, 2010.

Bifet, A. and Gavaldà, R. Learning from time-changing data with adaptive windowing. In *SIAM ICDM*, 2007.

Bifet, A., Holmes, G., Pfahringer, B., Kirkby, R., and Gavaldà, R. New ensemble methods for evolving data streams. In *Proc SIGKDD*. KDD '09. pp. 139–148, 2009.

Chew, C. and Eysenbach, G. Pandemics in the age of twitter: Content analysis of tweets during the 2009 h1n1 outbreak. *PLoS ONE* 5 (11): 13, 2010.

Easley, D. and Kleinberg, J. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, New York, NY, USA, 2010.

Hulten, G., Spencer, L., and Domingos, P. Mining time-changing data streams. In *Proc SIGKDD*. pp. 97–106, 2001.

Kivinen, J. and Mannila, H. The power of sampling in knowledge discovery. In *Proc SIGACT-SIGMOD-SIGART*. ACM, pp. 77–85, 1994.

Nakayama, H. and Yoshii, K. Active forgetting in machine learning and its application to financial problems. *Neural Networks, IEEE - INNS - ENNS International Joint Conference on* vol. 5, pp. 5123, 2000.

Silva, I. S., Gomide, J., Veloso, A., Meira, Jr., W., and Ferreira, R. Effective sentiment stream analysis with self-augmenting training and demand-driven projection. In *Proc SIGIR*. pp. 475–484, 2011.

Veloso, A. and Meira Jr., W. *Demand-Driven Associative Classification*. SpringerBriefs Computer Science, 2011.

Wu, S., Yang, C., and Zhou, J. Clustering-training for data stream mining. In *Proc ICDM*. pp. 653–656, 2006.

Zhang, Y.-s., Zhang, J.-p., Yang, J., and Yin, Z.-W. Svms' cooperative learning strategy based on mas to data streams mining. In *Proc ICICSE*. IEEE Computer Society, Washington, DC, USA, pp. 156–159, 2009.

Zhu, X., Zhang, P., Lin, X., and Shi, Y. Active learning from data streams. In *Proc ICDM*. pp. 757–762, 2007.