Detecção de Opinião em Mensagens Curtas usando Comitê de Classificadores e Indexação Semântica

Johannes V. Lochter, Rafael F. Zanetti, Tiago A. Almeida
Departamento de Computação (DComp)
Universidade Federal de São Carlos (UFSCar)
18052-780, Sorocaba, São Paulo, Brasil
jlochter@acm.org, ferrazzanett@wisc.edu, talmeida@ufscar.br

***Resumo*—A alta popularidade das redes sociais vem atraindo cada vez mais a atenção das empresas. O volume crescente de usuários e mensagens postadas tornam as redes sociais em ambientes frutíferos para detecção de necessidades, tendências, opiniões e outras informações relevantes, que podem alimentar bancos de dados de departamentos de marketing e vendas. Contudo, a maioria das redes sociais impõe limite no tamanho das mensagens, fazendo com que os usuários compactem seus textos por meio de abreviações, gírias e símbolos. Trabalhos recentes indicam avanços na redução do impacto que as mensagens curtas e ofuscadas têm em tarefas de categorização de textos, ao serem pré-processadas com dicionários semânticos e modelos ontológicos. Nesse sentido, é proposto neste trabalho um sistema de comitê de máquinas de classificação para encontrar automaticamente a melhor combinação entre métodos de classificação e técnicas recentes de processamento de linguagem natural, com o intuito de detectar a polaridade de mensagens curtas e ruidosas. Os experimentos realizados foram diligentemente projetados e analisados estatisticamente, indicando que o sistema proposto é superior aos demais métodos preditivos avaliados.**

I. Introdução

A inclusão digital vem permitindo que cada vez mais pessoas tenham acesso à Internet. Isso, aliado ao aumento no uso de dispositivos móveis, vêm promovendo o sucesso das redes sociais. Tais aplicações permitem, dentre muitas ações, compartilhar e ler informações. Dentre o altíssimo volume de informações compartilhadas diariamente, muitos usuários costumam disseminar opiniões, avaliar produtos e consultar avaliações feitas por outros usuários. Segundo pesquisa recentemente divulgada pela ComScore[1], análises e opiniões sobre produtos têm impacto significativo na decisão de compra. Consequentemente, as empresas vêm descobrindo a importância de identificar usuários formadores de opiniões e a necessidade de analisar grande quantidade de dados rapidamente para descobrir tendências e opiniões dos usuários.

Apesar da aplicação de métodos de classificação na detecção de opinião em mensagens de texto já ter sido explorada na literatura [1], [2], na maioria dos casos, ainda é desafiador identificar a polaridade de mensagens extraídas das redes sociais. Isso se deve ao fato das amostras normalmente serem curtas e ruidosas, repletas de gírias, abreviações e símbolos, que dificultam até mesmo o processo de tokenização.

Ruídos em mensagens compartilhadas nas redes sociais podem aparecer de diversas maneiras. A seguinte frase ilustra um exemplo: *"dz ne1 knw h2 ripair dis terrible LPT? :("*. Nela, há palavras grafadas incorretamente *"dz, ne1, knw, h2, dis"*, abreviação *"LPT"* e símbolo *":("*. Para que tal frase torne-se gramaticalmente correta na língua inglesa, é necessário empregar um processo de normalização e tradução para que cada gíria, símbolo e abreviação seja associado aos termos corretos da língua. Após essa etapa, a frase seria transcrita para *"Does anyone know how to repair this terrible printer? :("*, e o símbolo ao final indicaria a insatisfação do autor com relação ao produto.

Além das mensagens ruidosas, há outros problemas bem conhecidos na literatura, relacionados à extração de polaridade de mensagens de texto, como frases sarcásticas ou irônicas, discernimento de palavras ambíguas em um determinado contexto (polissemia) ou ocorrência de diferentes palavras com o mesmo significado (sinonímia). Há um consenso de que tratar estes casos pode conduzir a melhores resultados [3], [4]. Os exemplos a seguir refletem algumas dessas dificuldades.

- Em *"What a great car! It stopped working in two days."*, o adjetivo positivo *"great"* é usado para expressar uma opinião negativa ironicamente.
- Em *"This vacuum cleaner really sucks."*, a palavra *"sucks"* é empregada de forma positiva, embora ela seja normalmente empregada para expressar uma opinião negativa.
- Em *"I am going to read a book."* e *"I am unable to book that hotel."*, a palavra *"book"* possui diferentes significados ao longo da frase, caracterizando a polissemia.
- Em *"I bought a gift"* e *"I purchased a gift"*, as palavras *"bought"* e *"purchased"* têm o mesmo significado, caracterizando a sinonímia.

Problemas como sinonímia e polissemia podem ser minimizados através de indexação semântica, empregada para realizar a desambiguação de palavras. Tal técnica traça a relação de significado dos termos, permitindo encontrar termos com significados próximos ou polissêmicos no contexto da mensagem. Em geral, o emprego eficiente dos dicionários é condicionado à qualidade da extração dos termos das amostras [5], [6].

Após tratar os problemas de sinonímia e polissemia, muitas vezes os termos resultantes ainda não são suficientes para discernir a polaridade da mensagem. Trabalhos recentes recomendam empregar modelos ontológicos para analisar cada

[1] *Online Consumer-Generated Reviews Have Significant Impact on Offline Purchase Behavior*. Disponível em http://goo.gl/PRIHmS, acessado em 30/03/2015.

termo e associar novos conceitos, com mesmo significado, para enriquecer o conteúdo original da mensagem [7], [8].

Os conceitos obtidos através de modelos ontológicos e indexação semântica (chamada processo de expansão) são mais úteis para os métodos de classificação se forem associados a uma polaridade individualmente. Neste sentido, a aplicação de dicionários léxicos pode melhorar o desempenho dos classificadores [4], [8].

