Normalização Textual e Indexação Semântica Aplicadas na Filtragem de SMS Spam

Tiago P. Silva*, Igor Santos†, Tiago A. Almeida* e José M. Gómez Hidalgo‡
* *Universidade Federal de São Carlos (UFSCar), Sorocaba, São Paulo, Brasil*
Email: tpsilva@acm.org, talmeida@ufscar.br
† *Universidad de Deusto, Bilbao, Espanha*
Email: isantos@deusto.es
‡ *Optenet, Madrid, Espanha*
Email: jgomez@optenet.com

***Resumo*—Nos últimos anos, a popularização dos celulares e *smartphones* impulsionou o uso de SMS como forma alternativa e barata de comunicação. O crescimento de adeptos ao serviço aliado a alta confiança que os usuários possuem nesses tipos de mensagens, vêm atraindo a atenção de pessoas e empresas mal intencionadas, conhecidas como *spammers*. O spam nesse contexto representa um problema para os métodos tradicionais e já consolidados, pois tais técnicas, normalmente projetadas para processar e-mails, geralmente não obtém desempenho satisfatório quando aplicadas diretamente para classificar SMS, uma vez que essas mensagens tem tamanho reduzido e conteúdo normalmente repleto de gírias, símbolos e abreviações. Nesse cenário, este artigo apresenta um método baseado em normalização de textos e indexação semântica com o intuito de melhorar o desempenho de algoritmos de classificação tradicionais na filtragem de spam propagados via SMS. O método proposto é utilizado para normalizar os termos das mensagens e criar novos atributos, de forma a alterar e expandir as amostrar originais, com o objetivo de suavizar fatores que podem degradar o desempenho dos algoritmos de classificação, como redundâncias e inconsistências. Os experimentos foram conduzidos com uma base de dados real, pública e não codificada, além de vários métodos tradicionais de aprendizado de máquina. A análise estatística dos resultados indica que o emprego da técnica proposta, de fato, melhora a qualidade da predição das mensagens.**

***Keywords*-filtragem de *spam*; processamento de linguagem natural; aprendizado de máquina; classificação.**

I. Introdução

O serviço de mensagem curta, do inglês *Short Message Service* (SMS), possibilita a comunicação entre celulares ou telefones fixos através de mensagens de texto. Ele geralmente é utilizado como substituto das ligações em situações nas quais a comunicação por voz não é desejada. As mensagens são bastante populares em alguns lugares do mundo por serem mais baratas do que as ligações por voz.

Nos últimos anos, a indústria do SMS vêm se tornando enorme. De acordo com o relatório da empresa *Portio Research*[1], o faturamento mundial com SMS atingiu a marca de 128 bilhões de dólares em 2011, sendo que a receita estimada para 2016 é de mais de 153 bilhões de dólares. O mesmo documento indica que em 2011 foram enviadas mais de 7,8 trilhões de SMS no mundo todo e a estimativa é de que somente em 2014 sejam enviados 9,5 trilhões.

O aumento da popularidade do SMS fez com que suas taxas caíssem para menos de um centavo de dólar em mercados como a China, e chega a ser grátis em alguns países da Ásia. Além disso, em muitos outros mercados, o aumento explosivo na comunicação por mensagens de texto, acompanhado dos planos de telefonia com mensagens ilimitadas, diminuiu consideravelmente o custo de envio de SMS. Este fato, combinado com a confiança que os usuários tem em seus dispositivos móveis torna o ambiente propício para a disseminação de mensagens indesejadas. Como consequência, nos últimos anos os telefones móveis estão se tornando o principal alvo de *spammers*. SMS spam, também chamado de spam móvel, é o nome dado a qualquer mensagem de texto indesejada enviada para um celular ou telefone fixo. Esta prática, que se tornou muito popular em algumas partes da Ásia, está se espalhando rapidamente também nos países ocidentais[2].

