

Tiago Pasqualini da Silva

Normalização Textual e Indexação Semântica Aplicadas na Filtragem de SMS Spam

Sorocaba, SP

01 de Julho de 2016

Tiago Pasqualini da Silva

Normalização Textual e Indexação Semântica Aplicadas na Filtragem de SMS Spam

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação (PPGCC-So) da Universidade Federal de São Carlos como parte dos requisitos exigidos para a obtenção do título de Mestre em Ciência da Computação. Área de concentração: Inteligência Artificial e Banco de Dados.

Universidade Federal de São Carlos – UFSCar

Centro de Ciências em Gestão e Tecnologia – CCGT

Programa de Pós-Graduação em Ciência da Computação – PPGCC-So

Orientador: Prof. Dr. Tiago Agostinho de Almeida

Sorocaba, SP

01 de Julho de 2016

Pasqualini da Silva, Tiago

Normalização Textual e Indexação Semântica Aplicadas na Filtragem de
SMS Spam / Tiago Pasqualini da Silva. -- 2016.
63 f. : 30 cm.

Dissertação (mestrado)-Universidade Federal de São Carlos, campus
Sorocaba, Sorocaba

Orientador: Prof. Dr. Tiago Agostinho de Almeida

Banca examinadora: Prof. Dr. Akebo Yamakami, Profa. Dra. Sahudy
Montenegro González

Bibliografia

1. Filtragem de SMS spam. 2. Spam em dispositivos móveis. 3.
Categorização de texto. I. Orientador. II. Universidade Federal de São Carlos.
III. Título.

Ficha catalográfica elaborada pelo Programa de Geração Automática da Secretaria Geral de Informática (SIn).

DADOS FORNECIDOS PELO(A) AUTOR(A)



UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências em Gestão e Tecnologia
Programa de Pós-Graduação em Ciência da Computação

Folha de Aprovação

Assinaturas dos membros da comissão examinadora que avaliou e aprovou a Defesa de Dissertação de Mestrado do candidato Tiago Pasqualini da Silva, realizada em 01/07/2016:

Prof. Dr. Tiago Agostinho de Almeida
UFSCar

Prof. Dr. Akebo Yamakami
UNICAMP

Profa. Dra. Sabudy Montenegro Gonzalez
UFSCar

Aos meus pais João e Rosa.

Agradecimentos

Agradeço,

à UFSCar pela sua estrutura e ótimo ensino;

aos meus amigos por me motivarem a continuar em busca de mais esse título;

aos professores da UFSCar, em especial ao Prof. Tiago pela enorme paciência e pela motivação em terminar essa dissertação;

às empresas Fit e CESAR que trabalhei durante o mestrado, pela compreensão nas horas dedicadas à escrita dessa dissertação;

à todos que contribuíram de forma direta ou indireta com a realização deste trabalho.

“We should always be in pursuit of simplicity, in whatever form it takes.”
(Jeff Atwood)

Resumo

A popularização dos *smartphones* contribuiu para o crescimento do uso de mensagens SMS como forma alternativa de comunicação. O crescente número de usuários, aliado à confiança que eles possuem nos seus dispositivos tornam as mensagens SMS um ambiente propício aos *spammers*. Relatórios recentes indicam que o volume de spam enviados via SMS está aumentando vertiginosamente nos últimos anos. SMS spam representa um problema desafiador para os métodos tradicionais de detecção de spam, uma vez que essas mensagens são curtas e geralmente repletas de gírias, símbolos, abreviações e *emoticons*, que torna até mesmo a *tokenização* uma tarefa difícil. Diante desse cenário, esta dissertação propõe e avalia um método para normalizar e expandir amostras curtas e ruidosas de mensagens SMS de forma a obter atributos mais representativos e, com isso, melhorar o desempenho geral na tarefa de classificação. O método proposto é baseado em dicionários lexicográficos e semânticos e utiliza técnicas modernas de análise semântica e detecção de contexto. Ele é empregado para normalizar os termos que compõem as mensagens e criar novos atributos para alterar e expandir as amostras originais de texto com o objetivo de mitigar fatores que podem degradar o desempenho dos métodos de classificação, tais como redundâncias e inconsistências. A proposta foi avaliada usando uma base de dados real, pública e não codificada, além de vários métodos consagrados de aprendizado de máquina. Os experimentos foram conduzidos para garantir resultados estatisticamente corretos e indicaram que o método proposto pode de fato melhorar a detecção de spam em SMS.

Palavras-chave: Filtragem de SMS spam. Spam em dispositivos móveis. Categorização de texto. Aprendizado de máquina. Processamento de linguagem natural.

Abstract

The rapid popularization of smartphones has contributed to the growth of SMS usage as an alternative way of communication. The increasing number of users, along with the trust they inherently have in their devices, makes SMS messages a propitious environment for spammers. In fact, reports clearly indicate that volume of mobile phone spam is dramatically increasing year by year. SMS spam represents a challenging problem for traditional filtering methods nowadays, since such messages are usually fairly short and normally rife with slangs, idioms, symbols and acronyms that make even tokenization a difficult task. In this scenario, this thesis proposes and then evaluates a method to normalize and expand original short and messy SMS text messages in order to acquire better attributes and enhance the classification performance. The proposed text processing approach is based on lexicography and semantic dictionaries along with the state-of-the-art techniques for semantic analysis and context detection. This technique is used to normalize terms and create new attributes in order to change and expand original text samples aiming to alleviate factors that can degrade the algorithms performance, such as redundancies and inconsistencies. The approach was validated with a public, real and non-encoded dataset along with several established machine learning methods. The experiments were diligently designed to ensure statistically sound results which indicate that the proposed text processing techniques can in fact enhance SMS spam filtering.

Key-words: SMS spam filtering. Mobile phone spam. Text categorization. Machine learning. Natural language processing.

Lista de ilustrações

Figura 1 – Número de usuários (em milhões) dos principais mensageiros instantâneos, em abril de 2015.	30
Figura 2 – Principais tipos de SMS spam enviados nos Estados Unidos.	32
Figura 3 – Exemplo de spam enviado via SMS.	33
Figura 4 – Exemplo de SMS ruidoso, com símbolos, gírias e abreviações.	34
Figura 5 – A amostra original é processada pelos dicionários semânticos e técnicas de detecção de contexto. Cada dicionário/técnica cria uma nova amostra, normalizada ou expandida. Depois disso, dada uma regra de combinação, as amostras são unidas em uma mensagem de texto final com o mesmo conteúdo semântico da amostra original.	44
Figura 6 – Sistema <code>TextExpansion</code> disponível <i>online</i>	48
Figura 7 – Gráficos com os <i>rankings</i> médios, calculados com o Teste de Friedman utilizando os resultados da classificação de cada base de dados obtida com as diferentes regras de combinação. Um ponto mais próximo do centro indica que os <i>rankings</i> médios são menores, o que representa um resultado melhor.	55

Lista de tabelas

Tabela 1 – Exemplo de representação utilizada pelos métodos de aprendizado de máquina.	35
Tabela 2 – Representação vetorial das amostras (1) “ <i>I bought a car</i> ”, (2) “ <i>my car is white</i> ” e (3) “ <i>this wall is white</i> ”, utilizando o modelo <i>bag-of-words</i>	36
Tabela 3 – Representação vetorial das amostras (1) “ <i>joy is amazing</i> ”. (2) “ <i>amy is gr8</i> ”, (3) “ <i>deadpool is great</i> ”, (4) “ <i>this book is great</i> ” e (5) “ <i>book this hotel</i> ”, utilizando o modelo <i>bag-of-words</i>	36
Tabela 4 – Possíveis conjuntos de parâmetros que podem ser utilizados para gerar uma amostra expandida, considerando preservar os termos originais e os estágios de <i>Normalização</i> , <i>Geração de conceitos</i> e <i>Desambiguação</i> , respectivamente. Cada parâmetro é binário, sendo que ‘N’ indica que o estágio não será aplicado e ‘S’ indica que o estágio será aplicado.	46
Tabela 5 – Exemplo de normalização e expansão de uma amostra de SMS. <i>Termos normalizados</i> correspondem à saída do estágio de <i>Normalização</i> . <i>Conceitos</i> mostra todos os conceitos relacionados a cada palavra da amostra, obtidos no estágio de <i>Geração de conceitos</i> . <i>Conceitos selecionados</i> mostra os conceitos mais relevantes, selecionados de acordo com o contexto da amostra, obtidos no estágio de <i>Desambiguação</i> . A amostra <i>Final</i> é obtida através da <i>Regra de combinação R11</i> , que consiste na união das saídas dos estágios de <i>Normalização</i> e <i>Desambiguação</i>	47
Tabela 6 – Lista de métodos de classificação utilizados para avaliar se as bases de dados geradas com o método de expansão proposto obtém desempenhos superiores às bases de dados originais.	50
Tabela 7 – Resultados obtidos por cada método aplicado para classificar as amostras de SMS expandidas pelo método proposto utilizando todas possíveis regras de combinação apresentadas na Tabela 4. O melhor <i>MCC</i> obtido por cada método está destacado em negrito.	51

Tabela 8 – Posições calculadas utilizando o <i>Wilcoxon Signed-Ranks Test</i> . A coluna <i>Exp</i> apresenta os resultados obtidos usando a melhor regra de combinação para cada método de classificação; a coluna <i>Orig</i> mostra os resultados obtidos utilizando a base de dados original; a coluna <i>Dif</i> apresenta as diferenças entre os resultados obtidos com a base de dados <i>Original</i> e <i>Expandida</i> , respectivamente; e a coluna <i>Posição</i> mostra a posição de cada classificador no <i>ranking</i> . As primeiras linhas da tabela contém os resultados em que a base de dados <i>Original</i> obteve um resultado superior, já as linhas após a divisão contém os resultados em que a base de dados <i>Expandida</i> obteve um resultado superior.	53
Tabela 9 – Resultados obtidos usando o teste de Friedman aplicado nos grupos de métodos de classificação. A hipótese nula é rejeitada se F_F é maior do que o valor obtido ao calcular a média dos <i>rankings</i> médios, que nesse caso é 6.	54

Lista de abreviaturas e siglas

B.C4.5	<i>Boosting</i> do método de classificação C4.5
B.NB	<i>Boosting</i> do método de classificação Naïve Bayes
BICTW	<i>Binary Context Tree Weighting</i>
C4.5	Método de classificação baseado em árvore de decisão
DECTW	<i>Decomposed Context Tree Weighting</i>
DMC	<i>Dynamic Markov Compression</i>
<i>k</i> -NN	<i>k-Nearest Neighbors</i>
L.SVM	SVM Linear
LDB	<i>Lexical Database</i>
Logistic	Regressão Logística
LZ78	Algoritmo Lempel-Ziv 78
LZms	Algoritmo Improved Lempel-Ziv
MCC	<i>Matthews Correlation Coefficient</i>
NB	Naïve Bayes
PPM	<i>Prediction by Partial Match</i>
PST	<i>Probabilistic Suffix Trees Compression</i>
SMO	<i>Sequential Minimal Optimization</i>
SMS	<i>Short Message Service</i>
SVM	<i>Support Vector Machines</i>

Lista de símbolos

$\chi^2_{\mathbf{F}}$	Medida chi quadrado
$\mathbf{F}_{\mathbf{F}}$	Medida do Teste de Friedman

Sumário

	Prefácio	25
1	COMUNICAÇÃO MÓVEL	29
1.1	SMS Spam	30
1.2	Características das mensagens	33
2	REPRESENTAÇÃO COMPUTACIONAL	35
2.1	Representação computacional de textos	35
2.2	Limitações da representação de textos através de <i>bag-of-words</i>	36
3	TÉCNICAS DE NORMALIZAÇÃO LÉXICA E INDEXAÇÃO SEMÂNTICA	39
3.1	Normalização léxica	39
3.2	Indexação semântica	40
3.3	Técnicas de processamento de linguagem natural aplicadas na filtragem de SMS spam	42
4	TEXTEXPANSION	43
4.1	Normalização	44
4.2	Geração de conceitos	45
4.3	Desambiguação	45
4.4	Regras de combinação	46
4.5	Exemplo de expansão	46
4.6	Ferramenta desenvolvida	47
5	EXPERIMENTOS E RESULTADOS	49
5.1	Metodologia experimental	49
5.2	Resultados	50
5.3	Análise dos parâmetros	53
6	CONCLUSÃO	57
6.1	Publicações	58
6.1.1	Periódicos	58
6.1.2	Anais de congressos	58
	REFERÊNCIAS	59

Prefácio

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. Essas mensagens são bastante populares em alguns lugares do mundo por serem mais baratas do que as ligações por voz.

Recentemente, 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 ocidentais¹.

Filtros de spam já consolidados para detectar 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 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.

O uso excessivo de gírias, símbolos e abreviações, torna o vocabulário dessas mensagens muito variado e esparso, podendo causar redundâncias na sua representação. Além disso, dois importantes problemas bem conhecidos na área de processamento de linguagem natural também se aplicam a esse tipo de mensagem: a *polissemia* (quando uma palavra possui mais de um significado) e a *sinonímia* (quando várias palavras possuem o mesmo significado) (GABRILOVICH; MARKOVITCH, 2007). Um exemplo de mensagem com esses problemas pode ser observado na frase “*hey man, 5up*”. Nela, pode-se notar a presença de gírias, na palavra “*man*”, e de abreviações, no acrônimo “*5up*”, que significa “*what’s up*”, que também é uma gíria. Além disso, o exemplo apresenta os problemas de polissemia e sinonímia. A polissemia pode ser observada nas palavras “*man*” e “*up*”, que podem ser utilizadas com significados totalmente diferentes do contexto da mensagem; e a sinonímia pode ser observada na palavra “*hey*”, que pode ser substituída pelos sinônimos “*hi*” ou “*hello*”, sem mudar o contexto da informação.