As mensagens originais podem ser processadas por diversas técnicas distintas de processamento de linguagem natural (*natural language processing* – NLP), assim como a saída resultante de cada uma dessas técnicas pode ser apresentada a diferentes métodos de classificação. Consequentemente, para que as melhores combinações sejam feitas de forma automática, é natural empregar um comitê de máquinas de classificação. Tal recurso permite minimizar as deficiências de cada método, gerando hipóteses mais genéricas e melhorando o poder preditivo, ao combinar as saídas de diferentes técnicas de NLP com métodos consolidados de classificação [9], [10].

Nesse cenário, este artigo apresenta um sistema de comitê de classificadores e técnicas de NLP para detectar a polaridade de mensagens curtas e ofuscadas extraídas de redes sociais. O modelo proposto combina técnicas de normalização de texto com técnicas consideradas estado da arte em processamento de linguagem natural para melhorar a qualidade dos atributos que serão posteriormente empregados por métodos consolidados de categorização de texto. Os resultados apresentados demonstram que o sistema proposto é superior aos métodos disponíveis e conhecidos na literatura.

O restante deste artigo está estruturado da seguinte maneira: a Seção 2 apresenta uma breve revisão dos principais trabalhos relacionados. Os processos de normalização de texto e indexação semântica são apresentados na Seção 3. A Seção 4 descreve o sistema de comitê de máquinas de classificação. A metodologia experimental é detalhada na Seção 5. A Seção 6 apresenta os principais resultados obtidos e análises realizadas. Finalmente, as principais conclusões e direcionamentos para trabalhos futuros são oferecidos na Seção 7.

II. Trabalhos relacionados

Detecção de opinião é a tarefa de analisar grande quantidade de informação a partir de centenas (ou milhares) de mensagens para descobrir se existe um consenso a respeito de algo em discussão. Esse conhecimento, aliado a decisões rápidas, pode auxiliar departamentos de marketing das empresas, além de fortalecer tomadas de decisões [2], [4]. Essa tarefa está longe de ser plenamente solucionada por vários motivos, como a dificuldade de lidar com sarcasmo, ironia e sentenças com múltiplas polaridades. Além disso, outro problema bem conhecido está relacionado à baixa quantidade e qualidade de atributos extraídos das mensagens curtas e ruidosas, comumente empregadas em redes sociais e em comunicações usando dispositivos móveis. De maneira geral, ela afeta a representação computacional e prejudica o desempenho dos métodos tradicionais de classificação [11], [12], [13], [14].

Em categorização de textos curtos, um desafio que continua em aberto é a falta de informação referente ao conteúdo. O tamanho restritivo imposto pelo canal (Twitter, por exemplo), além de limitar consideravelmente o tamanho das mensagens, induz os usuários a fazerem uso de uma gramática arbitrária, repleta de gírias, símbolos e abreviações que podem acarretar problemas como polissemia e sinonímia. Polissemia é a capacidade de um termo ter vários significados, mas ser representado por apenas um atributo. Já a sinonímia é a capacidade de vários termos terem o mesmo significado e, portanto, serem representados por mais um de atributo. Nesse cenário, o uso de técnicas de indexação semântica e normalização léxica podem minimizar esses problemas e melhorar a qualidade da representação das mensagens [5], [8], [15].

Normalização léxica ou normalização textual é a tarefa de substituir variantes léxicas ou expressões grafadas incorretamente pelas suas formas canônicas, permitindo a execução de outras tarefas de processamento de texto. Muito similar a correção de grafia, diversos outros trabalhos presentes na literatura empregam esse procedimento para pré-processar as amostras antes da tokenização [16], [17].

Indexação semântica é a técnica para encontrar sinônimos (conceitos) para um determinado termo dentro do contexto no qual ele foi usado [18]. Por exemplo, a rede semântica WordNet representa sinônimos da seguinte forma: {*car, auto, automobile, machine, motorcar*} (*a motor vehicle with four wheels*) ou {*car, railcar, railway car, railroad car*} (*a wheeled vehicle adapted to the rails of railroad*) para a palavra "*car*".

As amostras resultantes da indexação semântica podem tornar difícil a identificação dos conceitos associados a cada palavra dentro do contexto da mensagem. Este problema pode ser sanado através da *desambiguação de conceitos* (*Word Sense Disambiguation* – WSD), um problema popular em linguística e usualmente tratado com técnicas de NLP [19]. Nesse quesito, neste trabalho foram empregados a rede semântica BabelNet [12] e um algoritmo de desambiguação não-supervisionado, seguindo o método de expansão semântica descrito em Gómez Hidalgo *et al*. [18].

Após os passos de normalização léxica e indexação semântica, as amostras originais curtas e ruidosas são tratadas e expandidas pela adição de novos conceitos relacionados ao contexto da mensagem. Portanto, além de ter seus termos normalizados, a amostra é enriquecida com informações que poderão auxiliar os métodos de classificação e aumentar o poder de predição [7], [8].

A abundância de técnicas de processamento de texto e métodos de classificação existentes demanda a escolha de alguma estratégia para combiná-los de tal forma a obter hipóteses robustas e genéricas.

Comitê de máquinas de classificação é uma técnica desenvolvida para obter hipóteses mais genéricas através da combinação de diferentes classificadores. Em um comitê, cada classificador é um membro com poder de voto e a decisão final depende dos votos de todos os membros, podendo cada um ter voto com peso diferente. Essa técnica é aplicada normalmente para minimizar as deficiências de cada método de classificação, evitando o super-ajustamento e a maldição da dimensionalidade [20]. Embora hajam diferentes tipos de comitês de classificadores, comumente eles possuem maior poder preditivo que métodos tradicionais isolados [21], [10].

Os comitês de máquinas de classificação com votação pon-

derada são encontrados com bastante frequência na literatura. A eficácia desses métodos geralmente está associada ao peso que cada classificador tem em seu voto, de forma que, em um dado cenário, um classificador com menor poder preditivo possui voto com peso inferior ao voto de um classificador com maior capacidade de predição [9], [22].