As mensagens de spam via SMS, além de serem indesejadas, podem ser custosas, pois existem planos de operadoras nos quais os usuários pagam para receber mensagens. Além disso, a filtragem automática de SMS spam ainda está engatinhando no cenário mundial, pois existem poucas opções de software disponíveis, e também há a preocupação de que mensagens de emergência possam ser bloqueadas. No entanto, muitas operadoras estão investindo em meios de atenuar este problema.

Da mesma forma que as operadoras estão enfrentando muitos problemas ao lidar com SMS spam, no meio acadêmico também há muitas dificuldades. Uma das preocupações é que filtros consolidados para bloquear spam via e-mail vêm apresentando desempenho degradado quando aplicados na filtragem de SMS spam. Isso ocorre devido ao tamanho limitado das mensagens, que possuem apenas 140 bytes, o que representa 160 caracteres. Além disso, tais mensagens são geralmente repletas de erros de digitação, gírias, símbolos, *emoticons*, e abreviações, que tornam até mesmo a *tokenização* uma tarefa difícil.

[1] *Mobile Messaging Futures 2012-2016.* Disponível em http://www.portioresearch.com/en/market-briefings/budget-reports/mobile-messaging-futures-2012-2016.aspx.

[2] Relatório anual da *Cloudmark*. Disponível em http://lx.pe/b71y.

Neste cenário, este trabalho apresenta um método de normalização textual e indexação semântica com o objetivo de melhorar o desempenho de classificadores na filtragem de SMS spam. A premissa básica é que este processamento pode aumentar a informação semântica das amostras e, consequentemente, melhorar a qualidade das predições. Para realizar a análise semântica das amostras, foi criado um processo em cascata, no qual são extraídas relações semânticas do dicionário léxico BabelNet [1] e é empregada uma etapa de desambiguação [2], de forma a obter atributos representativos. Posteriormente, o conteúdo do SMS original é expandido com as informações produzidas, que são utilizadas pelos métodos tradicionais de aprendizado de máquina.

O restante deste artigo está organizado da seguinte forma: na Seção II, são apresentados os trabalhos correlatos à proposta deste trabalho. A Seção III descreve o método de expansão utilizado. Na Seção IV, são detalhados o conjunto de dados, as medidas de desempenho e os parâmetros utilizados nos experimentos. A Seção V apresenta os resultados obtidos e a análise estatística realizada. Finalmente, na Seção VI, são oferecidas as principais conclusões e propostas para trabalhos futuros.

II. Trabalhos correlatos

Ao contrário do grande número de trabalhos disponíveis sobre classificação de spam via e-mail, existem poucas propostas para filtragem de SMS spam. A seguir, são brevemente descritos os trabalhos mais relevantes relacionados a esse tema.

Em [3], foi avaliado o emprego de diferentes classificadores Bayesianos para detecção de spam em telefones móveis. Os autores propuseram duas pequenas bases de dados de SMS spam e também verificaram o desempenho de diferentes técnicas de representação. Os resultados indicaram que os filtros Bayesianos podem ser utilizados na tarefa de filtragem de SMS spam, porém o desempenho obtido foi inferior ao de filtragem de spam via e-mail.

Em [4] e [5], foi analisado o problema da classificação de *spam* em três cenários distintos: SMS, comentários em blogs e somente através do assunto dos emails. A principal conclusão foi que mensagens curtas contém uma quantidade insuficiente de atributos para suportar o uso do *bag of words*. O uso de bigramas ortogonais, bigramas e trigramas melhorou o desempenho dos filtros avaliados. Porém, os autores evidenciaram que a presença de ruídos (abreviações, símbolos, gírias, etc) contribui negativamente para a qualidade das predições.

Em [6] e [7], foi apresentada uma nova base de dados pública de SMS spam, chamada *SMS Spam Collection*, composta por um número significativo de amostras reais. Os autores avaliaram o desempenho de diversos métodos tradicionais de aprendizado de máquina e concluíram que o SVM linear é o melhor *baseline* para comparações futuras.