Esse conjunto de problemas, aliado ao tamanho reduzido das amostras, faz com que os métodos de aprendizado de máquina já consolidados na filtragem de spam recebam pouca informação (normalmente de baixa qualidade) e, conseqüentemente, acabam tendo seu desempenho degradado. Na prática, isso pode impactar a eficácia de sistemas que manipulam tais tipos de dados.

¹ Relatório anual da *Cloudmark*. Disponível em <<http://goo.gl/5TFAMM>>. Acessado em: 30/05/2016

Pesquisas recentes na área de processamento de linguagem natural têm focado no combate à sinonímia e polissemia contida nos textos, porém pouca atenção se tem dado aos textos curtos e poluídos (GABRILOVICH; MARKOVITCH, 2005; HIDALGO; RODRÍGUEZ; PÉREZ, 2005; GABRILOVICH; MARKOVITCH, 2007; HERNAULT; BOLLEGALA; ISHIZUKA, 2010; LI et al., 2011). Contudo, apenas encontrar soluções para esses dois problemas pode não ser suficiente, já que as mensagens SMS normalmente são repletas de abreviações, gírias e símbolos. Dessa forma, a correção e normalização desses ruídos é um passo fundamental que deve ser aplicado antes do emprego de técnicas tradicionais de processamento de linguagem natural.

Diante desse cenário, esta dissertação visa procurar alternativas para maximizar o desempenho dos métodos de aprendizado de máquina quando aplicados na filtragem de SMS spam. Para combater os problemas mencionados, foi criado um sistema que faz uso de técnicas de normalização léxica e detecção de contexto, além de dicionários semânticos. A partir desses recursos, o sistema realiza traduções sucessivas das amostras para encontrar termos candidatos que são utilizados para expandir o conjunto original de atributos. Além disso, o modelo implementado é flexível, permitindo que sejam utilizados além dos dicionários propostos, outros dicionários criados pelos próprios usuários, possibilitando a aplicação do sistema em outros domínios.

Objetivos e contribuições

O principal objetivo desta dissertação é oferecer um sistema de normalização léxica e indexação semântica para tratar amostras curtas e ruidosas de texto. Ele pode ser utilizado para melhorar a representação computacional de mensagens, possibilitando assim obter melhores resultados em tarefas de aprendizado de máquina, tais como classificação e agrupamento. O sistema foi desenvolvido com o objetivo de melhorar a detecção de spam em SMS.

Entre as contribuições oferecidas neste trabalho, destacam-se:

- Introdução ao problema do SMS spam, suas características e dificuldades;
- Análise de diferentes técnicas consideradas o estado-da-arte para pré-processamento, normalização léxica e indexação semântica de mensagens de texto;
- Desenvolvimento e disponibilização de um sistema *online* para realizar o tratamento de amostras de texto.

Organização

Esta dissertação está organizada da seguinte maneira:

- No Capítulo 1, é apresentado o problema do spam disseminado por SMS e suas características.
- No Capítulo 2, são detalhadas as técnicas tradicionais de representação computacional de textos, bem como suas limitações.
- No Capítulo 3, são apresentados trabalhos relacionados à normalização textual, indexação semântica e filtragem de SMS spam.
- No Capítulo 4, é detalhado o sistema `TextExpansion`, principal proposta deste trabalho.
- No Capítulo 5, são apresentados os experimentos conduzidos neste trabalho, bem como os resultados obtidos.
- No Capítulo 6, são oferecidas as principais conclusões e direcionamentos para possíveis trabalhos futuros.

1 Comunicação móvel

Nos dias atuais, as redes sociais e a Internet estão presentes na vida de grande parte da população mundial. *Smartphones* estão sendo cada vez mais utilizados, chegando ao ponto de que praticamente todas as pessoas possuem um desses aparelhos.

A popularização desses aparelhos torna possível diversas formas de comunicação antes não existentes. Da mesma forma que os celulares trouxeram a possibilidade de fazer ligações por voz em qualquer lugar, os *smartphones*, devido a sua grande capacidade e facilidade de uso, possibilitaram a comunicação por texto de uma forma muito eficaz.

Segundo artigo publicado pela Forbes¹, a maioria das pessoas, principalmente aquelas com menos de 40 anos, prefere formas de comunicação por texto do que por voz, sejam via mensagens SMS, e-mails, mensageiros instantâneos ou mídias sociais. O artigo também indica que nos últimos anos, ao mesmo tempo que a média de chamadas por voz realizadas por mês tem diminuído, o número de mensagens de texto enviadas tem aumentado.

Segundo pesquisa realizada no final de 2014 nos Estados Unidos², os jovens estão entre os maiores usuários dessas formas de comunicação. A pesquisa aponta que 68% dos entrevistados entre 18 e 29 anos utilizam muito o celular para trocar mensagens de texto e 38% desses jovens utilizam muito mídias sociais através do celular. No caso dos entrevistados entre 30 e 49 anos, esses números caem para 47% e 20%, respectivamente.

A comunicação por texto está entre os principais usos dos *smartphones*, seja via SMS ou via mensageiros instantâneos. Segundo pesquisa realizada na Alemanha³, em 2013, mais de 60 bilhões de mensagens foram enviadas via SMS ou mensageiros instantâneos. A mesma pesquisa mostra que, em abril de 2015, o número de usuários do Whatsapp, o principal mensageiro instantâneo móvel atualmente, chegou a 800 milhões. A Figura 1 ilustra o número de usuários dos principais mensageiros instantâneos disponíveis, em abril de 2015.

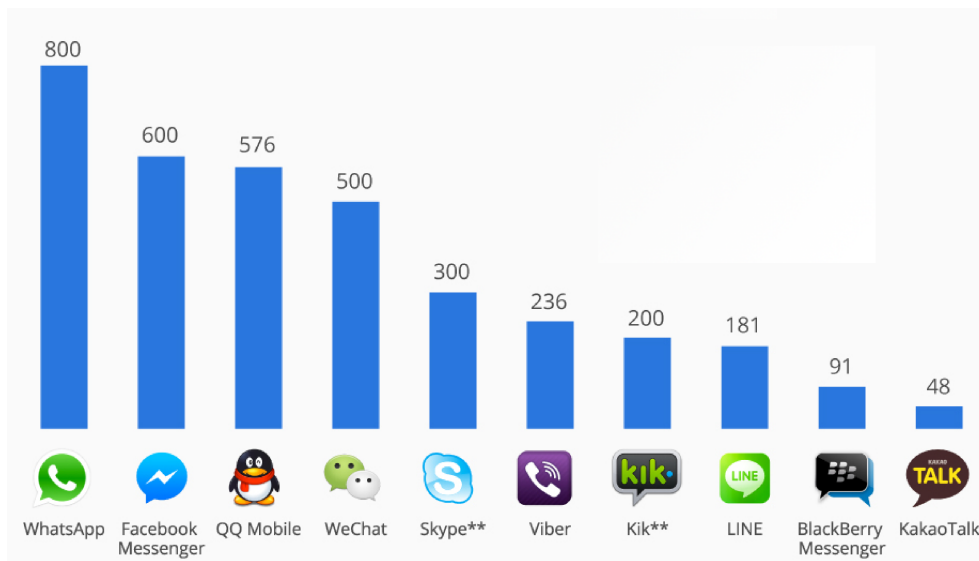
A crescente popularidade de mensagens de texto como nova forma de comunicação também atraiu seu emprego para a divulgação de spam. Consequentemente, nos últimos anos, mensageiros instantâneos e principalmente mensagens SMS vêm sendo utilizados para a disseminar spam.

¹ *Why Millennials Are Texting More And Talking Less*. Disponível em: <<http://goo.gl/1yCdY9>>. Acessado em: 30/05/2016.

² *The New Era of Communication Among Americans*. Disponível em: <<http://www.gallup.com/poll/179288/new-era-communication-americans.aspx>>. Acessado em: 30/05/2016.

³ *Instant messenger – an overview*. Disponível em: <<https://mdk.io/instant-messenger-an-overview/>>. Acessado em: 30/05/2016.

Figura 1 – Número de usuários (em milhões) dos principais mensageiros instantâneos, em abril de 2015.



Fonte: *WhatsApp jetzt mit 800 Millionen Nutzern*. Disponível em: <https://goo.gl/TrmwxO>. Acessado em: 30/05/2016.

1.1 SMS Spam

Spam é o nome dado a qualquer mensagem não desejada, geralmente com conteúdo publicitário, enviada em massa para diversos destinatários. Inicialmente, o e-mail foi o primeiro sistema de comunicação atacado massivamente, pois, uma vez que seu custo de envio é baixo, este ambiente se tornou ótimo para a proliferação de mensagens indesejadas. Com isso, mesmo que a maioria dessas mensagens fossem ignoradas, os *spammers* ainda conseguiam atingir um número considerável de usuários (GOODMAN; HECKERMAN; ROUNTHWAITE, 2005).

Na grande maioria dos casos, as mensagens de spam são incômodas e representam um problema tanto para os usuários quanto para os administradores de sistema, que buscam por formas de filtrar essas mensagens.

Segundo relatório de segurança da Cisco, em 2013, em um único dia chegou a ser enviado mais de 150 bilhões de spam via e-mail. O mesmo relatório aponta que os principais assuntos dessas mensagens são notificações falsas de pagamentos ou depósitos em bancos, compras de produtos online, anexos falsos e maliciosos, além de vários outros⁴.

A popularização da Internet e o surgimento de novos meios eletrônicos de comunicação abriram portas para que o spam se disseminasse também em outros cenários. Atualmente, pode-se encontrar spam em *chats*, redes sociais, mensagens SMS, blogs e diversos outros ambientes eletrônicos (GOODMAN; HECKERMAN; ROUNTHWAITE,

⁴ Cisco 2014 Annual Security Report. Disponível em: http://www.cisco.com/web/offer/gist_ty2_asset/Cisco_2014_ASR.pdf. Acessado em: 30/05/2016.

2005).

Um dos ambientes que passou a ser utilizado pelos *spammers* para a disseminação de mensagens indesejadas é o SMS (do inglês, *short message service*). Serviço de mensagens curtas é um meio de comunicação oferecido pelas operadoras de telefonia que permite a troca de mensagens de texto entre celulares ou telefones fixos. Ele é normalmente utilizado como alternativa às ligações de voz em situações onde a comunicação por voz é difícil ou não desejada. Tal forma de comunicação é muito popular em alguns lugares, uma vez que as mensagens de texto são significativamente mais baratas do que as ligações.

O SMS se tornou em uma indústria enorme, uma vez que as mensagens de texto ainda lideram os lucros das empresas de telefonia móvel, se desconsiderar o segmento de voz. De acordo com relatório da empresa Portio Research⁵, o lucro das mensagens móveis foi de mais de 128 bilhões de dólares em 2011, e em 2016 a previsão é de que seja maior que 153 bilhões de dólares. O mesmo documento mostra que, em 2011, mais de 7,8 trilhões de mensagens SMS foram enviadas ao redor do mundo, enquanto que em 2014 esse número subiu para mais de 9,5 trilhões.

A grande popularidade do SMS levou as empresas de telefonia a reduzirem os custos das mensagens para menos de US\$0.001 em mercados como a China, e até mesmo serem grátis em outros países. Entretanto, o crescimento das mensagens de texto, juntamente com os planos com mensagens ilimitadas, permitem que ataques maliciosos através de mensagens de texto custem pouco ou quase nada. Esse fato, combinado com a confiança que os usuários têm em seus dispositivos móveis, faz com que esse seja um ambiente muito propício para ataques. Como consequência disso, os celulares estão se tornando mais um alvo de mensagens indesejadas, com um número crescente de *spammers* utilizando mensagens de texto para conseguir propagar seus ataques. SMS spam (ou spam via SMS) é qualquer mensagem indesejada enviada para um telefone móvel. Essa prática, que se tornou muito popular em alguns países da Ásia, vem se espalhando pelos países ocidentais⁶.

Um relatório produzido pela empresa Cloudmark⁷ aponta que nos Estados Unidos é muito comum o uso de SMS spam para aplicar golpes nos usuários, sendo que 54% dessas mensagens visam tentar roubar informações de cartões de débito pré-pagos. A Figura 2 mostra um gráfico com os principais tipos de SMS spam enviados nos Estados Unidos, em Janeiro de 2013.

Segundo artigo publicado pelo The Economist⁸, “Usuários chineses recebem mais spam por SMS do que qualquer outro lugar no mundo. Só em 2013, os chineses receberam

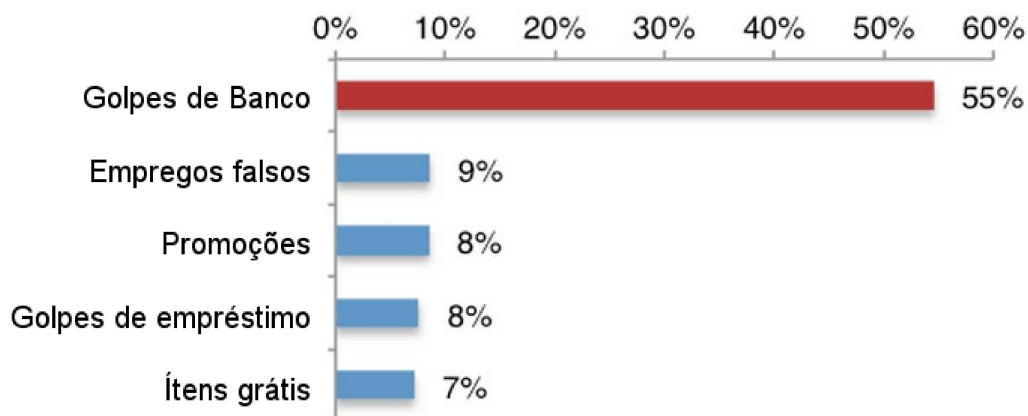
⁵ *Mobile Messaging Futures 2012–2016*. Disponível em: <<http://goo.gl/Wfb01z>>. Acessado em: 30/05/2016.