Como as mensagens de texto podem ser processadas por diversas técnicas de NLP que, por sua vez, podem ser combinadas com vários métodos de classificação recomendados para a detecção de polaridade, pressupõem-se que um sofisticado sistema de comitê de máquinas de classificação seja capaz de gerar hipóteses mais genéricas e, portanto, com maior capacidade preditiva.

III. Técnicas de processamento de texto

Em categorização de texto, o emprego direto da *bag of words* para representar mensagens curtas e repletas de gírias, símbolos e abreviações, geralmente conduz a resultados insatisfatórios [23], [24], [15]. Uma etapa de normalização léxica é frequentemente utilizada para transformar mensagens ofuscadas em mensagens gramaticalmente corretas. Em seguida, como as mensagens podem ser muito curtas, a pequena quantidade de atributos extraídos pode ser insuficiente para as técnicas de classificação, principalmente porque problemas de polissemia e sinonímia são bastante comuns nesse contexto. Desta forma, dicionários semânticos e técnicas consideradas estado da arte para detecção de contexto podem auxiliar na expansão da amostra original para uma nova amostra, com mais atributos e menos ambiguidade. Com esse intuito, foi desenvolvido um processo de normalização e expansão de texto[2], no qual a mensagem de entrada é processada em etapas, sendo que em cada uma delas é gerada uma saída diferente, que posteriormente são combinadas para formar uma nova amostra. Cada etapa é brevemente descrita a seguir.

A. Normalização léxica

Nesta etapa, foram empregados dois dicionários para normalizar e traduzir termos conhecidos como *Lingo*, nome dado ao conjunto de gírias e abreviações comumente utilizados na Internet, por palavras da língua inglesa. O primeiro é um dicionário padrão de inglês[3], usado para consultar se uma determinada palavra existe no idioma inglês ou não. Quando encontrada, a palavra é substituída por seu radical (por exemplo: "*having*" por "*have*"). O segundo dicionário é o de Lingo[4], que é utilizado para traduzir um termo Lingo para sua forma canônica em inglês. Caso não haja uma tradução para o termo de entrada, o termo original é mantido.

B. Geração de conceitos por indexação semântica

Os conceitos são provenientes do repositório BabelNet [12], um moderno e grande dicionário semântico da língua inglesa, composto por conceitos extraídos da WordNet e Wikipedia [12]. Como esta etapa demanda entradas em inglês, a *normalização léxica* é aplicada para garantir que os termos estão, de fato, em inglês. Contudo, o método ignora palavras que estejam em uma lista de *stopwords* (artigos e pronomes) para prevenir a execução exaustiva de buscas por substitutos de baixa qualidade preditiva, além de economizar tempo de processamento. Se a palavra estiver na lista de *stopwords*, ela é mantida. Caso contrário, a palavra é usada como campo de busca no dicionário semântico que, se encontrado, retorna uma lista de conceitos relacionados.

C. Desambiguação de conceitos

Como a quantidade de conceitos para cada termo pode ser excessiva, uma técnica de desambiguação foi implementada para selecionar o conceito mais relevante para cada palavra, de acordo com o seu contexto na mensagem original. Para esse propósito, foi implementado o método não-supervisionado de desambiguação proposto por Navigli e Ponzetto [5].

Após a execução de cada etapa de processamento da mensagem e com base em uma regra de combinação pré-definida pelo usuário, as diferentes saídas são mescladas criando uma amostra final normalizada, com mais atributos e, portanto, mais adequada para ser utilizada pelos método de classificação.

Para oferecer uma breve ideia do processo de expansão, é apresentado um exemplo na Tabela I. Considerando a amostra original "*plz lemme noe when u get der*", a saída de cada etapa é demonstrada. A *normalização léxica* troca as gírias e abreviações pelas palavras correspondentes em inglês. A *geração de conceitos* obtém um conjunto de conceitos para cada palavra da amostra original e a *desambiguação de conceitos* seleciona o conceito mais relevante para cada palavra, de acordo com o seu contexto na mensagem. Finalmente, a partir de uma regra de combinação escolhida pelo usuário [normalização léxica + desambiguação de conceitos], por exemplo, a amostra final normalizada e expandida será "*please let lease me know cognition when you get there*", a qual espera-se ser mais informativa e adequada para os métodos de aprendizado de máquina.

Tabela I. Exemplo de normalização léxica e expansão semântica de uma amostra de texto curta e ruidosa com a regra de combinação [normalização + desambiguação]

Original	*plz lemme noe when u get der*
Normalização	*please let me know when you get there*
Indexação Semântica	*please army_of_the_righteous lashkar-e-taiba lashkar-e-tayyiba lashkar-e-toiba let net_ball me knoe knowledge noesis when you get there*
Desambiguação	*please lease me cognition when you get there*
Amostra final	*please let lease me know cognition when you get there*

IV. Comitê de máquinas de classificação

O sistema de comitê de máquinas de classificação proposto é dividido em duas etapas: seleção de modelo e classificação.

Na etapa de *seleção de modelo*, inicialmente, os parâmetros de cada método de classificação que compõem o sistema são ajustados através de uma *grid search*. Como este processo pode consumir bastante tempo, apenas um sub-conjunto estratificado e aleatório de amostras é utilizado. Em seguida, as amostras de entrada são normalizadas e expandidas com as técnicas de normalização e expansão de texto ($E_1, \ldots, E_k$). Todas as possíveis regras de combinação são utilizadas, sendo

[2] `TextExpansion`: disponível em http://lasid.sor.ufscar.br/expansion/.

[3] *Freeling English dictionary*: disponível em: http://devel.cpl.upc.edu/freeling/.