Em [8], foi avaliado um software de detecção de SMS spam aplicado diretamente nos aparelhos celulares. Os resultados reportados indicam que o sistema proposto foi capaz de obter uma acurácia razoável, consumo de armazenamento mínimo e um tempo de processamento aceitável, sem a necessidade de um computador ou de uma grande base de dados para treinamento.

Dentre os trabalhos analisados, é praticamente unânime a constatação de que o tamanho reduzido e o conteúdo poluído das mensagens de SMS degradam consideravelmente o desempenho dos métodos tradicionais de aprendizado de máquina. Nesse cenário, este artigo propõem um método de normalização de texto e indexação semântica para processar os atributos originais extraídos das mensagens e fornecer mais informações para os classificadores. Tal método está relacionado com duas principais áreas de pesquisa:

1) uso de técnicas de linguagem natural para normalização léxica de texto [9]; e
2) uso de bases de dados léxicas e dicionários semânticos na representação de textos para classificação [10].

Normalização léxica é o nome dado à tarefa de traduzir variantes léxicas de palavras e expressões normalmente ofuscadas para sua forma canônica, de forma a facilitar o processamento do texto. Por exemplo, termos como "*goooood*" e "*b4*" podem ser traduzidos para as palavras inglesas "*good*" e "*before*", respectivamente.

A normalização léxica está relacionada à verificação ortográfica e, muitas abordagens na literatura compartilham técnicas para essa tarefa. Por exemplo, [11] e [12] propõem diversos modelos simples, no qual cada um captura uma forma particular de formação de variantes léxicas, como por fonética (por exemplo, *epik* – "*epic*") ou encurtamentos (por exemplo, *goin* – "*going*").

Os trabalhos disponíveis na literatura que estão mais relacionados com a proposta deste trabalho são [13], [14] e [15]. Em todos os casos, o problema é tratado como uma tarefa de tradução automática, no qual o objetivo é traduzir textos ruidosos para o inglês. Tais trabalhos usam modelos de linguagens sofisticados, treinados em amostras de texto ruidosas, enquanto que a abordagem proposta neste trabalho emprega um modelo relativamente simples de equivalência semântica e normalização textual.

Quanto ao *uso de bases de dados léxicas em classificação de texto*, existem vários trabalhos que empregam a base WordNet [16] em tarefas como recuperação de informação [17], categorização de texto [10], agrupamento de texto [18], entre outros. O uso dessas bases de dados léxicas adiciona a complexidade de identificar o significado correto (ou conceito apropriado) para cada palavra, um problema que é chamado de desambiguação (em inglês, *word sense disambiguation – WSD*) [2].

Neste trabalho, foi utilizada a base de dados léxica BabelNet [1], que por ser mais recente é bem maior e mais atualizada que a WordNet. Também foi utilizado o algoritmo de desambiguação proposto em [19], seguindo o método de expansão semântica descrito em [10], mas aplicado a documentos de texto ao invés de categorias.

III. MÉTODO DE EXPANSÃO

A aplicação direta de modelos superficiais de representação de texto, como o *bag of words* tradicional, têm sido apontada como um dos principais fatores na limitação de desempenho dos algoritmos de aprendizado de máquina em problemas de classificação de textos curtos [20], [21]. Para contornar essa deficiência, este trabalho propõem um método para normalizar o texto e expandir o volume de atributos representativos de cada amostra e, consequentemente, aumentar a capacidade de predição dos métodos de classificação quando aplicados na filtragem de SMS spam.

O método de expansão usa técnicas recentes para normalização léxica e detecção de contexto, além de dicionários semânticos para criar traduções das amostras originais. Neste trabalho, cada amostra foi expandida em três fases distintas:

- *Tradução de Lingo*: utilizada para traduzir palavras em Lingo, que é o nome dado às gírias e abreviações geralmente utilizadas em formas de comunicação online e móvel, para inglês.
- *Geração de conceitos*: utilizada para obter todos os conceitos relacionados a uma palavra, ou seja, cada possível significado desta palavra.
- *Desambiguação*: utilizada para encontrar o conceito mais relevante, de acordo com o contexto da mensagem, entre todos os conceitos relacionados a uma determinada palavra.