⁶ Relatório anual Cloudmark. Disponível em: <<http://goo.gl/5TFAMM>>. Acessado em: 30/05/2016.

⁷ *Go Phish, for Bank Cards*. Disponível em: <<https://goo.gl/Efjn8Y>>. Acessado em: 30/05/2016.

⁸ *Spam Messaging – 106 Ways to Annoy*. Disponível em: <<http://goo.gl/cjCRoS>>. Acessado em: 30/05/2016.

Figura 2 – Principais tipos de SMS spam enviados nos Estados Unidos.



Fonte: *Go Phish, for Bank Cards*. Disponível em: <<https://goo.gl/uxmzwj>>. Acessado em: 30/05/2016.

mais de 300 bilhões de spam por SMS, o que significa que na média, cada pessoa utilizando um aparelho celular recebeu um spam por dia. Usuários em grandes cidades como Pequim e Xangai recebem em média duas por dia, ou mais de 700 por ano, o que representa de um quinto a um terço de todas as mensagens de texto”. Para combater esse problema, em julho de 2015, o ministério chinês das indústrias e tecnologia de informação criou novas medidas, chamadas de Regras Administrativas para SMS, impondo regras gerais relacionadas ao marketing via SMS⁹. Além disso, como uma forma de combater a praga das mensagens de spam, o Ministério Vietnamita de Informação e Comunicação está considerando limitar os usuários a enviarem 50 mensagens por dia. A proposta de medida vai limitar um usuário a mandar no máximo 5 mensagens dentro de 5 minutos, 20 dentro de uma hora e não mais do que 50 dentro de 24 horas¹⁰.

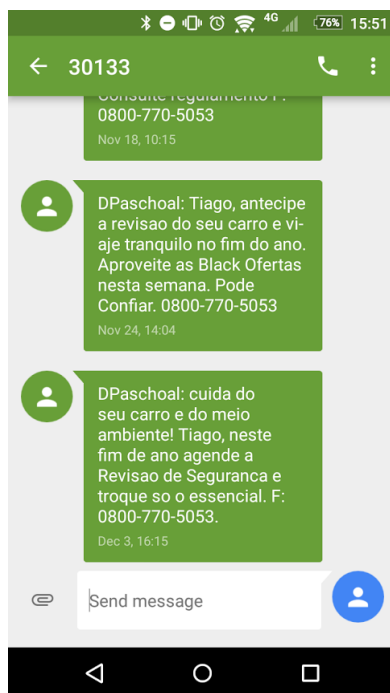
No Brasil, as próprias operadoras chegam a utilizar SMS para divulgar seus produtos e serviços, prática que gera desconforto entre os usuários. Além das operadoras, empresas utilizam o número do celular dos seus clientes como destinatário de suas mensagens de propagandas, como mostra a Figura 3.

Recentemente, o volume de spam também está crescendo em ambientes similares de comunicação por texto. Existem várias indicações de que aplicativos de mensagens instantâneas são o próximo alvo. Por exemplo, existem evidências de malas diretas e notícias falsas sendo circuladas no Whatsapp. A empresa de segurança Panda Labs identificou em 2015, na Espanha, uma das notícias falsas mais populares no Whatsapp, como sendo uma mensagem que promete novos *emojicons* ao clicar e enviar a mesma notícia falsa para dez

⁹ *China Clarifies Requirements for Marketing via SMS*. Disponível em: <<http://goo.gl/ffZceS>>. Acessado em: 30/05/2016.

¹⁰ *Govt to regulate bulk texting*. Disponível em: <<http://goo.gl/LHBP4V>>. Acessado em: 30/05/2016.

Figura 3 – Exemplo de spam enviado via SMS.



amigos¹¹. Existem também mensagens parecidas propagando vírus¹². O Skype também tem sido reportado pelos usuários como alvo frequente de spam¹³. Para evitar esse tipo de problema, o aplicativo Facebook Messenger adicionou uma funcionalidade que permite ao usuário reportar uma mensagem como spam¹⁴.

Além de serem indesejadas, SMS spam também pode ser custar caro, uma vez que alguns usuários podem pagar para receber essas mensagens. Além disso, existe um número muito limitado de softwares que realizam a filtragem de spam em celulares e um ponto preocupante é que mensagens legítimas, como por exemplo mensagens de emergência, podem acabar sendo bloqueadas.

1.2 Características das mensagens

Da mesma forma que as operadoras estão enfrentando problemas com SMS spam, pesquisadores dessa área também estão enfrentando dificuldades. Uma das preocupações é que os filtros de spam consolidados para email têm seu desempenho altamente degradado quando utilizados para filtrar spam em mensagens móveis. Isso acontece devido ao tamanho reduzido das mensagens, que é limitado a 160 caracteres. Além disso, essas mensagens

¹¹ *Las 5 estafas de WhatsApp más famosas de 2015*. Disponível em: <<http://goo.gl/LY9gm7>>. Acessado em: 30/05/2016.

¹² *WhatsApp: Nuevos emoticonos son spam para secuestrar tu agenda de contactos*. Disponível em: <<http://goo.gl/6pNrY1>>. Acessado em: 30/05/2016.

¹³ *Spoofed message from contact*. Disponível em: <<http://goo.gl/fw5wl4>>. Acessado em: 30/05/2016.

¹⁴ *How do I report a message as spam?* Disponível em: <<https://goo.gl/9qjkIr>>. Acessado em: 30/05/2016.

geralmente são repletas de gírias, símbolos, *emoticons* e abreviações, que fazem até mesmo a *tokenização* se tornar uma tarefa difícil. A Figura 4 ilustra uma mensagem SMS repleta de símbolos gírias e abreviações.

Figura 4 – Exemplo de SMS ruidoso, com símbolos, gírias e abreviações.

Hey ru guna b @ da bday party
2day???

Fonte: *Texting Rewrites the Rules of Grammar*. Disponível em:
<<http://goo.gl/zeSXVm>>. Acessado em: 30/05/2016.

O ruído presente nessas mensagens pode aparecer de diversas formas. Um exemplo disso é a frase: “*Plz, call me bak asap... Ive gr8 news! :)*”. Nesse exemplo, pode-se perceber a presença de palavras com a grafia incorreta: “*Plz, bak, Ive, gr8*”, palavras abreviadas: “*asap*”, e de símbolos “*:)*”. Para transcrever essa frase no idioma inglês com a gramática correta, seria necessário um dicionário de *Lingo*¹⁵, além de um dicionário de inglês tradicional, para associar cada gíria, símbolo ou abreviação para o termo correto. Depois de uma etapa de normalização, a frase de entrada seria transcrita para “*Please, call me back as soon as possible... I have great news! :)*”.

Além das mensagens ruidosas, existem outros problemas já conhecidos, como as palavras ambíguas no seu contexto (polissemia) e diferentes palavras com o mesmo significado (sinonímia), que podem prejudicar o desempenho de métodos de aprendizado de máquina tradicionais quando aplicados em problemas de categorização de texto.

Mesmo depois de lidar com problemas de polissemia e sinonímia, a quantidade de atributos pode não ser suficiente para classificar corretamente uma mensagem SMS como spam ou legítima, uma vez que as mensagens são geralmente muito curtas, limitadas em apenas 160 caracteres.

¹⁵ Lingo é uma linguagem abreviada, comumente utilizada em mensagens escritas em celulares e na Internet, como SMS, chats, e-mails, blogs e redes sociais.

2 Representação computacional

Para entender porque as características das mensagens SMS impactam nas tarefas de aprendizado, é necessário primeiramente compreender como os métodos de classificação representam essas mensagens computacionalmente, afinal tais métodos não são desenvolvidos para trabalhar diretamente com textos em linguagem natural.

Em um problema de aprendizado de máquina, cada amostra do conjunto de dados é representado como um vetor de atributos, onde cada coordenada desse vetor representa o valor de um determinado atributo. Supondo uma determinada aplicação onde se deseja classificar objetos de acordo com quatro atributos: as medidas em centímetros da altura, largura e profundidade e sua cor. Nesse caso, o vetor de atributos deverá conter os valores em centímetros da altura, largura e profundidade para cada um dos objetos, além de sua cor. A Tabela 1 apresenta um exemplo composto por um conjunto de 5 amostras.

Tabela 1 – Exemplo de representação utilizada pelos métodos de aprendizado de máquina.

Objeto	Altura	Largura	Profundidade	Cor
Caixa	30	30	20	Azul
Celular	12	7	1	Verde
Monitor	40	25	6	Branco
Notebook	25	15	3	Preto
Mouse	10	6	4	Cinza

No exemplo acima, o vetor de atributos da amostra Caixa, seria (30, 30, 20, *Azul*), o da amostra Celular seria (12, 7, 1, *Verde*), e assim por diante. Através do exemplo, pode-se perceber que esses vetores são compatíveis com atributos quantitativos e qualitativos. Textos, por sua vez, como são dados não-estruturados, não possuem atributos claramente definidos e, portanto demandam pré-processamento para que possam ser representados em vetores de atributos, conforme é apresentado a seguir.

2.1 Representação computacional de textos

Existem diversas formas de representar textos em linguagem natural em vetores de atributos, de forma que os métodos de aprendizado de máquina possam utilizá-los. Entre os modelos disponíveis, o mais simples e amplamente utilizado é o modelo chamado *bag-of-words*, que relaciona a presença de termos em uma amostra com um vocabulário conhecido (FRAKES; BAEZA-YATES, 1992).

No modelo *bag-of-words*, cada amostra de texto da base de dados é dividida em cada uma de suas palavras, de forma a obter cada palavra da amostra separadamente. A

partir dessa divisão, é formado o vocabulário, que basicamente consiste em cada possível palavra existente nessa base de dados. Por exemplo, supondo uma base de dados com as amostras: (1) “*I bought a car*”, (2) “*my car is white*” e (3) “*this wall is white*”, ao dividir cada amostra de texto nas suas palavras, o vocabulário seria composto pelas palavras: *a*, *bought*, *car*, *I*, *is*, *my*, *this*, *wall* e *white*.

Depois de formado o vocabulário de uma base de dados, é possível representar cada amostra de texto através de um vetor que indica a ocorrência ou não de determinada palavra na mensagem. Para criar esse vetor de atributos das amostras é utilizado o valor 1 quando a palavra do vocabulário está presente na amostra e o valor 0 quando a palavra não está presente. A Tabela 2 mostra os vetores de atributos para as amostras do exemplo acima.

Tabela 2 – Representação vetorial das amostras (1) “*I bought a car*”, (2) “*my car is white*” e (3) “*this wall is white*”, utilizando o modelo *bag-of-words*.

	<i>a</i>	<i>bought</i>	<i>car</i>	<i>I</i>	<i>is</i>	<i>my</i>	<i>this</i>	<i>wall</i>	<i>white</i>
(1)	1	1	1	1	0	0	0	0	0
(2)	0	0	1	0	1	1	0	0	1
(3)	0	0	0	0	1	0	1	1	1

2.2 Limitações da representação de textos através de *bag-of-words*

O modelo *bag-of-words*, apesar de conseguir representar os textos em uma forma que os métodos de aprendizado de máquina possam utilizá-los, possui algumas limitações (GABRILOVICH; MARKOVITCH, 2005; GABRILOVICH; MARKOVITCH, 2007).

Para exemplificar, considere as amostras (1) “*joy is amazing*”. (2) “*amy is gr8*”, (3) “*deadpool is great*”, (4) “*this book is great*” e (5) “*book this hotel*”. A Tabela 3 contém seus vetores de atributos.

Tabela 3 – Representação vetorial das amostras (1) “*joy is amazing*”. (2) “*amy is gr8*”, (3) “*deadpool is great*”, (4) “*this book is great*” e (5) “*book this hotel*”, utilizando o modelo *bag-of-words*.

	<i>amazing</i>	<i>amy</i>	<i>book</i>	<i>deadpool</i>	<i>gr8</i>	<i>great</i>	<i>hotel</i>	<i>is</i>	<i>joy</i>	<i>this</i>
(1)	0	1	0	0	1	0	0	1	0	0
(2)	0	0	0	1	0	1	0	1	0	0
(3)	1	0	0	0	0	0	0	1	1	0
(4)	0	0	1	0	0	1	0	1	0	1
(5)	0	0	1	0	0	0	1	0	0	1

Em primeiro lugar, o modelo é muito sensível a variações no vocabulário, como efeito disso, palavras escritas com erros ou abreviadas acabam sendo interpretadas como atributos diferentes pelo modelo, o que pode confundir os métodos de classificação. Isso pode ser percebido nas palavras “*great*” e “*gr8*” no exemplo acima. Apesar de serem a

mesma palavra, porém uma abreviada e outra escrita da forma correta, o modelo acaba criando dois atributos completamente diferentes. Como consequência disso, os métodos de aprendizado de máquina não observam a relação existente entre essas palavras.

Quando as amostras de texto são curtas, o vocabulário acaba se tornando muito grande, o que deixa os vetores de atributos extremamente esparsos. No exemplo apresentado na Tabela 3, apenas cinco amostras resultaram em dez atributos. Em uma base de dados real, com milhares de amostras, o número de atributos pode ser enorme, porém alguns dos atributos acabam sendo utilizados por apenas uma amostra. Mesmo no exemplo reduzido, isso pode ser percebido em diversos atributos: “*amazing*”, “*amy*”, “*deadpool*”, “*gr8*”, “*hotel*” e “*joy*”.