[4] *NoSlang Lingo dictionary*: disponível em: http://www.noslang.com/dictionary/full/.

que cada uma delas produz um conjunto de saídas diferente, que são posteriormente usadas para treinar e avaliar os métodos de classificação ($C_1, \ldots, C_n$). Em seguida, o desempenho alcançado por cada possível combinação expansor-classificador ($E_p \rightarrow C_j$) é avaliado para definir qual regra de combinação é a mais adequada para cada método de classificação do sistema ($E^*_{c_j} = \max(E_p, C_j)\ \forall\ p \in \{1, \ldots, k\}$, $j \in \{1, \ldots, n\}$). Finalmente, um peso w_j (grau de confiança) é calculado para cada combinação j, baseado na sua acurácia individual comparada com aquela que obteve melhor acurácia no sistema (Eq. 1). A Figura 1 ilustra essa etapa.

$$w_j = \frac{1}{\left|\log_2\left(\frac{Acc_j}{max(Acc_j)+T}\right)\right|}, \quad 1 \leq j \leq n. \tag{1}$$

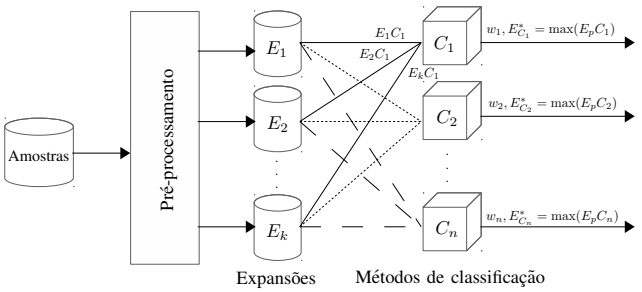

Figura 1. Na etapa de seleção de modelo, o conjunto de dados original é processado pelas técnicas de normalização e expansão de texto ($E_1, \ldots, E_k$). Em seguida, cada base resultante da expansão é usada no treinamento de cada método de classificação ($C_1, \ldots, C_n$). Para cada método de classificação, é selecionada a melhor combinação expansor-classificador, $E^*_{c_j} = \max(E_p, C_j)\ \forall\ p \in \{1, \ldots, k\}$, $j \in \{1, \ldots, n\}$ e calculado um peso w_j, correspondente ao grau de confiança de tal combinação.

Na Equação 1, uma constante $0 < T \leq 1$ foi adicionada para controlar o balanceamento entre os maiores e menores pesos (w_j). Quanto menor o valor de T, maior é a diferença dos pesos entre os classificadores com melhor e pior desempenho, com base nas acurácias obtidas. Assim, conforme T tende a 0, maior é a diferença do peso do voto w_j do classificador com maior acurácia ($Acc_j = max(Acc_j)$) e todos os demais classificadores com $Acc_j < max(Acc_j)$. Por outro lado, conforme T tende a 1, menor será a diferença do peso de voto entre todos os classificadores. A Figura 2 ilustra como a escolha do valor de T pode influenciar na diferença entre os pesos dos votos dos classificadores com diferentes desempenhos obtidos na etapa de seleção de modelo.

Finalizada a etapa de seleção, o melhor modelo é escolhido e os métodos de classificação são treinados. Em seguida, na *etapa de classificação*, as amostras são pré-processadas pelas técnicas de normalização e indexação semântica selecionadas na etapa anterior. As saídas de cada uma dessas técnicas são mescladas de acordo com a regra de combinação mais adequada para cada método de classificação, gerando a amostra de entrada que será processada por cada classificador que, por sua vez, emite uma predição com um certo grau de confiança (w). O rótulo final é então computado pelo voto majoritário ponderado, conforme ilustrado na Figura 3.

Em resumo, neste trabalho foi projetado um sistema de máquinas de classificação adequado para manipular mensagens

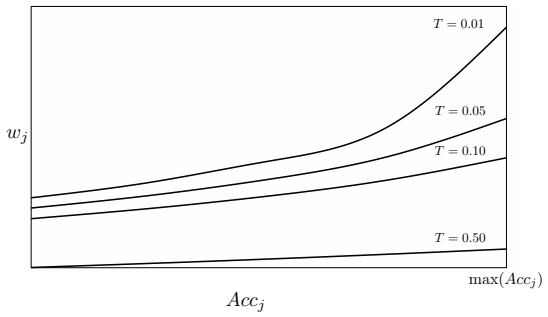

Figura 2. A constante T influencia na diferença entre os pesos dos votos dos classificadores com desempenhos diferentes na etapa de seleção de modelo. Quanto menor o valor de T, maior será a diferença de peso entre os classificadores com maior e menor acurácia.

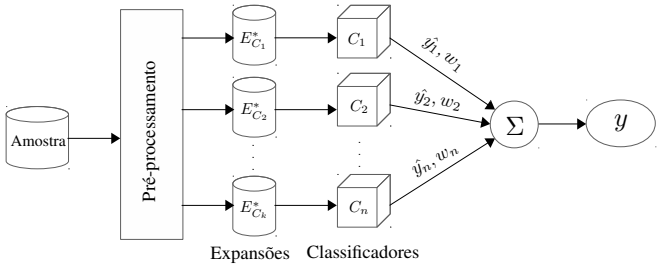

Figura 3. Após a etapa de seleção de modelo, o sistema associa quais técnicas de normalização e expansão de texto (e regra de combinação) (E^*_p) são mais adequadas para cada método de classificação (C_j), e o treinamento é então realizado. Em seguida, na etapa de classificação, dada uma amostra de entrada, ela é pré-processada e classificada por cada modelo que envia ao concentrador (Σ) sua predição ($\hat{y_j}$) com um grau de confiança (w_j). O resultado final é então computado com base no voto majoritário ponderado.

de texto curtas e ruidosas, como as comumente utilizadas nas redes sociais e dispositivos móveis. O método proposto seleciona automaticamente a melhor combinação entre técnicas de normalização e expansão de texto e métodos de classificação, considerados estado da arte tanto em processamento de linguagem natural quanto em aprendizado de máquina. Com isso, a abordagem proposta é capaz de gerar hipóteses mais robustas e genéricas que podem conduzir a desempenhos superiores aos métodos isolados e disponíveis na literatura.