As fases de *Geração de conceitos* e *Desambiguação* utilizam a base de dados léxica BabelNet, que é um grande repositório semântico [1]. Enquanto que a fase de *Geração de conceitos* consiste em trocar uma determinada palavra por todos os conceitos relacionados a ela, a fase de *Desambiguação* seleciona automaticamente o conceito mais relevante para cada palavra. Esta escolha é feita por meio de análise semântica, através do contexto em que a palavra é encontrada na amostra.

O método proposto expande uma amostra de texto processando cada *token* da mensagem nas fases descritas, gerando novas amostras expandidas. Dessa forma, através de uma regra de combinação pré-definida, a amostra final expandida é então obtida na saída do método. A Figura 1 ilustra o processo.

A seguir, são detalhadas cada uma das três possíveis etapas de expansão:

A. Tradução de Lingo

Nesta fase são utilizados dois dicionários. O primeiro é um dicionário de inglês, que é utilizado para verificar se uma palavra qualquer pertence a língua inglesa. O segundo é o dicionário de Lingo em si, que é utilizado para traduzir um termo de Lingo (gírias, abreviações, símbolos, etc) para o inglês. O processo consiste em procurar cada *token* extraído da amostra no dicionário de inglês, que neste caso é utilizado

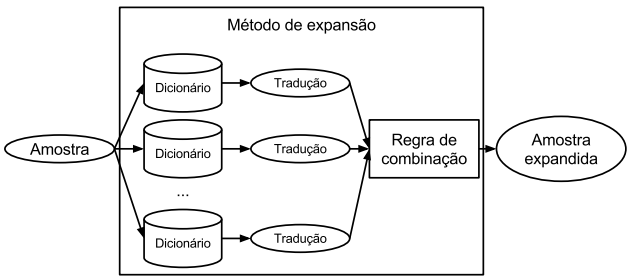

Figura 1. A amostra original é processada por dicionários semânticos e técnicas de detecção de contexto, no qual cada um gera uma nova versão traduzida ou expandida da amostra. Depois disso, de acordo com uma regra de combinação, as amostras expandidas são combinadas para gerar a amostra final.

o dicionário *Freeling*[3]. Se for encontrado, o processo parte para o próximo *token*, pois já trata-se de uma palavra da língua inglesa. Caso contrário, o método utiliza o dicionário de Lingo, que neste caso é o dicionário NoSlang[4] para tentar traduzir o *token*.

B. Geração de conceitos

Os conceitos semânticos são obtidos do repositório BabelNet e, portanto, uma vez que o repositório exige como entrada uma palavra em inglês, o método não utiliza os termos originais nesta fase. Ao invés disso, o método primeiramente realiza a *tradução de Lingo* para certificar que cada *token* foi traduzido para o inglês. Em seguida, para evitar possíveis ruídos nos conceitos, o método exclui palavras presentes em uma lista de *stopwords*, que contém artigos, pronomes e preposições. As palavras restantes na amostra são então analisadas semanticamente para encontrar seus conceitos.

C. Desambiguação

Como a fase de *geração de conceitos* pode produzir um número muito grande de conceitos para cada palavra, foi implementada uma fase de desambiguação, que utiliza o algoritmo proposto em [22]. Em resumo, a técnica procura pelo conceito mais relevante, de acordo com o contexto da amostra. Basicamente, para cada uma das palavras na amostra de entrada, o algoritmo utiliza técnicas de análise semântica e contextual para atribuir notas a todos os conceitos obtidos do repositório BabelNet. O método então seleciona o conceito com a maior nota para ser a tradução da palavra. Tal nota é obtida através do cálculo do número de distâncias definido na rede semântica presente no BabelNet.

Exemplo de expansão

A Tabela I apresenta um exemplo de expansão para uma amostra de texto curto e ruidoso. Considerando a mensagem original "*plz lemme noe when u get der*", são exibidas

[3]*Dicionário de inglês Freeling*. Disponível em: http://devel.cpl.upc.edu/freeling/.