Outra desvantagem é a sua sensibilidade com polissemia e sinonímia. No exemplo oferecido, a sinonímia está presente nas amostras (1), (2) e (3). Elas apresentam comentários positivos sobre filmes, porém, como as palavras utilizadas são diferentes (“*deadpool*”, “*joy*” e “*amy*” para se referir aos nomes dos filmes; “*great*”, “*gr8*” e “*amazing*” como os adjetivos), os métodos de aprendizado de máquina iriam interpretá-las como sendo amostras completamente diferentes.

A polissemia pode ser percebida nas amostras (4) e (5). A palavra “*book*” foi utilizada com a mesma grafia nas duas amostras, porém seu significado é diferente. Nesse caso, como a palavra utilizada é a mesma, ambas as amostras possuem o mesmo atributo. Com isso, perde-se a informação do significado da palavra no contexto da amostra, o que também pode prejudicar o desempenho dos métodos de aprendizado de máquina.

Tanto a sinonímia quanto a polissemia podem ter seus efeitos minimizados utilizando indexação semântica para seleção de conceitos (desambiguação) (NAVIGLI; PONZETTO, 2012b; TAIEB; AOUICHA; HAMADOU, 2013). Tal técnica associa os significados às palavras, procurando termos similares através do contexto das mensagens. De forma geral, a eficácia dessa desambiguação depende da qualidade dos termos extraídos das amostras. Entretanto, ferramentas existentes para processamento de linguagem natural podem não ser adequadas para lidar com mensagens curtas, exigindo outras ferramentas mais apropriadas para esse contexto (BONTCHEVA et al., 2013; MAYNARD; BONTCHEVA; ROUT, 2012; DERCZYNSKI et al., 2013). Trabalhos recentes recomendam a utilização de modelos de ontologia para analisar cada termo de forma a encontrar novos termos relacionados (com o mesmo significado) para enriquecer a amostra original e aumentar o número de atributos (KONTOPOULOS et al., 2013; NASTASE; STRUBE, 2013).

De forma geral, as desvantagens do modelo *bag-of-words*, basicamente, podem ser resumidas à falta de informação semântica nas amostras, o que acaba sendo agravado quando utilizado com amostras curtas de texto, como as presentes nas mensagens SMS. O modelo leva apenas em consideração as palavras individuais presentes na amostra e não consegue utilizar as informações semânticas de cada palavra. Como consequência

disso, bases de dados com um vocabulário muito grande e amostras muito curtas acabam resultando em um desempenho insatisfatório quando utilizado na classificação de mensagens de texto curtas e ruidosas.

3 Técnicas de normalização léxica e indexação semântica

Esta dissertação está diretamente relacionada com três áreas de pesquisa:

1. O uso de técnicas de linguagem natural para normalização léxica de mensagens de *chats* e mídias sociais (HAN; COOK; BALDWIN, 2013);
2. O uso de bases de dados léxicas e dicionários semânticos na representação de textos para classificação (HIDALGO; RODRÍGUEZ; PÉREZ, 2005); e
3. A aplicação em si, ou seja, filtragem de SMS spam com base no conteúdo das mensagens (HIDALGO et al., 2006; ALMEIDA; HIDALGO; YAMAKAMI, 2011).

Cada uma dessas áreas é brevemente apresentada a seguir.

3.1 Normalização léxica

Normalização léxica é a tarefa de substituir variantes léxicas de palavras e expressões normalmente ofuscadas em textos ruidosos pelas suas formas canônicas, para permitir que esses textos possam ser submetidos a tarefas de processamento. Por exemplo, termos como “*gooooood*” e “*b4*” devem ser substituídos pelas palavras em inglês “*good*” e “*before*”, respectivamente.

Normalização léxica está muito relacionada à correção ortográfica e, de fato, muitas abordagens existentes na literatura compartilham técnicas com essa tarefa. Por exemplo, (COOK; STEVENSON, 2009) e (XUE et al., 2011) propõem muitos modelos de erros simples, onde cada um deles captura uma forma em que variantes léxicas são formadas, como variações na ortografia conforme o som (por exemplo, *epik* – “*epic*”) ou abreviações (por exemplo, *goïn* – “*going*”).

Os trabalhos mais parecidos com a proposta apresentada nesta dissertação são os de (AW et al., 2006), (HENRÍQUEZ; HERNÁNDEZ, 2009) e (KAUFMANN; KALITA, 2010), que resolveram o problema da normalização léxica através de traduções, cujo objetivo é traduzir textos ruidosos para o idioma inglês. Tais trabalhos usam modelos de linguagem sofisticados, treinados com amostras de textos ruidosos, enquanto que a abordagem apresentada nesta dissertação utiliza uma tradução relativamente simples de palavras e um modelo de normalização.

3.2 Indexação semântica

Quanto ao uso de bases de dados léxicas (LDBs) para classificação de textos, existe uma longa história de abordagens utilizando a base de dados léxica WordNet (MILLER, 1995) em tarefas como recuperação de informação (GONZALO et al., 1998), categorização de textos (HIDALGO; RODRÍGUEZ; PÉREZ, 2005), agrupamento de textos (HOTHO; STAAB; STUMME, 2003) e mineração de opiniões (WEICHSELBRAUN; GINDL; SCHARL, 2014). Para tarefas supervisionadas, existem duas abordagens principais para usar dicionários de conceitos como o WordNet, ConceptNet (WU; TSAI, 2014) ou BabelNet (HIDALGO; RODRÍGUEZ; PÉREZ, 2005):

- *Indexação semântica*¹: é a troca de palavras em documentos de texto ou em nomes de categorias pelos seus sinônimos, de acordo com os conceitos dessas palavras. Por exemplo, conceitos no WordNet são representados como conjuntos de sinônimos (*synsets*), como {*car, auto, automobile, machine, motorcar*} (*motor vehicle with four wheels*) ou {*car, railcar, railway car, railroad car*} (*vehicle with wheels adapted for rails*) para a palavra “*car*”.
- *Indexação de conceitos*: é a troca (ou adição) de palavras pelos conceitos em si nos documentos de texto. Por exemplo, os dois *synsets* do WordNet apresentados anteriormente possuem os códigos 02961779 e 02963378 como substantivos. Como consequência, qualquer ocorrência da palavra “*car*” pode ser substituída pelo código correspondente ao *synset* apropriado.

Os conceitos em bases de dados léxicas podem expandir sobre sequências de palavras (colocações ou expressões com mais de uma palavra) e podem estar relacionados entre si. No ConceptNet, por exemplo, a expressão “Ford Escort” (na qual cada palavra individual tem seu próprio significado) corresponde ao conceito “ford escort”. Esse conceito é relacionado ao conceito “*car*” pela relação “é um” (*A é um B*). Essas relações tornam as bases de dados léxicas em redes semânticas, sendo que essas relações também podem ser utilizadas para classificação de textos (SCOTT; MATWIN, 1998). É importante ressaltar que alguns autores consideram que trabalhar com expressões de várias palavras é uma abordagem semântica, enquanto que trabalhar com as palavras separadamente é uma abordagem baseada em palavras (CAMBRIA; WHITE, 2014). Na discussão sobre indexação semântica e indexação de conceitos, as abordagens semânticas e baseadas em palavras são instâncias da indexação semântica, a menos que as expressões sejam identificadas por seus códigos de conceitos, o que levaria à indexação de conceitos.

¹ Essa abordagem também é chamada de Busca de Expansão no trabalho de (HIDALGO; RODRÍGUEZ; PÉREZ, 2005). Ela é aplicada a nomes de categorias, mas de forma geral pode ser aplicada a qualquer tipo de texto, especificamente documentos a serem categorizados.

Em ambos os casos, os documentos precisam ser indexados e um processo de treinamento é geralmente aplicado para gerar um classificador, usando algoritmos de aprendizado de máquina como os algoritmos utilizados neste trabalho. Entretanto, o uso de conceitos de bases de dados léxicas adiciona uma certa complexidade para identificar o significado correto (ou os conceitos apropriados) para cada ocorrência de uma palavra. Nesse contexto, esse problema é chamado de desambiguação, ou em inglês, *Word Sense Disambiguation* (WSD).

Existem várias abordagens para realizar desambiguação, uma vez que é trata-se de tarefa comum e muito utilizada em tarefas de processamento de linguagem natural (COLLOBERT; WESTON, 2008). Entre as abordagens, pode-se notar que as duas principais envolvem 1) utilizar aprendizado de máquina em uma coleção de texto manualmente desambiguada, como SemCor (KILGARRIFF; ENGLAND; ROSENZWEIG, 2000), também chamada de desambiguação supervisionada ou, 2) utilizar a informação contida em dicionários (como as palavras e suas definições) ou em bases de dados léxicas (como as relações semânticas do WordNet) de forma a definir distâncias entre os conceitos e utilizá-los para classificar os potenciais conceitos para uma palavra em um determinado contexto (NAVIGLI; LAPATA, 2010; NAVIGLI; PONZETTO, 2012b; TAIEB; AOUICHA; HAMADOU, 2014) – também conhecida como desambiguação não supervisionada.

Neste trabalho foi utilizada a base de dados léxica BabelNet (NAVIGLI; LAPATA, 2010), que é relativamente mais completa, mais recente e menos utilizada do que WordNet em tarefas de classificação de texto. Além disso, foi aplicado um algoritmo de desambiguação não supervisionado, utilizando o método da expansão semântica descrito no trabalho de (HIDALGO; RODRÍGUEZ; PÉREZ, 2005), mas aplicado a documentos de texto em vez de nomes de categorias.

O BabelNet tem como objetivo ser um grande dicionário enciclopédico. Para isso, ele combina os recursos da base de dados léxica WordNet (MILLER, 1995) com os recursos da enciclopédia Wikipédia. No WordNet, os conceitos são representados em conjuntos de sinônimos, chamados de *synsets*. Por exemplo, a palavra “play” no contexto artístico é representada pelo *synset* {*drama, dramatic play, play*}, já a mesma palavra no contexto de brincadeiras infantis é representada pelo *synset* {*child’s play, play*}. No caso da Wikipédia, os autores utilizaram cada artigo da enciclopédia como um conceito e fizeram um mapeamento entre dos conceitos da Wikipédia com os conceitos do WordNet (NAVIGLI; PONZETTO, 2012a).

Para relacionar os conceitos e tornar possível o uso da desambiguação, o BabelNet utiliza um grafo que conecta os conceitos através de uma relação semântica. No caso do WordNet, os autores utilizaram as relações já presentes na base de dados. Já no caso da Wikipédia, os autores utilizaram os *links* entre os artigos como relações. Dessa forma, o BabelNet consegue extrair relações semânticas do tipo “é uma”, “é parte de”, entre outras.

Por exemplo, no caso do conceito “*play*” descrito acima existe uma relação do tipo “*é uma*” com o conceito “*dramatic composition*” (NAVIGLI; PONZETTO, 2012a).

3.3 Técnicas de processamento de linguagem natural aplicadas na filtragem de SMS spam

Com relação à tarefa em si, conhecida como filtragem de SMS Spam, muitas abordagens baseadas na filtragem de spam em e-mail foram aplicadas nessa tarefa. Mesmo assim, a abordagem predominante ainda é a análise do SMS com base no conteúdo, essencialmente replicando filtros de spam bayesianos (ALMEIDA; HIDALGO; YAMAKAMI, 2011; DELANY; BUCKLEY; GREENE, 2012; HIDALGO; ALMEIDA; YAMAKAMI, 2012; TANG; PEI; LUK, 2014). Nesses trabalhos, as mensagens são representadas pelas palavras que as compõem (*bag-of-words*), e métodos de aprendizado de máquina são aplicados nessa representação de forma a obter um classificador automático que é capaz de predizer se uma nova mensagem é spam ou legítima.

(CORMACK; HIDALGO; SANZ, 2007) estudaram o problema de filtragem de spam com base no conteúdo de mensagens curtas de texto que surgiram de três diferentes contextos: SMS, comentários em *blogs* e conteúdos resumidos de e-mails, como aqueles mostrados em clientes de e-mails. Suas principais conclusões foram que mensagens curtas não contém um número suficiente de atributos para suportar adequadamente classificadores de spam baseados em *bag-of-words* ou bigramas, e conseqüentemente, o desempenho dos filtros pode ser melhorado ao se expandir o conjunto de atributos para incluir bigramas esparsos ortogonais (SIEFKES et al., 2004) e também ao se adicionar bigramas ou trigramas de palavras.

Outros autores propuseram formas adicionais de representação de texto. Por exemplo, (LIU; WANG, 2010) apresentaram um método *online* de classificação de textos baseada em índices, que utiliza representação por trigramas. (SOHN et al., 2012) propuseram um método que emprega atributos estilísticos na representação das mensagens, enquanto que (XU et al., 2012) fizeram uso de outros atributos além do conteúdo, como a data, hora e tráfego na rede na mesma abordagem de aprendizado. Contudo, não foi encontrado nenhum trabalho na literatura que tenha empregado informações semânticas ou conceituais na representação de SMS para a tarefa de filtragem de spam.

4 TextExpansion

Conforme já discutido, modelos superficiais de representação de texto como o simples *bag-of-words* têm sido apontados como os principais limitadores do desempenho de algoritmos de aprendizado de máquina em problemas de categorização de texto (GABRILOVICH; MARKOVITCH, 2007). Com o objetivo de melhorar a detecção de spam em SMS, é proposta uma abordagem de pré-processamento de texto composta por técnicas para normalizar, expandir e gerar melhores representações para textos curtos e ruidosos, de forma a produzir atributos mais representativos e melhorar o desempenho da classificação.