V. METODOLOGIA

Para dar credibilidade aos resultados obtidos e para tornar os experimentos completamente reprodutíveis, a metodologia utilizada neste trabalho é detalhada a seguir.

A. Bases de dados e representação

Foram empregadas sete bases de dados reais, públicas e não-codificadas. Na Tabela II, é apresentada a quantidade de amostras em cada classe, bem como o tema relacionado ao objeto de interesse de cada base. Cada amostra foi rotulada como positiva ou negativa, de acordo com a opinião que a mensagem expressa com relação ao objeto de interesse.

As quatro primeiras bases de dados foram pré-processadas de acordo com os procedimentos descritos por Saif *et al*. [29]. As mensagens das últimas três bases de dados listadas na

Tabela II. BASES DE DADOS USADAS NA AVALIAÇÃO DO SISTEMA PROPOSTO.

Base de dados	# Positivas	# Negativas	Tema
STS-Test [25]	181	177	Geral
HCR [26]	537	886	Medicina
OMD [27]	709	1195	Política
SS-Tweet [28]	1252	1037	Geral
iPhone6	371	161	*Smartphone*
Archeage	724	994	Jogo
Hobbit3	354	168	Filme

Tabela II foram coletadas do Twitter e rotuladas pelos autores deste trabalho, no segundo semestre de 2014, utilizando uma ferramenta de rotulação colaborativa, desenvolvida para essa finalidade, chamada *Labeling*[5]. Todas as bases criadas estão publicamente disponíveis em http://dcomp.sor.ufscar.br/talmeida/sentcollections/.

As amostras foram representadas computacionalmente por unigramas gerados a partir do conteúdo das mensagens. Os símbolos comumente utilizados no Twitter foram preservados, tais como números e sinalizadores de *hashtags* e *retweets*. Além disso, dois novos atributos foram criados para indicar o número de termos positivos e negativos presentes na mensagem. Para isso, cada termo da amostra foi procurado em um dicionário léxico[6], e se tal termo estiver associado a um valor positivo, o atributo referente ao contador de termos positivos é incrementado em um. Por outro lado, caso o termo esteja associado a um valor negativo, então o contador de termos negativos é incrementado em um. Segundo Mostafa [4] e Nastase e Strube [8], abordagens semelhantes vêm apresentando bons resultados em tarefas de detecção de opinião.

B. Métodos de classificação

O sistema proposto é formado por métodos de classificação (Tabela III) apontados por Wu *et al*. [30] como os melhores disponíveis para tarefas de classificação e mineração de dados. Tais métodos utilizam representação e estratégia de seleção de hipótese diferentes, tais como probabilidade, otimização, distância e árvores. Com isso, é esperado que o comitê de máquinas de classificação seja flexível e capaz de gerar hipóteses genéricas e robustas, já que o próprio comitê encontra de forma automática a melhor estratégia e modelo para cada base de dados.

Tabela III. MÉTODOS DE CLASSIFICAÇÃO EMPREGADOS NO SISTEMA PROPOSTO.

Métodos de classificação
Naïve Bayes Bernoulli (NB-B) [31]
Naïve Bayes Multinomial (NB-M) [31]
Naïve Bayes Gaussiano (NB-G) [31]
Máquinas de vetores de suporte (SVM) [32], [33]
Árvores de decisão (C4.5) [34]
N-vizinhos próximos (k-NN) [35]
Boosted C4.5 (B.C4.5) [36]
Regressão logística (RL) [33]

C. Avaliação do sistema

Para cada base de dados, inicialmente foram selecionadas aleatoriamente 20% das amostras para a etapa de seleção de modelo, e as demais 80% foram empregadas no treinamento e teste do modelo selecionado.

Na seleção do modelo, a base de dados de entrada foi processada e expandida pelas possíveis técnicas de processamento de texto, normalização e indexação semântica, resultando em um conjunto de bases expandidas (uma para cada possível regra de combinação). Em seguida, para cada base expandida, foram ajustados os valores dos parâmetros dos métodos de classificação através de *grid search*. Nessa etapa, a seleção de valores foi feita com base na F-medida, usando validação cruzada com 5 partições. Posteriormente, após os parâmetros terem sido ajustados, o sistema determinou qual base expandida obteve o melhor desempenho para cada método de classificação. Em outras palavras, nesta etapa foi selecionada a regra de combinação mais adequada para cada método de classificação do sistema. Finalmente, a constante T, usada para ajustar a diferença entre os pesos dos votos dos classificadores (Eq. 1), foi empiricamente definida com valor igual a 0,05.

A maior parte do conjunto original dos dados (80% das amostras) foi usada para treinar o modelo selecionado e testá-lo. Essas amostras foram aleatoriamente divididas em duas partes: conjunto de treino (75%) e conjunto de teste (25%).

Para comparar o desempenho alcançado pelo comitê proposto, os mesmos passos e condições foram usados para cada método de classificação individualmente. Assim, para prover uma comparação justa, cada método de classificação também foi combinado com a regra de combinação que obteve melhor desempenho, além de ter seus parâmetros também ajustados por *grid search*, através do mesmo protocolo experimental empregado com o comitê de máquinas.

Para não correr o risco de algum método ter seu desempenho tendencioso de acordo com a configuração dos conjuntos de treino e teste, esses procedimentos foram repetidos dez vezes para cada método, com seleção aleatória e estratificada das amostras.

VI. RESULTADOS

A Tabela IV apresenta os resultados obtidos pelos métodos de classificação para cada base de dados. O desempenho de cada técnica foi avaliado pela média e desvio padrão da F-medida obtida para os dez testes executados. Para cada base, os melhores resultados estão destacados em negrito.