[4]*NoSlang*: Dicionário de gírias. Disponível em: http://www.noslang.com/dictionary/full/.

as saídas de cada uma das três fases de expansão. Assumindo que a regra de combinação seja {*Tradução de Lingo* + *Desambiguação*}, a amostra expandida resultante seria "*please favor let me know cognition when you get there*", que poderia ser utilizada por algoritmos de aprendizado de máquina e, possivelmente obtendo melhor desempenho de classificação do que com a amostra original.

Tabela I
EXEMPLO DE EXPANSÃO DE UMA AMOSTRA DE TEXTO CURTO. CADA LINHA REPRESENTA A SAÍDA OBTIDA EM UMA FASE DO PROCESSO DE EXPANSÃO E A ÚLTIMA LINHA APRESENTA A AMOSTRA EXPANDIDA, OBTIDA ATRAVÉS DA COMBINAÇÃO DAS SAÍDAS DE TRADUÇÃO DE LINGO E DESAMBIGUAÇÃO.

Original	*plz lemme noe when u get der*
Tradução Lingo	*please let me know when you get there*
Geração de conceitos	*care delight favor gratify like please please_(album) please_(toni_braxton_song) please_(u2_song) satisfy wish army_of_the_pure army_of_the_righteous lashkar-e-taiba lashkar-e-tayyiba lashkar-e-toiba let let_(rave_master) net_ball acknowledge cognition distinguish experience know knowledge noesis when you get get_(animal) get_(conflict) there*
Desambiguação	*favor let me cognition when you get there*
Amostra final	*please favor let me know cognition when you get there*

Conforme exposto na Tabela I, a *Tradução de Lingo* substitui as gírias, símbolos e abreviações por palavras correspondentes em inglês. Enquanto que a *Geração de conceitos* obtém todos os conceitos relacionados à cada uma das palavras traduzidas da amostra original, a *Desambiguação* mantém somente os conceitos que são semanticamente relevantes à amostra original. Finalmente, utilizando a saída combinada, problemas semânticos tradicionais, como polissemia e sinonímia [21] podem ser evitados e, consequentemente, resultados melhores podem ser atingidos ao utilizar técnicas tradicionais de aprendizado de máquina.

É importante observar que, como o processo de expansão é parametrizado pela regra de combinação das saídas dos três estágios (com ou sem: *Tradução de Lingo*, *Geração de conceitos* e *Desambiguação*), além da permanência ou não dos *tokens* originais, existem dez possibilidades distintas de realizar a expansão da amostra original.

IV. EXPERIMENTOS

Para avaliar a eficácia do método de expansão proposto, ele foi aplicado em um problema de classificação chamado filtragem de spam em SMS. Para esta tarefa, foi utilizada a base de dados pública *SMS Spam Collection* [6], que é composta por 5.574 mensagens reais e não codificadas, escritas em inglês, previamente rotuladas como legítimas (*ham*) ou *spam*. É importante ressaltar que os criadores da base de dados mostraram que os métodos de classificação tradicionais podem ter o desempenho prejudicado, uma vez que as mensagens originais são relativamente curtas (limitadas a um máximo de 160 caracteres) e seu conteúdo é repleto de gírias e abreviações [6]. Essas mesmas características podem ser encontradas em redes sociais, fóruns, *chats*, entre outras formas de comunicação *on-line*.

Nos experimentos, foram utilizadas todas as possíveis regras de combinação para o método de expansão proposto, gerando uma base de dados expandida para cada conjunto de parâmetros. Dessa forma, foram avaliadas a base de dados original e dez expandidas, totalizando onze bases. Além disso, para cada uma das bases de dados geradas, foi avaliado o desempenho de diversas técnicas tradicionais de aprendizado de máquina, com o intuito de verificar se as técnicas propostas de expansão podem prover melhorias no desempenho dos métodos.