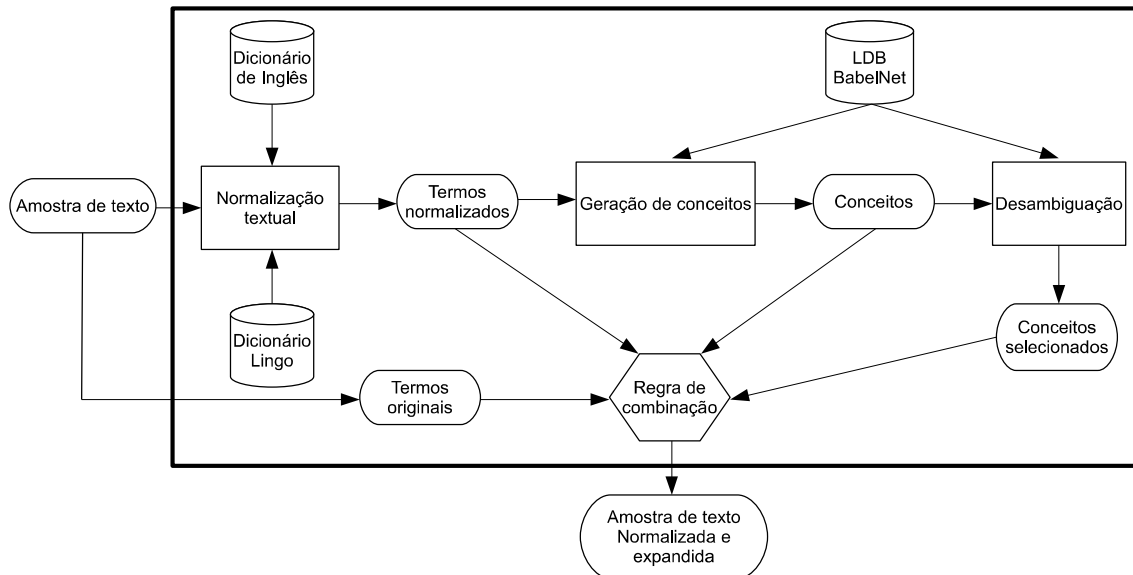
O método de expansão proposto combina técnicas recentes de normalização léxica e detecção de contexto, além de dicionários semânticos. Nessa abordagem, cada amostra de texto é processada em três estágios diferentes, sendo que cada um deles, por sua vez, produz uma nova representação:

1. *Normalização*: utilizada para normalizar e traduzir palavras de Lingo para o idioma inglês.
2. *Geração de conceitos*: utilizada para obter todos os conceitos relacionados a uma palavra, ou seja, cada possível significado de uma determinada palavra.
3. *Desambiguação*: utilizada para encontrar o conceito mais relevante de acordo com o contexto da mensagem, dentre todos os conceitos relacionados a uma determinada palavra.

Os estágios de *Geração de conceitos* e *Desambiguação* são realizados utilizando a base de dados léxica BabelNet, que é o maior repositório semântico atualmente disponível (NAVIGLI; LAPATA, 2010; NAVIGLI; PONZETTO, 2012b). Enquanto que a *Geração de conceitos* consiste em substituir uma determinada palavra por cada um dos conceitos relacionados, a etapa de *Desambiguação* seleciona automaticamente o conceito mais relevante para cada palavra. Isso é feito através de uma análise semântica, realizada pelo algoritmo não supervisionado de desambiguação descrito em (NAVIGLI; PONZETTO, 2012b).

A abordagem de pré-processamento de texto proposta neste trabalho expande uma amostra de texto primeiramente dividindo-a em *tokens*, depois processando cada um dos *tokens* nos estágios descritos, gerando uma nova amostra expandida e normalizada. Dessa forma, dada uma regra de combinação predefinida, as amostras expandidas são então combinadas em uma saída final, que pode ser processada pelos métodos de aprendizado de máquina no lugar da amostra original. A Figura 5 ilustra o processo completo.

Figura 5 – A amostra original é processada pelos dicionários semânticos e técnicas de detecção de contexto. Cada dicionário/técnica cria uma nova amostra, normalizada ou expandida. Depois disso, dada uma regra de combinação, as amostras são unidas em uma mensagem de texto final com o mesmo conteúdo semântico da amostra original.



Nas próximas seções, são apresentados cada um dos estágios da expansão.

4.1 Normalização

Nesse estágio, são utilizados dois dicionários. O primeiro é um dicionário de inglês, que é utilizado para verificar se um determinado termo é uma palavra do idioma inglês, para então normalizá-lo para sua raiz (por exemplo, “*is*” → “*be*” e “*going*” → “*go*”). O segundo dicionário é o de Lingo, que é utilizado para traduzir uma palavra de Lingo para inglês. O processo de tradução inicia procurando cada palavra da amostra no dicionário de inglês. Nesse caso, o método utiliza o dicionário *Freeling*¹. Se a palavra está presente no dicionário, ela é então normalizada para sua raiz. Caso contrário, a palavra é então consultada no dicionário de Lingo, que nesse caso é o dicionário *NoSlang*². Se o dicionário de Lingo não contém uma tradução para a palavra, o termo original é mantido.

¹ Dicionário de inglês *Freeling*. Disponível em: <<http://devel.cpl.upc.edu/freeling/>>. Acessado em 30/05/2016.

² *NoSlang: Internet Slang Dictionary & Translator*. Disponível em: <<http://www.noslang.com/dictionary/full/>>. Acessado em 30/05/2016.

4.2 Geração de conceitos

Os conceitos são obtidos da base de dados léxica BabelNet. Nessa etapa são obtidos todos os conceitos de uma determinada palavra. Como os conceitos são representados no BabelNet como um conjunto de sinônimos, nessa etapa são utilizados todos os sinônimos desses conceitos de forma a se obter todos os possíveis sinônimos de todos os possíveis significados dessa palavra. No caso da palavra “*play*” do exemplo acima, os dois *synsets* da palavra seriam utilizados e, portanto, os conceitos obtidos para a palavra “*play*” seriam: {*drama*, *dramatic play*, *child’s play*}.

Como o BabelNet recebe como entrada palavras no idioma inglês, o método primeiramente aplica a *Normalização* para garantir que cada palavra é de fato do idioma inglês. Depois disso, o método remove as palavras que pertencem a uma lista de *stopwords*, que contém artigos, pronomes, preposições e palavras comuns³. O restante das palavras é então utilizado em um sistema de busca para encontrar seus conceitos.

4.3 Desambiguação

Uma vez que o estágio de *Geração de conceitos* pode obter uma quantidade enorme de conceitos para cada palavra na amostra original, foi implementada a etapa de desambiguação, que utiliza o algoritmo proposto por (NAVIGLI; PONZETTO, 2012b). Basicamente, esse algoritmo procura pelos conceitos mais relevantes, de acordo com o contexto da amostra.

O algoritmo utiliza o grafo do BabelNet que conecta os *synsets* entre si. Para realizar a desambiguação, o algoritmo faz buscas em profundidade no grafo de conceitos para cada *synset* obtido. Depois disso, o algoritmo une esses caminhos obtidos em um único grafo e pontua os *synsets* de acordo com as distâncias nesse grafo gerado, sendo que quanto menor a distância, maior a sua nota. O algoritmo, então, seleciona, para cada palavra, os *synsets* com as maiores notas. Como cada *synset* contém mais de uma palavra, o algoritmo utiliza como conceito selecionado apenas a primeira palavra do *synset* selecionado.

Utilizando o mesmo exemplo anterior, no caso da amostra “*this next play is great*”, o *synset* escolhido para a palavra “*play*” seria {*drama*, *dramatic play*, *play*} e, portanto, o conceito utilizado na etapa de desambiguação para a palavra “*play*” seria “*drama*”.

³ As *stopwords* utilizadas são: {*a*, *an*, *are*, *as*, *at*, *be*, *by*, *for*, *from*, *had*, *has*, *have*, *he*, *how*, *i*, *in*, *is*, *it*, *of*, *on*, *or*, *she*, *that*, *the*, *they*, *this*, *to*, *too*, *was*, *we*, *were*, *what*, *when*, *where*, *who*, *whose*, *will*, *with*, *you*}.

4.4 Regras de combinação

Como é possível utilizar os termos originais e três estágios de expansão para gerar as bases de dados expandidas, existem quatro parâmetros que definem a regra de combinação. Tais parâmetros são basicamente respostas para as seguintes perguntas, respectivamente: 1. *O método deverá manter os termos originais?*, 2. *O método deverá aplicar a normalização léxica?*, 3. *O método deverá gerar conceitos?*, 4. *O método deverá aplicar a desambiguação?*. Como cada parâmetro é binário: ‘N’ caso a resposta seja negativa e ‘S’ caso a resposta seja positiva, isso resulta em onze possíveis combinações de parâmetros para expandir cada amostra. A Tabela 4 apresenta todos os possíveis conjuntos de parâmetros que podem ser utilizados na regra de combinação.

Tabela 4 – Possíveis conjuntos de parâmetros que podem ser utilizados para gerar uma amostra expandida, considerando preservar os termos originais e os estágios de *Normalização*, *Geração de conceitos* e *Desambiguação*, respectivamente. Cada parâmetro é binário, sendo que ‘N’ indica que o estágio não será aplicado e ‘S’ indica que o estágio será aplicado.

Regras	Termos Originais	Termos Normalizados	Conceitos	Conceitos Selecionados
R1	S	N	N	N
R2	S	S	N	N
R3	S	N	S	N
R4	S	N	S	S
R5	S	S	S	N
R6	S	S	S	S
R7	N	S	N	N
R8	N	N	S	N
R9	N	N	S	S
R10	N	S	S	N
R11	N	S	S	S

Algumas combinações de parâmetros, como por exemplo apenas **Conceito selecionados** não fazem sentido como regras de combinação. Isso acontece porque o parâmetro **Conceito selecionados** sempre depende do parâmetro **Conceitos**, portanto algumas possíveis combinações de parâmetros não são regras de combinação.

4.5 Exemplo de expansão

A Tabela 5 apresenta um exemplo de expansão de uma amostra de SMS. Ela ilustra a saída obtida em cada um dos três estágios para a mensagem original “*Plz, call me bak asap... Ive gr8 news! :)*”. Então, supondo que a regra de combinação seja a R11 [*Normalização + Desambiguação*], a amostra obtida no final do processo seria “*please call phone_call me back as soon as possible i have great big news news_program :)*”, que

poderia ser utilizada por algoritmos de aprendizado de máquina e possivelmente melhorar o desempenho de classificação, uma vez que essa amostra evita problemas comuns de representação.

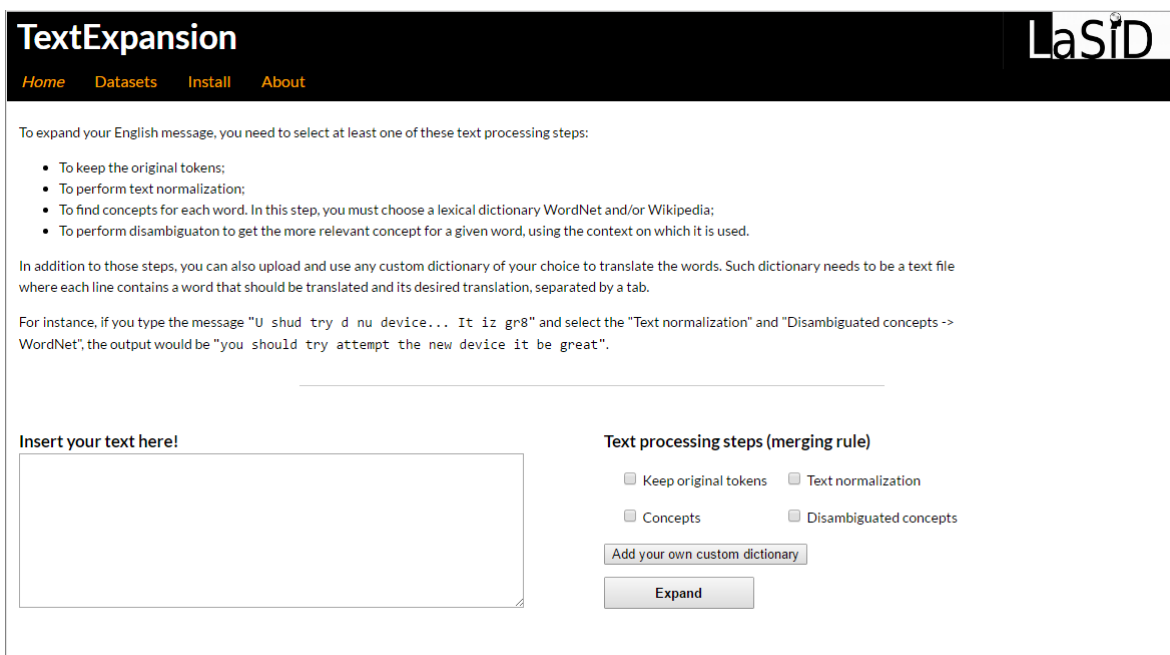
Tabela 5 – Exemplo de normalização e expansão de uma amostra de SMS. *Termos normalizados* correspondem à saída do estágio de *Normalização*. *Conceitos* mostra todos os conceitos relacionados a cada palavra da amostra, obtidos no estágio de *Geração de conceitos*. *Conceitos selecionados* mostra os conceitos mais relevantes, selecionados de acordo com o contexto da amostra, obtidos no estágio de *Desambiguação*. A amostra *Final* é obtida através da *Regra de combinação R11*, que consiste na união das saídas dos estágios de *Normalização* e *Desambiguação*.