Sob as mesmas condições, é evidente que o sistema de comitê de máquinas proposto obteve desempenho superior a qualquer um dos métodos de classificação individuais avaliados. Porém, para garantir que os resultados não foram obtidos por acaso, foi realizado o teste estatístico não-paramétrico de Friedman [37], seguindo a metodologia descrita em Japkowicz e Shah [38].

O teste de Friedman avalia se a hipótese nula, que neste caso afirma não haver diferenças entre os resultados obtidos, pode ser rejeitada com base no *ranking* computado através do desempenho obtido por cada método de classificação para cada base de dados.

Seja k a quantidade de métodos de classificação, n o número de base de dados e R a soma dos *rankings* de cada método, a medida χ^2_F foi calculada de acordo com a Equação 2.

[5]*Labeling*. Disponível em http://lasid.sor.ufscar.br/ml-tools/.

[6]O dicionário léxico utilizado está disponível em http://goo.gl/czIfkd.

Tabela IV. MÉDIA E DESVIO PADRÃO DA F-MEDIDA OBTIDA POR CADA MÉTODO DE CLASSIFICAÇÃO AVALIADO.

Métodos	Archeage	HCR	Hobbit	IPhone6	OMD	SS-Tweet	STS-Test
B. C4.5	0,79±0,04	0,68±0,04	0,88±0,04	0,68±0,06	0,74±0,03	0,56±0,02	0,81±0,06
C4.5	0,75±0,02	0,64±0,03	0,86±0,02	0,68±0,05	0,68±0,04	0,54±0,03	0,75±0,06
Comitê	**0,87±0,02**	**0,73±0,03**	**0,92±0,02**	**0,74±0,04**	**0,81±0,02**	**0,61±0,02**	**0,86±0,03**
KNN	0,72±0,03	0,63±0,03	0,84±0,04	0,65±0,07	0,73±0,05	0,53±0,06	0,78±0,06
RL	0,84±0,03	0,68±0,03	0,91±0,03	0,70±0,05	0,77±0,02	0,58±0,02	0,80±0,03
NB-B	0,86±0,02	0,70±0,03	0,87±0,05	0,74±0,03	0,78±0,02	0,60±0,02	0,82±0,05
NB-G	0,79±0,02	0,59±0,02	0,77±0,08	0,69±0,04	0,63±0,04	0,53±0,03	0,75±0,04
NB-M	0,84±0,02	0,68±0,03	0,88±0,03	0,70±0,05	0,78±0,04	0,59±0,04	0,82±0,05
SVM-L	0,84±0,03	0,69±0,03	0,91±0,03	0,71±0,04	0,77±0,01	0,58±0,03	0,81±0,03
SVM-P	0,77±0,02	0,65±0,03	0,83±0,03	0,63±0,09	0,71±0,04	0,54±0,03	0,73±0,05
SVM-R	0,82±0,02	0,69±0,02	0,86±0,07	0,65±0,07	0,74±0,05	0,54±0,09	0,76±0,07

$$\chi_F^2 = \left[\frac{12}{nk(k+1)} \sum_{j=1}^{k} R_j^2 \right] - 3n(k+1). \qquad (2)$$

A hipótese nula pode ser rejeitada se χ_F^2 estiver em uma distribuição χ^2, com $k-1$ graus de confiança para $n > 15$ e $k > 5$. Para valores menores de n e k, a medida χ^2 aproximada é imprecisa, sendo necessário procurar valores em tabelas de χ_F^2 específicas para o teste de Friedman [38]. Assim, considerando um intervalo de confiança $\alpha = 0,01$, $n = 7$ e $k = 11$, o valor crítico é dado por 23,209. Portanto, como χ_F^2 = 57,6363, conclui-se que há uma diferença significativa entre os desempenhos obtidos pelos métodos de classificação avaliados e, portanto, a hipótese nula pode ser rejeitada.

Em seguida, o teste de Nemenyi [39] foi empregado para comparar o desempenho dos métodos em pares. Nesse cenário, para cada par de métodos, um valor q foi calculado com base na distribuição do *ranking* de desempenho dos métodos de classificação avaliados (Eq. 3). Nesse caso, a hipótese nula, que indica não haver diferença no desempenho entre os dois métodos, pode ser rejeitada se q for maior que o valor crítico $q\alpha$, obtido para o grau de confiança α desejado.

$$q = \frac{\overline{R_{j1}} - \overline{R_{j2}}}{\sqrt{\frac{k(k+1)}{6n}}}. \qquad (3)$$

O resultado do teste de Nemenyi demonstrou que o sistema de comitê de máquinas de classificação proposto neste trabalho é significativamente superior aos demais métodos avaliados ($p < 0,001$). É possível afirmar que sob as mesmas condições e bases de dados, com 99,9% de grau de confiança, o desempenho do sistema proposto é estatisticamente superior a qualquer outro método avaliado neste trabalho.

Finalmente, na Tabela V é apresentada a média do tempo de execução (em segundos) demandado para concluir as etapas seleção de modelo, treinamento e teste de cada método de classificação para cada base de dados. Os menores valores estão destacados em negrito. Tais experimentos foram realizados em um computador com processador Intel Xeon X3430 (quatro núcleos - 2,4GHz) e 8GB de RAM.

Conforme esperado, embora o comitê de máquinas de classificação tenha obtido resultados de predição estatisticamente superiores aos métodos individuais, ele consumiu maior tempo de processamento. Dado que o comitê combina várias técnicas de processamento de texto com diversos métodos de classificação, é natural que ele necessite de mais tempo para realizar as etapas de seleção de modelo e treinamento. No entanto, é importante destacar que em cenários nos quais esses processos, computacionalmente mais caros, possam ser executados *offline*, o comitê de máquinas de classificação é altamente recomendado devido ao seu grande poder de generalização e predição.