A Tabela II apresenta os métodos de classificação que foram empregados. É importante notar que, para dar credibilidade aos experimentos, foram avaliados métodos que utilizam diferentes estratégias de geração de hipóteses, como distância, árvores, otimização, entre outros.

Tabela II
TÉCNICAS DE CLASSIFICAÇÃO USADAS PARA VERIFICAR SE O MÉTODO DE EXPANSÃO PODE MELHORAR A QUALIDADE DAS PREDIÇÕES.

Regressão logística (Logistic) [23]
k-vizinhos mais próximos (k-NN) [3]
Naive Bayes (NB) [24]
Classificação baseada em regras (PART) [3]
Otimização mínima sequencial (SMO) [25]
SVM linear (L.SVM) [26]
Árvores de decisão (C4.5) [3]
Boosting de árvores de decisão (B.C4.5) [3]
Boosted Naive Bayes (B.NB) [27]
Bagging de árvores de decisão (Bagging) [28]

Todos os métodos avaliados estão disponíveis na biblioteca de aprendizado de máquina WEKA [29] e em todos os experimentos, foram empregados os parâmetros padrões, com exceção do algoritmo k-NN, no qual foram avaliados $k = 1$, 3 e 5.

Os experimentos foram realizados por meio de validação cruzada com cinco partições. Para *tokenizar* as mensagens, foram utilizados como delimitadores pontos, vírgulas, tabulações e espaços. Para comparar os resultados, foi empregado o Coeficiente de Correlação de Matthews (do inglês, *Matthews Correlation Coefficient – MCC*), que avalia a qualidade de uma classificação binária. O MCC retorna um valor real entre -1 e $+1$, no qual um coeficiente igual a $+1$ indica uma classificação perfeita; 0 uma classificação aleatória; e -1 uma classificação inversa [24].

V. RESULTADOS

Para cada método de classificação avaliado (Tabela II), foram coletados os valores de MCC obtidos com a base de dados original (*Original*), bem como com a base obtida pelo processo de expansão (*Expansão*) cuja regra de combinação obteve o melhor desempenho para cada classificador em questão. Tal processo pode ser visto como um *tunning* de

parâmetro na etapa de expansão e o emprego do valor que obteve o melhor desempenho.

Os resultados foram analisados estatisticamente para verificar se o processo de expansão realmente oferece ganho de desempenho às técnicas de classificação. Para tal propósito, foi utilizado o teste não-paramétrico de Mann-Whitney (*Wilcoxon Signed-Ranks Test*) [30]. Esse teste cria um *ranking* dos algoritmos de acordo com as diferenças absolutas entre os resultados obtidos com cada uma das bases de dados. Em seguida, são calculadas as somas das posições cuja diferença foi negativa e positiva.

A Tabela III apresenta o valor do MCC obtido por cada um dos classificadores nas bases de dados *Original* e *Expansão*, juntamente com a diferença entre os resultados e o *ranking* dos algoritmos.

Tabela III
MCC OBTIDO POR CADA UM DOS CLASSIFICADORES NAS BASES DE DADOS ORIGINAL E EXPANSÃO, DIFERENÇA ENTRE OS RESULTADOS OBTIDOS USANDO AS DUAS BASES DE DADOS E POSIÇÃO (RANK) DO ALGORITMO NO TESTE DE MANN-WHITNEY.

Classificador	**MCC**		**Diferença**	**Rank**
	Orig.	**Exp.**		
SMO	0,929	0,927	0,002	1,5
L.SVM	0,929	0,927	0,002	1,5
NB	0,864	0,870	-0,006	3
B.C4.5	0,915	0,922	-0,007	4,5
Bagging	0,833	0,840	-0,007	4,5
B.NB	0,903	0,912	-0,009	6
1-NN	0,771	0,800	-0,029	7
PART	0,819	0,851	-0,032	8
C4.5	0,802	0,838	-0,036	9
Logistic	0,638	0,715	-0,077	10
3-NN	0,572	0,707	-0,135	11
5-NN	0,448	0,595	-0,147	12

Para computar as diferenças, foram utilizados os índices $R+$ e $R-$, que correspondem à soma das posições cuja diferença é positiva e negativa, respectivamente. Neste casso, $R+ = 3$ e $R- = 75$.