Termos originais	<i>Plz, call me bak asap... Ive gr8 news! :)</i>
Termos normalizados	<i>please call me back as soon as possible i have great news :)</i>
Conceitos	<i>please birdsong call call_option caller caller-out claim cry margin_call outcry phone_call shout song telephone_call vociferation yell me backbone backrest binding book_binding cover dorsum rachis rear spinal_column spine vertebral_column as soon as possible i have great news news_program news_show newsworthiness tidings word :)</i>
Conceitos selecionados	<i>please phone_call me as soon as possible i have big news_program :)</i>
Regra de combinação: R11	<i>please, call phone_call me back as soon as possible i have great big news news_program :)</i>

Como mostrado na Tabela 5, o estágio de *Normalização* substitui as gírias e abreviações por suas palavras correspondentes em inglês. Enquanto que a etapa de *Geração de conceitos* obteve todos os conceitos de cada palavra na amostra original, a etapa de *Desambiguação* manteve apenas os conceitos que são semanticamente relevantes à amostra original. Finalmente, utilizando a amostra resultante, espera-se que seja possível evitar problemas conhecidos, como polissemia e sinonímia e, conseqüentemente, espera-se obter melhores resultados ao utilizar técnicas tradicionais de aprendizado de máquina.

4.6 Ferramenta desenvolvida

O sistema implementado está disponível tanto para uso *online* quanto para *download* e instalação. A versão *online* permite que o usuário faça a expansão de bases de dados de texto utilizando os quatro estágios descritos e também permite a utilização de dicionários customizados do usuário. A Figura 6 apresenta o sistema *online*, sendo que tanto o sistema *online* quanto as instruções de *download* e instalação estão disponíveis em: <<http://lasid.sor.ufscar.br/expansion/>>.

Figura 6 – Sistema TextExpansion disponible *online*.

The screenshot shows the TextExpansion web application interface. At the top, there is a black header with the title "TextExpansion" in white and the LaSiD logo on the right. Below the header, there are navigation links: "Home", "Datasets", "Install", and "About".

The main content area contains the following text:

To expand your English message, you need to select at least one of these text processing steps:

- To keep the original tokens;
- To perform text normalization;
- To find concepts for each word. In this step, you must choose a lexical dictionary WordNet and/or Wikipedia;
- To perform disambiguation to get the more relevant concept for a given word, using the context on which it is used.

In addition to those steps, you can also upload and use any custom dictionary of your choice to translate the words. Such dictionary needs to be a text file where each line contains a word that should be translated and its desired translation, separated by a tab.

For instance, if you type the message "U shud try d nu device... It iz gr8" and select the "Text normalization" and "Disambiguated concepts -> WordNet", the output would be "you should try attempt the new device it be great".

Below the text, there is a large text input field labeled "Insert your text here!". To the right of the input field, there is a section titled "Text processing steps (merging rule)" with four checkboxes: "Keep original tokens", "Text normalization", "Concepts", and "Disambiguated concepts". Below these checkboxes, there is a button labeled "Add your own custom dictionary" and a button labeled "Expand".

5 Experimentos e resultados

Neste capítulo, são detalhados os experimentos realizados para validar o sistema proposto. A Seção 5.1 discute a metodologia experimental empregada, bem como as medidas de desempenho e métodos de classificação utilizados. A Seção 5.2 apresenta os resultados obtidos, bem como a análise estatística realizada. Finalmente, a Seção 5.3 oferece a análise das possíveis combinações de parâmetros que podem compor a regra de combinação.

5.1 Metodologia experimental

Para avaliar se o método de expansão proposto pode de fato melhorar o desempenho da tarefa de classificação, foi utilizada a base de dados pública *SMS Spam Collection* (ALMEIDA; HIDALGO; YAMAKAMI, 2011), que é composta por 5.574 mensagens reais de texto em inglês e não codificadas, sendo que 4.827 delas estão categorizadas como legítimas (*ham*), e 747 como spam. É importante destacar que os criadores da base demonstraram que abordagens já consolidadas de classificação de texto têm seu desempenho severamente degradado quando utilizadas para classificar as mensagens originais de SMS, uma vez que elas são pequenas e repletas de gírias, símbolos e abreviações. Essas mesmas características podem ser encontradas em mensagens postadas em redes sociais, fóruns, *chats*, entre outros.

Foram utilizadas todas as onze possíveis regras de combinação descritas na Seção 4.4 para gerar as bases de dados dos experimentos. Também, foi avaliado o desempenho de diversos algoritmos conhecidos de aprendizado de máquina com cada uma das bases de dados geradas, de forma a verificar se a expansão pode melhorar o desempenho na tarefa de classificação. A Tabela 6 apresenta todos os métodos de classificação que foram avaliados. Para dar credibilidade aos resultados obtidos, foram avaliados 17 métodos que utilizam diferentes estratégias de classificação, tais como compressão, distância, árvores e otimização. A maioria dos métodos selecionados é listada como as melhores técnicas de classificação e mineração de dados atualmente disponíveis (WU et al., 2008).

Os métodos avaliados estão disponíveis na biblioteca de aprendizado de máquina WEKA (HALL et al., 2009). Os algoritmos de classificação baseados em compressão foram implementados e estão disponíveis publicamente no pacote `CompressionTextClassifier`¹.

Em todos os experimentos, os métodos de classificação foram utilizados com seus

¹ Os classificadores baseados em compressão também estão disponíveis em: <<http://paginaspersonales.deusto.es/isantos/resources/CompressionTextClassifier-0.4.3.zip>>, compatível com WEKA versão 3.7 ou superiores. Acessado em 30/05/2016.

Tabela 6 – Lista de métodos de classificação utilizados para avaliar se as bases de dados geradas com o método de expansão proposto obtém desempenhos superiores às bases de dados originais.

Técnicas de classificação avaliadas	
Ávores	Bagging of Decision Trees (Bagging) (BREIMAN, 1996) Boosted C4.5 (B.C4.5) (HIDALGO et al., 2006) C4.5 (QUINLAN, 1993)
Compressão	Binary Context Tree Weighting (BICWT) (WILLEMS, 1998) Decomposed Context Tree Weighting (DECTW) (VOLF, 2002) Improved Lempel-Ziv Algorithm (LZms) (NISENSEN et al., 2003) Lempel-Ziv 78 Algorithm (LZ78) (BEGLEITER; EL-YANIV; YONA, 2004) Markov Compression (DMC) (CORMACK; HORSPOOL, 1987) Prediction by Partial Match (PPM) (CLEARY; WITTEN, 1984) Probabilistic Suffix Trees Compression (PST) (RON; SINGER; TISHBY, 1996)
Distância	K -Nearest Neighbors (KNN) (AHA; KIBLER, 1991)
Otimização	Linear SVM (L.SVM) (FORMAN; SCHOLZ; RAJARAM, 2009) Logistic regression (Logistic) (FRIEDMAN; HASTIE; TIBSHIRANI, 1998) Sequential Minimal Optimization (SMO) (PLATT, 1998)
Probabilidade	Boosted Naïve Bayes (B.NB) (FREUND; SCHAPIRE, 1996) Naïve Bayes (NB) (ALMEIDA; ALMEIDA; YAMAKAMI, 2011)
Regras	PART Decision List (PART) (FRANK; WITTEN, 1998)

parâmetros padrões, com exceção do algoritmo dos K -vizinhos mais próximos, que foi avaliado com $K = 1, 3$ e 5 , e para todas as técnicas baseadas em compressão, que foram avaliadas com $C = 0$ e 1 . Isso indica se a adaptação do modelo utilizando teste de instância é aplicado (1) ou não (0) (BEGLEITER; EL-YANIV; YONA, 2004).

Os experimentos foram conduzidos utilizando validação cruzada com 5 -fold e as mensagens foram *tokenizadas* utilizando como delimitadores: pontos, vírgulas, tabulações e espaços. Para comparar os resultados, foi utilizado o Coeficiente de Correlação de Matthews (em inglês, *Matthews Correlation Coefficient* – *MCC*). Ele que é comumente utilizado em aprendizado de máquina como medida da qualidade de classificação binária. Tal coeficiente retorna um valor real entre -1 e $+1$, sendo que um coeficiente igual a $+1$ indica uma predição perfeita, 0 uma predição aleatória, e -1 uma predição inversa (MATTHEWS, 1975). O *MCC* provê uma forma mais balanceada de medir o desempenho de métodos de classificação, se comparado a outras alternativas, especialmente quando as classes são desbalanceadas (ALMEIDA; ALMEIDA; YAMAKAMI, 2011).

5.2 Resultados

A Tabela 7 apresenta os resultados obtidos por cada um dos métodos de classificação avaliados. As amostras originais foram expandidas pelo método proposto utilizando todos os possíveis conjuntos de parâmetros. Os resultados estão ordenados alfabeticamente de acordo com o nome dos métodos e o resultado com maior *MCC* para cada técnica está destacado em negrito.

Tabela 7 – Resultados obtidos por cada método aplicado para classificar as amostras de SMS expandidas pelo método proposto utilizando todas possíveis regras de combinação apresentadas na Tabela 4. O melhor *MCC* obtido por cada método está destacado em negrito.

Métodos		Regras de combinação										
		R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11
Árvores	Bagging	0.833	0.836	0.837	0.825	0.840	0.831	0.832	0.826	0.816	0.837	0.835
	B.C4.5	0.915	0.911	0.917	0.914	0.922	0.912	0.922	0.921	0.910	0.918	0.918
	C4.5	0.802	0.797	0.833	0.825	0.833	0.831	0.816	0.814	0.813	0.838	0.826
Compressão	BICTW C 0	0.014	0.034	-0.027	0.060	-0.023	0.025	-0.043	-0.032	-0.024	-0.011	-0.011
	BICTW C 1	-0.128	-0.122	0.066	-0.185	-0.019	-0.044	-0.083	0.030	0.093	0.010	0.023
	DECTW C 0	0.781	0.797	0.264	0.722	0.274	0.729	0.783	0.331	0.747	0.211	0.749
	DECTW C 1	0.939	0.938	0.749	0.922	0.747	0.916	0.942	0.749	0.923	0.751	0.915
	DMC C 0	0.939	0.938	0.819	0.932	0.816	0.928	0.934	0.818	0.924	0.813	0.932
	DMC C 1	0.797	0.809	0.764	0.846	0.770	0.842	0.805	0.760	0.818	0.771	0.834
	LZ78 C 0	0.876	0.894	0.735	0.880	0.727	0.882	0.893	0.719	0.867	0.728	0.875
	LZ78 C 1	0.876	0.894	0.735	0.880	0.727	0.882	0.893	0.719	0.867	0.728	0.875
	LZms C 0	0.921	0.918	0.764	0.889	0.765	0.903	0.920	0.737	0.902	0.770	0.896
	LZms C 1	0.921	0.918	0.764	0.889	0.765	0.903	0.920	0.737	0.902	0.770	0.896
Distância	PPM C 0	0.929	0.934	0.735	0.916	0.739	0.919	0.935	0.730	0.923	0.739	0.910
	PPM C 1	0.582	0.558	0.395	0.511	0.336	0.480	0.581	0.394	0.486	0.364	0.489
	PST C 0	0.800	0.805	0.546	0.799	0.546	0.805	0.806	0.529	0.810	0.547	0.805
	PST C 1	0.902	0.915	0.713	0.891	0.712	0.893	0.910	0.705	0.906	0.713	0.890
	1-NN	0.771	0.772	0.798	0.778	0.799	0.778	0.772	0.800	0.773	0.798	0.777
Otimização	3-NN	0.572	0.557	0.705	0.580	0.703	0.577	0.574	0.696	0.572	0.707	0.575
	5-NN	0.448	0.454	0.595	0.463	0.592	0.476	0.459	0.589	0.472	0.592	0.475
	L.SVM	0.929	0.925	0.921	0.923	0.922	0.921	0.927	0.917	0.919	0.922	0.927
Probabilidade	Logistic	0.638	0.646	0.702	0.715	0.704	0.704	0.668	0.679	0.679	0.715	0.704
	SMO	0.929	0.925	0.921	0.923	0.922	0.922	0.927	0.917	0.919	0.921	0.927
	B.NB	0.903	0.912	0.832	0.897	0.823	0.902	0.889	0.838	0.893	0.842	0.883
Regras	NB	0.864	0.855	0.744	0.870	0.744	0.863	0.869	0.737	0.867	0.740	0.863
	PART	0.819	0.830	0.846	0.827	0.850	0.851	0.838	0.831	0.839	0.847	0.821

Para cada método de classificação avaliado, foi escolhida a regra de combinação que obteve o melhor resultado de acordo com o *MCC*. Esse resultado foi chamado de *Expansão*. Isso é o equivalente a realizar um ajuste de parâmetros no método de expansão para cada método de classificação avaliado (como a tradicional busca em *grid*). Também, foram selecionados os resultados obtidos com a base de dados original, que foram chamados de *Original*.

Para verificar se as amostras expandidas podem de fato melhorar o desempenho dos métodos de classificação para essa aplicação, é necessário certificar que os resultados obtidos com as bases de dados criadas pelo método proposto são estatisticamente superiores aos resultados obtidos com a base de dados original. Apesar de existirem diversos testes que poderiam ser utilizados para realizar tal análise, o *Wilcoxon Signed-Ranks Test* é conhecido por ser mais robusto que as demais alternativas (DEMSAR, 2006).

O teste de Wilcoxon faz um *ranking* das diferenças de desempenho de ambas as bases de dados para cada técnica de classificação e compara as posições das diferenças positivas e negativas. A Tabela 8 apresenta o *MCC* obtido por cada método de classificação com as bases de dados *Original* e *Expansão*, além das suas diferenças.

Depois disso, é necessário calcular os índices $R+$ e $R-$ que correspondem à soma das posições nas quais as diferenças são negativas e positivas, respectivamente. Nesse caso, $R+ = 21$ e $R- = 330$.