VII. CONCLUSÕES

Detectar automaticamente opiniões expressas em mensagens curtas e ruidosas postadas em redes sociais ainda é um desafio tanto na área de aprendizado de máquina, quanto em processamento de linguagem natural. Até mesmo os métodos considerados estado da arte em classificação encontram pelo menos duas grandes dificuldades ao lidar com esse tipo de problema: 1) a pequena quantidade de atributos extraídos por mensagem e 2) a baixa qualidade desses atributos, ocasionada pelo uso constante de abreviações, gírias e símbolos que podem gerar problemas de polissemia e sinonímia.

Para contornar essas dificuldades, neste trabalho foi proposto um sistema de comitê de máquinas de classificação, que combina automaticamente técnicas consolidadas de processamento de texto, normalização e indexação semântica com os métodos de classificação considerados o estado da arte.

Nesse cenário, foi desenvolvido e apresentado um sistema de normalização de texto e expansão por indexação semântica, que funciona com base em dicionários semânticos e léxicos, juntamente com técnicas de análise semântica e detecção de contexto. As amostras são inicialmente normalizadas e, posteriormente, novos atributos são gerados por meio de conceitos usados para expandir e enriquecer a amostra original. Com isso, problemas de polissemia e sinonímia podem ser evitados.

O sistema proposto foi avaliado com sete bases de dados públicas e os desempenhos obtidos foram comparados com os métodos de classificação individuais. Os resultados da análise estatística indicaram que, sob as mesmas condições e com grau de confiança igual à 99,9%, o comitê de classificadores foi estatisticamente superior a qualquer outro método de classificação avaliado. Contudo, ele demanda maior tempo computacional nas etapas de seleção de modelo e treinamento.

Atualmente, o sistema proposto está sendo avaliado em contextos com características semelhantes aos descritos neste trabalho, tais como filtragem de comentários curtos, além de outras aplicações Web, como comentários postados no Youtube, fóruns e blogs. Como trabalho futuro, pretende-se projetar uma versão paralelizada do sistema para acelerar as etapas que demandam maior custo computacional.

Tabela V. MÉDIA E DESVIO PADRÃO DO TEMPO DE EXECUÇÃO (EM SEGUNDOS) CONSUMIDO NAS ETAPAS DE SELEÇÃO DE MODELO, TREINAMENTO E TESTE DE CADA MÉTODO DE CLASSIFICAÇÃO PARA CADA BASE DE DADOS.

Métodos	Archeage	HCR	Hobbit	IPhone6	OMD	SS-Tweet	STS-Test
B. C4.5	2,30±1,85	2,90±2,55	2,50±0,92	1,50±1,43	8,00±4,12	16,60±9,81	1,20±1,33
C4.5	0,70±0,46	0,40±0,49	0,40±0,49	0,20±0,40	0,40±0,49	1,40±0,49	0,20±0,40
Comitê	93,10±23,74	41,10±5,50	7,60±1,36	7,70±2,19	78,60±17,56	386,10±49,31	6,50±0,92
KNN	8,20±2,27	2,70±1,10	0,80±0,60	0,70±0,46	7,90±0,94	29,40±7,49	0,70±0,46
RL	0,50±0,50	0,30±0,46	0,10±0,30	0,10±0,30	**0,30±0,46**	1,60±0,49	**0,00±0,00**
NB-B	5,40±1,62	2,50±0,50	1,40±0,49	1,40±0,49	3,90±0,94	13,00±1,79	1,50±0,50
NB-G	**0,40±0,49**	**0,10±0,30**	**0,00±0,00**	**0,00±0,00**	**0,30±0,46**	1,50±0,81	**0,00±0,00**
NB-M	1,00±0,63	0,40±0,49	0,20±0,40	0,60±0,49	0,60±0,49	2,50±0,92	0,20±0,40
SVM-L	0,70±0,64	0,50±0,50	**0,00±0,00**	0,40±0,49	0,70±0,46	**1,30±0,46**	0,20±0,40
SVM-P	56,50±14,81	24,50±4,41	1,90±0,83	2,30±1,10	43,90±12,09	248,00±34,09	1,90±0,54
SVM-R	24,20±6,03	10,60±1,96	0,60±0,66	0,90±0,70	21,90±4,41	110,40±16,17	0,80±0,60

AGRADECIMENTOS

Os autores são gratos a FAPESP (processo 2014/01237-7) e Capes pelo apoio financeiro ao desenvolvimento desse projeto.

REFERÊNCIAS

[1] K. Denecke, "Using SentiWordNet for multilingual sentiment analysis," in *Proc.of the 2008 ICDEW*, Cancun, Mexico, 2008, pp. 507–512.

[2] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? sentiment classification using machine learning techniques," in *Proc. of the 2002 ACL EMNLP*, Philadelphia, USA, 2002, pp. 79–86.

[3] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," *Found. and Trends in Inform. Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.

[4] M. M. Mostafa, "More than words: Social networks' text mining for consumer brand sentiments," *Exp. Systems with Appl.*, vol. 40, no. 10, pp. 4241–4251, 2013.

[5] R. Navigli and S. P. Ponzetto, "Multilingual WSD with just a few lines of code: the BabelNet API," in *Proc. of the 2012 ACL*, Jeju Island, South Korea, 2012, pp. 67–72.

[6] M. A. H. Taieb, M. B. Aouicha, and A. B. Hamadou, "Computing semantic relatedness using wikipedia features," *Knowledge-Based Systems*, vol. 50, no. 9, pp. 260–278, 2013.

[7] E. Kontopoulos, C. Berberidis, T. Dergiades, and N. Bassiliades, "Ontology-based sentiment analysis of Twitter posts," *Exp. Systems with Appl.*, vol. 40, no. 10, pp. 4065–4074, 2013.

[8] V. Nastase and M. Strube, "Transforming Wikipedia into a large scale multilingual concept network," *Art. Intell.*, vol. 194, pp. 62–85, 2013.