O objetivo é verificar se a hipótese nula pode ser rejeitada, o que neste caso indica que não existem diferenças estatísticas entre os resultados obtidos com a base de dados expandida e a original. Para o teste de Mann-Whitney, a hipótese nula é rejeitada com $\alpha = 0,05$, isto é, com um nível de confiança de 95%, quando $z \leq -1,96$. A equação de z é dada por:

$$z = \frac{T - \frac{1}{4}N(N+1)}{\sqrt{\frac{1}{24}N(N+1)(2N+2)}},$$

sendo $T = min(R+, R-)$ e N o número de métodos avaliados (a mesma técnica com diferentes parâmetros é contabilizada como um método distinto).

Neste caso, como $T = 3$ e $N = 12$, então $z = -2,77$. Isso significa que a hipótese nula pode ser claramente rejeitada e que os resultados obtidos pelos classificadores utilizando as amostras expandidas são estatisticamente superiores aos resultados obtidos com as amostras originais. Portanto, para a base de dados empregada, o emprego do método de expansão proposto resultou no aumento do desempenho dos classificadores.

VI. CONCLUSÕES E TRABALHOS FUTUROS

Este artigo apresentou um problema de classificação conhecido como filtragem de spam em SMS. Este problema tem se tornado um desafio para os métodos tradicionais de aprendizado de máquina, pois essas mensagens, além de serem curtas e possuírem poucos atributos, são geralmente repletas de gírias, símbolos e acrônimos. Consequentemente, nesses cenários, mesmo os métodos mais estabelecidos podem ter o desempenho prejudicado.

Como uma forma de contornar este problema, foi utilizado um método de normalização de textos que fornece informação semântica às amostras de SMS. O método utilizado é baseado em dicionários semânticos e lexicográficos, além de técnicas de análise semântica e detecção de contexto. Ele foi utilizado para normalizar os termos extraídos das mensagens e criar novos atributos, de forma a modificar e expandir as amostras originais, com o objetivo de amenizar fatores como redundâncias e inconsistências.

Foram analisados os desempenhos obtidos por métodos consolidados de classificação sob uma base de dados real, pública e não codificada de amostras de textos curtos e ruidosos. Finalmente, a análise estatística dos resultados indicou que o uso do método de expansão pode efetivamente oferecer melhorias no desempenho das técnicas de aprendizado de máquina nesta aplicação.

Trabalhos futuros compreendem a avaliação da técnica de expansão em aplicações com características similares à estudada neste trabalho, como classificação de comentários postados em blogs e redes sociais, ambos baseados no conteúdo das mensagens. Além disso, prevê-se o emprego do método de expansão em diferentes tarefas de aprendizado, como agrupamento e recomendação.

AGRADECIMENTOS

Os autores são gratos a Fapesp e CNPq pelo apoio financeiro ao desenvolvimento desse projeto.

REFERÊNCIAS

[1] R. Navigli and S. P. Ponzetto, "BabelNet: Building a very large multilingual semantic network," in *Proc. of the 48th ACL*, Uppsala, Sweden, 2010, pp. 216–225.

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

[3] J. M. Gómez Hidalgo, G. Cajigas Bringas, E. Puertas Sanz, and F. Carrero García, "Content Based SMS Spam Filtering," in *Proc. of the 2006 ACM DOCENG*, Amsterdam, The Netherlands, 2006, pp. 107–114.

[4] G. V. Cormack, J. M. Gómez Hidalgo, and E. Puertas Sanz, "Feature Engineering for Mobile (SMS) Spam Filtering," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, USA, 2007, pp. 871–872.

[5] ——, "Spam Filtering for Short Messages," in *Proceedings of the 16th ACM Conference on Conference on information and Knowledge Management*, Lisbon, Portugal, 2007, pp. 313–320.