O objetivo é checar se a hipótese nula pode ser rejeitada, o que nesse caso indica que existe uma diferença estatística entre os resultados obtidos com a base de dados expandida e com a base de dados original. Para o *Wilcoxon Signed-Ranks Test*, a hipótese nula pode ser rejeitada com $\alpha = 0.05$, isto é, com um grau de confiança de 95%, quando $z \leq -1.96$. A equação para calcular z é:

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

onde $T = \min(R+, R-)$ e N é a quantidade de métodos de classificação avaliados (os mesmos métodos porém com parâmetros diferentes também devem ser considerados).

Nesse caso, $T = 21$ e $N = 26$, então $z = -5.55$, o que significa que a hipótese nula é rejeitada. Portanto, pode-se concluir que os resultados obtidos pelos métodos de classificação utilizando as amostras expandidas são estatisticamente superiores aos resultados obtidos com as amostras originais. Isso significa que, para tal aplicação, a abordagem de pré-processamento de texto proposta pode de fato prover melhorias no desempenho dos métodos de classificação.

Tabela 8 – Posições calculadas utilizando o *Wilcoxon Signed-Ranks Test*. A coluna *Exp* apresenta os resultados obtidos usando a melhor regra de combinação para cada método de classificação; a coluna *Orig* mostra os resultados obtidos utilizando a base de dados original; a coluna *Dif* apresenta as diferenças entre os resultados obtidos com a base de dados *Original* e *Expandida*, respectivamente; e a coluna *Posição* mostra a posição de cada classificador no *ranking*. As primeiras linhas da tabela contém os resultados em que a base de dados *Original* obteve um resultado superior, já as linhas após a divisão contém os resultados em que a base de dados *Expandida* obteve um resultado superior.

Classificador	MCC		Dif	Posição
	Orig	Exp		
LZms C 0	0.921	0.920	0.001	2
LZms C 1	0.921	0.920	0.001	2
DMC C 0	0.939	0.938	0.001	2
PPM C 1	0.582	0.581	0.001	4
SMO	0.929	0.927	0.002	5.5
L.SVM	0.929	0.927	0.002	5.5
DECTW C 1	0.939	0.942	-0.003	7
PPM C 0	0.929	0.935	-0.006	8.5
NB	0.864	0.870	-0.006	8.5
B.C4.5	0.915	0.922	-0.007	10.5
Bagging	0.833	0.840	-0.007	10.5
B.NB	0.903	0.912	-0.009	12
PST C 0	0.800	0.810	-0.010	13
PST C 1	0.902	0.915	-0.013	14
DECTW C 0	0.781	0.797	-0.016	15
LZ78 C 0	0.876	0.894	-0.018	16.5
LZ78 C 1	0.876	0.894	-0.018	16.5
1-NN	0.771	0.800	-0.029	18
PART	0.819	0.851	-0.032	19
C4.5	0.802	0.838	-0.036	20
BICTW C 0	0.014	0.060	-0.046	21
DMC C 1	0.797	0.846	-0.049	22
Logistic	0.638	0.715	-0.077	23
3-NN	0.572	0.707	-0.135	24
5-NN	0.448	0.595	-0.147	25
BICTW C 1	-0.128	0.093	-0.221	26

5.3 Análise dos parâmetros

Para saber se existe uma escolha de regra de combinação estatisticamente superior às demais para todos os métodos de classificação avaliados, foi realizada outra análise estatística sobre todas as possíveis bases de dados expandidas, sendo que cada uma foi criada utilizando uma regra de combinação diferente. Porém, o teste de Friedman (DEMSAR, 2006) indicou que a hipótese nula não pode ser rejeitada, portanto, não existe diferença estatística entre os resultados encontrados com as diferentes regras de combinação.

Mesmo assim, foi analisado se há alguma escolha de regra de combinação que oferece resultados estatisticamente superiores para grupos específicos de métodos de classificação. Para isso, os algoritmos foram agrupados de acordo com as estratégias de classificação. Os grupos foram definidos como apresentados a seguir.

- *Compressão*: BICTW, DMC, DECTW, LZ78, LZms, PPM e PST;
- *Árvores*: Bagging, B.C4.5 e C4.5;
- *Otimização*: Logistic, L.SVM e SMO;
- *Distância*: 1-NN, 3-NN e 5-NN;
- *Probabilidade*: B.NB e NB.

A Tabela 9 apresenta os resultados obtidos aplicando o teste de Friedman em cada um dos grupos. Como a hipótese nula pode ser rejeitada se $F_F > 6$, então para dois dos cinco grupos analisados existe apenas uma regra de combinação que oferece resultados estatisticamente superiores que qualquer outra: métodos baseados em compressão e métodos baseados em distância.

Tabela 9 – Resultados obtidos usando o teste de Friedman aplicado nos grupos de métodos de classificação. A hipótese nula é rejeitada se F_F é maior do que o valor obtido ao calcular a média dos *rankings* médios, que nesse caso é 6.

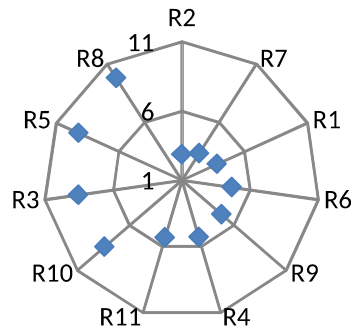
Grupo	χ_F^2	F_F
<i>Compressão</i>	80.61	17.65
<i>Árvores</i>	19.05	3.48
<i>Otimização</i>	12.40	1.40
<i>Distância</i>	27.39	21.02
<i>Probabilidade</i>	15.63	3.58

A Figura 7 apresenta os gráficos dos *rankings* médios, calculados com o Teste de Friedman para cada grupo de métodos de classificação. Na imagem, um ponto mais próximo do centro significa que o *ranking* médio é menor, o que indica que este conjunto obteve resultados superiores aos pontos mais distantes do centro.

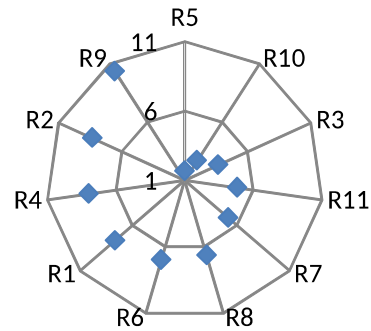
Para o grupo *Compressão*, existem evidências estatísticas de que a melhor regra de combinação é a R2, que consiste em manter os *termos originais* e aqueles obtidos após a *normalização léxica*. Entretanto, para o grupo *Distância*, o melhor desempenho foi obtido aplicando a *normalização léxica* e *geração de conceitos*, representadas pela regra de combinação R3.

Apesar dos resultados obtidos pelos métodos de classificação dos grupos *Árvores*, *Otimização* e *Probabilidade* não terem rejeitado a hipótese nula, os resultados mostram

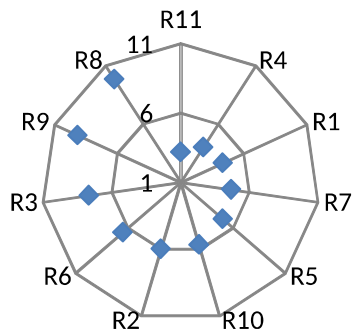
Figura 7 – Gráficos com os *rankings* médios, calculados com o Teste de Friedman utilizando os resultados da classificação de cada base de dados obtida com as diferentes regras de combinação. Um ponto mais próximo do centro indica que os *rankings* médios são menores, o que representa um resultado melhor.



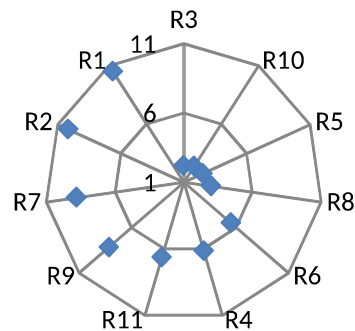
(a) Compressão



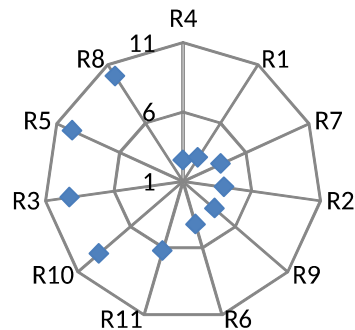
(b) Árvores



(c) Otimização



(d) Distância



(e) Probabilidade

que algumas opções de regras de combinação são, na média, melhores do que as amostras originais. De fato, os algoritmos de classificação baseados em árvores obtiveram resultados melhores com a regra de combinação R5, que emprega os *termos originais* combinados com os obtidos nas etapas de *normalização léxica* e *geração de conceitos*. Já os métodos baseados em probabilidade obtiveram resultados melhores com a regra R4, que resulta nos *termos originais* combinados com aqueles obtidos através da *geração de conceitos* e *desambiguação*. Finalmente, os métodos baseados em otimização obtiveram os melhores desempenhos utilizando a regra R11, que emprega a *normalização léxica* e *desambiguação*.

Para a aplicação proposta, tal análise demonstra que não existe uma única regra de combinação estatisticamente superior para todas as abordagens de classificação avaliadas. Portanto, não é possível selecionar uma regra de combinação que obtenha o melhor desempenho para todos os métodos. Entretanto, uma vez que a regra de combinação ideal é encontrada, utilizar as bases de dados pré-processadas pelo sistema de expansão proposto claramente melhora o desempenho dos métodos de classificação, quando comparado com os resultados obtidos com as amostras originais.

6 Conclusão

A tarefa de filtrar spam em SMS ainda é um desafio nos dias atuais. Dois problemas principais dificultam a aplicação de métodos de classificação tradicionais nesse problema: (1) o pequeno número de atributos que podem ser extraídos de cada mensagem e (2) o fato das mensagens serem repletas de gírias, abreviações e símbolos. Esses problemas tornam extremamente esparsa o vocabulário das bases de dados dessa natureza. Além disso, os problemas de polissemia e sinonímia, já conhecidos na tarefa de classificação de texto, são agravados devido a limitação no tamanho das amostras.

De forma a mitigar esses problemas, foi proposto um sistema de processamento de texto para normalizar, corrigir e expandir amostras curtas, com o objetivo de melhorar o desempenho de técnicas de classificação quando aplicadas para filtrar SMS. O método de expansão é baseado em dicionários lexicográficos e semânticos, além de técnicas recentes de análise semântica e desambiguação. O método foi utilizado para normalizar os termos e criar novos atributos, de forma a modificar e expandir as amostras de texto originais com o objetivo de aliviar os fatores que podem degradar o desempenho, como redundâncias e inconsistências.

A abordagem proposta foi avaliada com uma base de dados pública, real e não codificada, além de diversos algoritmos de classificação conhecidos. Também, foi realizada análise estatística nos resultados obtidos, o que indicou que o método de expansão pode de fato trazer melhorias no desempenho dos métodos de classificação. Portanto, filtros de spam atualmente em uso, podem ter seu desempenho melhorado através do uso do método de processamento proposto.

O plano atual é avaliar o método de expansão em aplicações com características similares às aplicações apresentadas neste estudo, tais como filtragem de comentários e filtragem de spam em redes sociais, sendo que já existe um trabalho que utiliza o método de expansão para detecção de polaridade em redes sociais ([LOCHTER, 2015](#)). Além disso, também planeja-se utilizar o método proposto em problemas diferentes, como agrupamento e recomendação.

Apesar dos bons resultados obtidos, o método de expansão possui limitações. O método é altamente dependente dos dicionários utilizados e, portanto, a qualidade das expansões e o tempo necessário para expandir depende dos dicionários utilizados. Além disso, o método está restrito ao idioma inglês e possui alto custo de manutenção.

Para melhorar o método proposto, planeja-se utilizar técnicas de seleção de termos para reduzir automaticamente a quantidade de conceitos obtidos pela base de dados léxica BabelNet, com o objetivo de atenuar o ruído que, em alguns casos, pode ser criado durante

o estágio de geração de conceitos. Além disso, como trabalho futuro, planeja-se avaliar outros dicionários semânticos e léxicos do idioma inglês, tornar o método capaz de processar textos em outros idiomas e avaliar o uso de comitês, para se aproveitar dos diferentes parâmetros na mesma aplicação, eliminando a necessidade de identificar e selecionar a melhor regra de combinação.

6.1 Publicações

Durante o período de curso de mestrado, os seguintes trabalhos foram publicados:

6.1.1 Periódicos

- ALMEIDA, T.A.; SILVA, T. P. ; SANTOS, I. ; GOMEZ HIDALGO, J. M. . Text Normalization and Semantic Indexing to Enhance Instant Messaging and SMS Spam Filtering. *Knowledge-Based Systems*, 2016.
- ALMEIDA, T.A.; GOMEZ HIDALGO, J. M. ; SILVA, T. P. . Towards SMS Spam Filtering: Results under a New Dataset. *International Journal of Information Security Science*, v. 2, p. 1-18, 2013.

6.1.2 Anais de congressos

- SILVA, T. P. ; SANTOS, I. ; ALMEIDA, T.A. ; GOMEZ HIDALGO, J. M. . Normalização Textual e Indexação Semântica Aplicadas na Filtragem de SMS Spam. *Anais do XI Encontro Nacional de Inteligência Artificial e Computacional (ENIAC'14)*, v. 1. p. 1-6. São Carlos, SP. Brasil. 2014. Premiada com o *The Best Paper Award*.