[9] R. Xia, C. Zong, and S. Li, "Ensemble of feature sets and classification algorithms for sentiment classification," *Information Sciences*, vol. 181, no. 6, pp. 1138 – 1152, 2011.

[10] G. Wang, J. Sun, J. Ma, K. Xu, and J. Gu, "Sentiment classification: The contribution of ensemble learning," *Decision Support Systems*, vol. 57, pp. 77 – 93, 2014.

[11] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," Stanfod University, Tech. Rep., 2009.

[12] R. Navigli and M. Lapata, "An experimental study of graph connectivity for unsupervised word sense disambiguation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 32, no. 4, pp. 678–692, 2010.

[13] T. C. Alberto and T. A. Almeida, "Aprendizado de Máquina Aplicado na Detecção Automática de Comentários Indesejados," in *Proc. of the 10th ENIAC*, Fortaleza, Brazil, 2013, pp. 1–6.

[14] T. C. Alberto, J. V. Lochter, and T. A. Almeida, "Post or block? advances in automatically filtering undesired comments," *Journal of Intelligent & Robotic Systems*, pp. 1–15, 2014.

[15] T. P. Silva, I. Santos, T. A. Almeida, and J. M. Gómez Hidalgo, "Normalização Textual e Indexação Semântica Aplicadas na Filtragem de SMS Spam," in *Proc. of the 11st ENIAC*, São Carlos, Brazil, 2014.

[16] P. Cook and S. Stevenson, "An unsupervised model for text message normalization," in *Proc. of the 2009 CALC*. Association for Computational Linguistics, 2009, pp. 71–78.

[17] Z. Xue, D. Yin, B. D. Davison, and B. Davison, "Normalizing Microtext," in *Proc. of the 2011 AAAI*, San Francisco, USA, 2011.

[18] J. M. Gómez Hidalgo, M. Buenaga Rodríguez, and J. C. Cortizo Pérez, "The role of word sense disambiguation in automated text categorization," in *Proc. of the 10th NLDB*, Alicante, Spain, 2005.

[19] E. Agirre and P. Edmonds, *Word sense disambiguation*. Springer, 2006.

[20] T. G. Dietterich, "Ensemble methods in machine learning," in *Proc. of 2000 MCS*, ser. MCS '00, Cagliari, Italy, 2000, pp. 1–15.

[21] L. K. Hansen and P. Salamon, "Neural network ensembles," *IEEE Trans. Pattern Anal. Machine Intelligence*, vol. 12, no. 10, pp. 993–1001, 1990.

[22] H. Kim, H. Kim, H. Moon, and H. Ahn, "A weight-adjusted voting algorithm for ensembles of classifiers," *Journal of the Korean Statistical Society*, vol. 40, no. 4, pp. 437–449, 2011.

[23] E. Gabrilovich and S. Markovitch, "Feature Generation for Text Categorization Using World Knowledge," in *Proc. of the 19th IJCAI*, Edinburgh, Scotland, 2005, pp. 1048–1053.

[24] ——, "Harnessing the expertise of 70,000 human editors: Knowledge-based feature generation for text categorization," *Journal of Machine Learning Research*, vol. 8, pp. 2297–2345, 2007.

[25] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment analysis of twitter data," in *Proc. of 2011 LSM*, Portland, USA, 2011, pp. 30–38.

[26] M. Speriosu, N. Sudan, S. Upadhyay, and J. Baldridge, "Twitter polarity classification with label propagation over lexical links and the follower graph," in *Proc. of 2011 EMNLP*, Edinburgh, Scotland, 2011, pp. 53–63.

[27] D. A. Shamma, L. Kennedy, and E. F. Churchill, "Tweet the debates: Understanding community annotation of uncollected sources," in *Proc. of 2009 WSM*, Beijing, China, 2009, pp. 3–10.

[28] M. Thelwall, K. Buckley, and G. Paltoglou, "Sentiment strength detection for the social web," *J. Am. Soc. Inf. Sci. Technol.*, vol. 63, no. 1, pp. 163–173, 2012.

[29] H. Saif, M. Fernández, Y. He, and H. Alani, "Evaluation datasets for twitter sentiment analysis: a survey and a new dataset, the sts-gold," in *Proc. of the 1st ESSEM*, Turin, Italy, 2013.

[30] X. Wu, V. Kumar, J. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. McLachlan, A. Ng, B. Liu, P. Yu, Z.-H. Zhou, M. Steinbach, D. Hand, and D. Steinberg, "Top 10 algorithms in data mining," *KAIS*, vol. 14, no. 1, pp. 1–37, 2008.

[31] T. Almeida, J. Almeida, and A. Yamakami, "Spam filtering: How the dimensionality reduction affects the accuracy of naive bayes classifiers," *JISA*, vol. 1, no. 3, pp. 183–200, 2011.

[32] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[33] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd ed. New York, NY, USA: Prentice Hall, 1998.

[34] J. Quinlan, *C4.5: programs for machine learning*, 1st ed. San Mateo, CA, USA: Morgan Kaufmann, 1993.

[35] D. Aha, D. Kibler, and M. Albert, "Instance-based learning algorithms," *Machine Learning*, vol. 6, no. 1, pp. 37–66, 1991.

[36] Y. Freund and R. Schapire, "Experiments with a new boosting algorithm," in *Proc. of the 13rd ICML*, Bari, Italy, 1996, pp. 148–156.

[37] M. Friedman, "A comparison of alternative tests of significance for the problem of m rankings," *Ann. Math. Statist.*, vol. 11, no. 1, 1940.

[38] N. Japkowicz and M. Shah, *Evaluating Learning Algorithms - A Classification Perspective*. Cambridge University Press, 2011.

[39] P. F. Nemenyi, "Distribution-free multiple comparisons," Ph.D. dissertation, Princeton University, 1963.