[6] T. A. Almeida, J. M. Gómez Hidalgo, and A. Yamakami, "Contributions to the study of SMS spam filtering: new collection and results," in *Proc. of the 11th ACM DOCENG*, Mountain View, California, USA, 2011, pp. 259–262.

[7] J. M. Gómez Hidalgo, T. A. Almeida, and A. Yamakami, "On the Validity of a New SMS Spam Collection," in *Proc. of the 2012 IEEE ICMLA*, Boca Raton, FL, USA, 2012, pp. 240–245.

[8] M. Taufiq Nuruzzaman, C. Lee, M. F. A. b. Abdullah, and D. Choi, "Simple sms spam filtering on independent mobile phone," *Security and Communication Networks*, vol. 5, no. 10, pp. 1209–1220, 2012.

[9] B. Han, P. Cook, and T. Baldwin, "Lexical Normalization for Social Media Text," *ACM Trans. Intell. Syst. Technol.*, vol. 4, no. 1, pp. 1–27, 2013.

[10] 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, pp. 298–309.

[11] 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.

[12] Z. Xue, D. Yin, B. D. Davison, and B. Davison, "Normalizing Microtext," in *Proc. of the 2011 AAAI*. Association for the Advancement of Artificial Intelligence, 2011, pp. 74–79.

[13] A. Aw, M. Zhang, J. Xiao, and J. Su, "A Phrase-based Statistical Model for SMS Text Normalization," in *Proc. of the 2006 COLING/ACL*. Association for Computational Linguistics, 2006, pp. 33–40.

[14] C. Henríquez and A. Hernández, "A ngram-based statistical machine translation approach for text normalization on chat-speak style communications," in *Proc. of the 2009 CAW2*, 2009.

[15] J. Kaufmann and J. Kalita, "Syntactic Normalization of Twitter Messages," in *Proc. of the 2010 ICON*, 2010.

[16] G. A. Miller, "Wordnet: a lexical database for English," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[17] J. Gonzalo, F. Verdejo, I. Chugur, and J. Cigarran, "Indexing with WordNet synsets can improve text retrieval," in *Proc. of the 1998 COLING/ACL*, 1998.

[18] A. Hotho, S. Staab, and G. Stumme, "Ontologies improve text document clustering," in *Proc. of the 3rd ICDM*. IEEE, 2003, pp. 541–544.

[19] 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.

[20] 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.

[21] ——, "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.

[22] R. Navigli and S. P. Ponzetto, "Multilingual WSD with just a few lines of code: The BabelNet API," in *Proc. of the 50th ACL*, Jeju Island, Korea, 2012, pp. 67–72.

[23] J. Friedman, T. Hastie, and R. Tibshirani, "Additive Logistic Regression: a Statistical View of Boosting," Stanford University, Stanford University, Tech. Rep., 1998.

[24] T. A. Almeida, J. Almeida, and A. Yamakami, "Spam Filtering: How the Dimensionality Reduction Affects the Accuracy of Naive Bayes Classifiers," *Journal of Internet Services and Applications*, vol. 1, no. 3, pp. 183–200, 2011.

[25] J. Platt, "Fast Training of Support Vector Machines using Sequential Minimal Optimization," in *Advances in Kernel Methods - Support Vector Learning*, B. Schoelkopf, C. Burges, and A. Smola, Eds. MIT Press, 1998. [Online]. Available: http://research.microsoft.com/\~jplatt/smo.html

[26] J. M. Gómez Hidalgo, "Evaluating Cost-Sensitive Unsolicited Bulk Email Categorization," in *Proc. of the 17th ACM SAC*, Madrid, Spain, 2002, pp. 615–620.

[27] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *Proc. of the 13rd ICML*. San Francisco: Morgan Kaufmann, 1996, pp. 148–156.

[28] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[29] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update," *SIGKDD Explorations*, vol. 11, no. 1, 2009.

[30] J. Demsar, "Statistical Comparisons of Classifiers over Multiple Data Sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, Dec. 2006.