Referências

- AHA, D.; KIBLER, D. Instance-based learning algorithms. *Machine Learning*, v. 6, p. 37–66, 1991. Citado na página 50.
- ALMEIDA, T. A.; ALMEIDA, J.; YAMAKAMI, A. Spam Filtering: How the Dimensionality Reduction Affects the Accuracy of Naive Bayes Classifiers. *Journal of Internet Services and Applications*, v. 1, n. 3, p. 183–200, 2011. Citado na página 50.
- ALMEIDA, T. A.; HIDALGO, J. M. G.; YAMAKAMI, A. Contributions to the study of SMS spam filtering: new collection and results. In: *Proc. of the 11th ACM DOCENG*. Mountain View, California, USA: [s.n.], 2011. p. 259–262. Citado 3 vezes nas páginas 39, 42 e 49.
- AW, A. et al. A Phrase-based Statistical Model for SMS Text Normalization. In: *Proc. of the 2006 COLING/ACL*. [S.l.]: Association for Computational Linguistics, 2006. p. 33–40. Citado na página 39.
- BEGLEITER, R.; EL-YANIV, R.; YONA, G. On prediction using variable order Markov models. *Journal of Artificial Intelligence Research*, v. 22, p. 385–421, 2004. Citado na página 50.
- BONTCHEVA, K. et al. Twitie: An open-source information extraction pipeline for microblog text. In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing*. Hissar, Bulgaria: [s.n.], 2013. (RANLP’13), p. 83–90. Citado na página 37.
- BREIMAN, L. Bagging predictors. *Machine Learning*, v. 24, n. 2, p. 123–140, 1996. Citado na página 50.
- CAMBRIA, E.; WHITE, B. Jumping nlp curves: a review of natural language processing research [review article]. *Computational Intelligence Magazine, IEEE*, IEEE, v. 9, n. 2, p. 48–57, 2014. Citado na página 40.
- CLEARY, J.; WITTEN, I. Data compression using adaptive coding and partial string matching. *IEEE Transactions on Communications*, IEEE, v. 32, n. 4, p. 396–402, 1984. Citado na página 50.
- COLLOBERT, R.; WESTON, J. L. A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. In: *Proceedings of the 25th International Conference on Machine Learning*. [S.l.: s.n.], 2008. p. 160–167. Citado na página 41.
- COOK, P.; STEVENSON, S. An unsupervised model for text message normalization. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proc. of the 2009 CALC*. [S.l.], 2009. p. 71–78. Citado na página 39.
- CORMACK, G. V.; HIDALGO, J. M. G.; SANZ, E. P. Spam Filtering for Short Messages. In: *Proc. of the 16th ACM CIKM*. Lisbon, Portugal: [s.n.], 2007. p. 313–320. Citado na página 42.

- CORMACK, G. V.; HORSPOOL, R. N. S. Data compression using dynamic Markov modelling. *The Computer Journal*, Br Computer Soc, v. 30, n. 6, p. 541–550, 1987. Citado na página 50.
- DELANY, S. J.; BUCKLEY, M.; GREENE, D. Sms spam filtering: Methods and data. *Expert Systems with Applications*, v. 39, n. 10, p. 9899–9908, 2012. Citado na página 42.
- DEMSAR, J. Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, v. 7, p. 1–30, dez. 2006. ISSN 1532-4435. Citado 2 vezes nas páginas 52 e 53.
- DERCZYNSKI, L. et al. Microblog-genre noise and impact on semantic annotation accuracy. In: *Proceedings of the 24th ACM Conference on Hypertext and Social Media*. Paris, France: [s.n.], 2013. (HT'13), p. 21–30. Citado na página 37.
- FORMAN, G.; SCHOLZ, M.; RAJARAM, S. Feature Shaping for Linear SVM Classifiers. In: *Proc. of the 15th ACM SIGKDD*. Paris, France: [s.n.], 2009. p. 299–308. Citado na página 50.
- FRAKES, W. B.; BAEZA-YATES, R. Information retrieval: data structures and algorithms. Prentice Hall PTR, 1992. Citado na página 35.
- FRANK, E.; WITTEN, I. H. Generating Accurate Rule Sets Without Global Optimization. In: *Proc. of the 15th ICML*. Madison, WI, USA: [s.n.], 1998. p. 144–151. Citado na página 50.
- FREUND, Y.; SCHAPIRE, R. E. Experiments with a new boosting algorithm. In: *Proc. of the 13rd ICML*. San Francisco: Morgan Kaufmann, 1996. p. 148–156. Citado na página 50.
- FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. *Additive Logistic Regression: a Statistical View of Boosting*. Stanford University, 1998. Citado na página 50.
- GABRILOVICH, E.; MARKOVITCH, S. Feature Generation for Text Categorization Using World Knowledge. In: *Proc. of the 19th IJCAI*. Edinburgh, Scotland: [s.n.], 2005. p. 1048–1053. Citado 2 vezes nas páginas 26 e 36.
- GABRILOVICH, E.; MARKOVITCH, S. Harnessing the Expertise of 70,000 Human Editors: Knowledge-Based Feature Generation for Text Categorization. *Journal of Machine Learning Research*, v. 8, p. 2297–2345, 2007. Citado 4 vezes nas páginas 25, 26, 36 e 43.
- GONZALO, J. et al. Indexing with WordNet synsets can improve text retrieval. In: *Proc. of the 1998 COLING/ACL*. [S.l.: s.n.], 1998. Citado na página 40.
- GOODMAN, J.; HECKERMAN, D.; ROUNTHWAITE, R. Stopping spam. *Scientific American*, Nature Publishing Group, v. 292, n. 4, p. 42–49, 2005. Citado 2 vezes nas páginas 30 e 31.
- HALL, M. et al. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, v. 11, n. 1, 2009. Citado na página 49.

- HAN, B.; COOK, P.; BALDWIN, T. Lexical Normalization for Social Media Text. *ACM Trans. Intell. Syst. Technol.*, ACM, New York, NY, USA, v. 4, n. 1, p. 1–27, 2013. Citado na página 39.
- HENRÍQUEZ, C.; HERNÁNDEZ, A. A ngram-based statistical machine translation approach for text normalization on chat-speak style communications. In: *Proc. of the 2009 CAW2*. [S.l.: s.n.], 2009. Citado na página 39.
- HERNAULT, H.; BOLLEGALA, D.; ISHIZUKA, M. A semi-supervised approach to improve classification of infrequent discourse relations using feature vector extension. In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP'10)*. Stroudsburg, PA, USA: [s.n.], 2010. p. 399–409. Citado na página 26.
- HIDALGO, J. M. G.; ALMEIDA, T. A.; YAMAKAMI, A. On the Validity of a New SMS Spam Collection. In: *Proc. of the 2012 IEEE ICMLA*. Boca Raton, FL, USA: [s.n.], 2012. p. 240–245. Citado na página 42.
- HIDALGO, J. M. G. et al. Content Based SMS Spam Filtering. In: *Proc. of the 2006 ACM DOCENG*. Amsterdam, The Netherlands: [s.n.], 2006. p. 107–114. Citado 2 vezes nas páginas 39 e 50.
- HIDALGO, J. M. G.; RODRÍGUEZ, M. B.; PÉREZ, J. C. C. The role of word sense disambiguation in automated text categorization. In: *Proc. of the 10th NLDB*. Alicante, Spain: [s.n.], 2005. p. 298–309. Citado 4 vezes nas páginas 26, 39, 40 e 41.
- HOTH, A.; STAAB, S.; STUMME, G. Ontologies improve text document clustering. In: *IEEE. Proc. of the 3rd ICDM*. [S.l.], 2003. p. 541–544. Citado na página 40.
- KAUFMANN, J.; KALITA, J. Syntactic Normalization of Twitter Messages. In: *Proc. of the 2010 ICON*. [S.l.: s.n.], 2010. Citado na página 39.
- KILGARRIFF, A.; ENGLAND, B.; ROSENZWEIG, J. English Senseval: Report and Results. In: *Proc. of the 2nd LREC*. [S.l.: s.n.], 2000. p. 1239–1244. Citado na página 41.
- KONTOPOULOS, E. et al. Ontology-based sentiment analysis of Twitter posts. *Expert Systems with Applications*, v. 40, n. 10, p. 4065–4074, 2013. Citado na página 37.
- LI, Y. et al. A framework for semisupervised feature generation and its applications in biomedical literature mining. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, v. 8, n. 2, p. 294–307, 2011. Citado na página 26.
- LIU, W.; WANG, T. Index-based Online Text Classification for SMS Spam Filtering. *Journal of Computers*, v. 5, n. 6, p. 844–851, 2010. Citado na página 42.
- LOCHTER, J. V. *Máquinas de Classificação para Detectar Polaridade de Mensagens de Texto em Redes Sociais*. Dissertação (Mestrado) — Universidade Federal de São Carlos, 2015. Citado na página 57.
- MATTHEWS, B. W. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, Elsevier, v. 405, n. 2, p. 442–451, 1975. Citado na página 50.

- MAYNARD, D.; BONTCHEVA, K.; ROUT, D. Challenges in developing opinion mining tools for social media. In: *Proceedings of @NLP can u tag #usergeneratedcontent?!* Istanbul, Turkey: [s.n.], 2012. (LREC'12). Citado na página 37.
- MILLER, G. A. Wordnet: a lexical database for English. *Communications of the ACM*, ACM, v. 38, n. 11, p. 39–41, 1995. Citado 2 vezes nas páginas 40 e 41.
- NASTASE, V.; STRUBE, M. Transforming Wikipedia into a large scale multilingual concept network. *Artificial Intelligence*, v. 194, n. 1, p. 62–85, 2013. Citado na página 37.
- NAVIGLI, R.; LAPATA, M. An experimental study of graph connectivity for unsupervised word sense disambiguation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, IEEE, v. 32, n. 4, p. 678–692, 2010. Citado 2 vezes nas páginas 41 e 43.
- NAVIGLI, R.; PONZETTO, S. P. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, v. 193, p. 217–250, 2012. Citado 2 vezes nas páginas 41 e 42.
- NAVIGLI, R.; PONZETTO, S. P. Multilingual WSD with just a few lines of code: the BabelNet API. In: *Proceedings of the Association for Computational Linguistics 2012 System Demonstrations*. Jeju Island, South Korea: [s.n.], 2012. (ACL '12), p. 67–72. Citado 4 vezes nas páginas 37, 41, 43 e 45.
- NISENSEN, M. et al. Towards behaviorometric security systems: Learning to identify a typist. In: *Proc. of the PKDD'03*. [S.l.]: Springer, 2003. p. 363–374. Citado na página 50.
- PLATT, J. Fast Training of Support Vector Machines using Sequential Minimal Optimization. In: SCHOELKOPF, B.; BURGESS, C.; SMOLA, A. (Ed.). *Advances in Kernel Methods - Support Vector Learning*. [S.l.]: MIT Press, 1998. Citado na página 50.
- QUINLAN, J. R. *C4.5: Programs for Machine Learning*. [S.l.]: Morgan Kaufmann Publishers Inc., 1993. Citado na página 50.
- RON, D.; SINGER, Y.; TISHBY, N. The power of amnesia: Learning probabilistic automata with variable memory length. *Machine Learning*, Springer, v. 25, n. 2, p. 117–149, 1996. Citado na página 50.
- SCOTT, S.; MATWIN, S. Text classification using wordnet hypernyms. In: *Use of WordNet in natural language processing systems: Proceedings of the conference*. [S.l.: s.n.], 1998. p. 38–44. Citado na página 40.
- SIEFKES, C. et al. Combining Winnow and Orthogonal Sparse Bigrams for Incremental Spam Filtering. In: *Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases*. New York, NY, USA: [s.n.], 2004. (PKDD '04), p. 410–421. Citado na página 42.
- SOHN, D.-N. et al. Content-based Mobile Spam Classification Using Stylistically Motivated Features. *Pattern Recogn. Lett.*, Elsevier Science Inc., New York, NY, USA, v. 33, n. 3, p. 364–369, 2012. ISSN 0167-8655. Citado na página 42.
- TAIEB, M. A. H.; AOUICHA, M. B.; HAMADOU, A. B. Computing semantic relatedness using wikipedia features. *Knowledge-Based Systems*, v. 50, n. 9, p. 260–278, 2013. Citado na página 37.

- TAIEB, M. A. H.; AOUICHA, M. B.; HAMADOU, A. B. A new semantic relatedness measurement using wordnet features. *Knowledge and Information Systems*, v. 41, n. 1, p. 467–497, 2014. Citado na página 41.
- TANG, G.; PEI, J.; LUK, W.-S. Email mining: tasks, common techniques, and tools. *Knowledge and Information Systems*, v. 41, n. 1, p. 1–31, 2014. Citado na página 42.
- VOLF, P. A. J. *Weighting techniques in data compression: Theory and algorithms*. Tese (Doutorado) — Technische Universiteit Eindhoven, 2002. Citado na página 50.
- WEICHSELBRAUN, A.; GINDL, S.; SCHARL, A. Enriching semantic knowledge bases for opinion mining in big data applications. *Knowledge-Based Systems*, v. 69, p. 78–85, 2014. ISSN 0950-7051. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0950705114001695>>. Citado na página 40.
- WILLEMS, F. The context-tree weighting method: Extensions. *IEEE Transactions on Information Theory*, IEEE, v. 44, n. 2, p. 792–798, 1998. Citado na página 50.
- WU, C.-E.; TSAI, R. T.-H. Using relation selection to improve value propagation in a conceptnet-based sentiment dictionary. *Knowledge-Based Systems*, v. 69, p. 100–107, 2014. ISSN 0950-7051. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0950705114001737>>. Citado na página 40.
- WU, X. et al. Top 10 algorithms in data mining. *Knowledge and Information Systems*, v. 14, n. 1, p. 1–37, 2008. Citado na página 49.
- XU, Q. et al. SMS Spam Detection Using Noncontent Features. *IEEE Intelligent Systems*, v. 27, n. 6, p. 44–51, 2012. Citado na página 42.
- XUE, Z. et al. Normalizing Microtext. In: ASSOCIATION FOR THE ADVANCEMENT OF ARTIFICIAL INTELLIGENCE. *Proc. of the 2011 AAI*. [S.l.], 2011. p. 74–79. Citado na página 39.