Johannes Von Lochter

Máquinas de Classificação para Detectar Polaridade de Mensagens de Texto em Redes Sociais

Sorocaba, SP

18 de Novembro de 2015

Johannes Von Lochter

Máquinas de Classificação para Detectar Polaridade de Mensagens de Texto em Redes Sociais

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação (PPGCCS) 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 – PPGCCS

Orientador: Prof. Dr. Tiago Agostinho de Almeida

Sorocaba, SP

18 de Novembro de 2015

Johannes Von Lochter

Máquinas de Classificação para Detectar Polaridade de Mensagens de Texto em Redes Sociais/ Johannes Von Lochter. – Sorocaba, SP, 18 de Novembro de 2015-

83 p. : il. (algumas color.) ; 30 cm.

Orientador: Prof. Dr. Tiago Agostinho de Almeida

Dissertação (Mestrado) – Universidade Federal de São Carlos – UFSCar
Centro de Ciências em Gestão e Tecnologia – CCGT
Programa de Pós-Graduação em Ciência da Computação – PPGCCS, 18 de Novembro de 2015.

1. Palavra-chave1. 2. Palavra-chave2. I. Orientador. II. Universidade xxx. III. Centro de xxx. IV. Título

CDU 02:141:005.7

Johannes Von Lochter

Máquinas de Classificação para Detectar Polaridade de Mensagens de Texto em Redes Sociais

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação (PPGCCS) 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.

Trabalho aprovado. Sorocaba, SP, 18 de Novembro de 2015:

Prof. Dr. Tiago Agostinho de Almeida
Orientador

Prof. Dr. Akebo Yamakami
Membro Externo

Prof[a]. Dr[a]. Sahudy Montenegro González
Membro Interno

Sorocaba, SP
18 de Novembro de 2015

Aos meus pais João e Silvia,
e minha irmã Rebekka.

Agradecimentos

Agradeço,

a Deus pela vida e pela oportunidade de realizar um sonho;

aos meus pais João e Silvia por me apoiarem na decisão de continuar a estudar;

a minha irmã Rebekka por ter tido paciência quando eu não pude estar presente;

a amigos especiais, Edeli, Heitor e Nerian, por me motivarem nas dificuldades;

aos companheiros de laboratório pelas dúvidas e críticas, e pela paciência quando eu quis dar palpite no trabalho deles;

e, por último, mas não menos importante, ao Prof. Tiago pela oportunidade, paciência e ensinamentos, por ter sido orientador, professor, e psicólogo nas horas livres. Graças a ele, até recado de geladeira eu reescrevo dez vezes para ter certeza de que ficou claro.

"It's the questions we can't answer that teach us the most. They teach us how to think. If you give a man an answer, all he gains is a little fact. But give him a question and he'll look for his own answers."

(Patrick Rothfuss)

Resumo

A popularidade das redes sociais tem atraído a atenção das empresas. O crescimento do número de usuários e das mensagens enviadas por dia transforma esse ambiente em uma rica fonte de informações para descoberta de necessidades, tendências, opiniões e outras informações que podem auxiliar departamentos de vendas e marketing. Contudo, a maioria das redes sociais impõe limite no tamanho das mensagens, o que leva os usuários a usarem abreviações e gírias para compactarem o texto. Trabalhos na literatura demonstraram avanço na minimização do impacto de mensagens ruidosas nas tarefas de categorização textual através da utilização de dicionários semânticos e modelos ontológicos. Com a aplicação destes, as amostras são normalizadas e expandidas antes de serem apresentadas aos métodos preditivos. Assim, nesta dissertação é proposto um comitê de máquinas de classificação utilizando técnicas de processamento de linguagem natural para detectar opiniões automaticamente em mensagens curtas de texto em inglês. Os resultados apresentados foram validados estatisticamente e indicaram que o sistema proposto obteve capacidade preditiva superior aos métodos preditivos isolados.

Palavras-chaves: Detecção de opinião; Classificação; Processamento de linguagem natural; Aprendizado de máquina.

Abstract

The popularity of social networks have attracted attention of companies. The growing amount of connected users and messages posted per day make these environments fruitful to detect needs, tendencies, opinions, and other interesting information that can feed marketing and sales departments. However, the most social networks impose size limit to messages, which lead users to compact them by using abbreviations, slangs, and symbols. Recent works in literature have reported advances in minimizing the impact created by noisy messages in text categorization tasks by means of semantic dictionaries and ontology models. They are used to normalize and expand short and messy text messages before using them with a machine learning approach. In this way, we have proposed an ensemble of machine learning methods and natural language processing techniques to find the best way to combine text processing approaches with classification methods to automatically detect opinion in short english text messages. Our experiments were diligently designed to ensure statistically sound results, which indicate that the proposed system has achieved a performance higher than the individual established classifiers.

Key-words: opinion detection; classification; natural language processing; machine learning.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

Acc	Acurácia
B. C4.5	*Boosting* de classificador baseado em árvore de decisão
C4.5	Classificador baseado em árvore de decisão
F_1	F-medida
FA	Florestas aleatórias
k-NN	*k-Nearest Neighbors* (k-Vizinhos Próximos)
NB-B	*Naïve* Bayes Bernoulli
NB-G	*Naïve* Bayes Gauss
NB-M	*Naïve* Bayes Multinomial
NLP	*Natural Language Processing* (Processamento de linguagem natural)
RAM	*Random Access Memory* (Memória de Acesso Aleatório)
RL	Regressão logística
SVM	*Support-Vector Machines* (Máquinas de vetores de suporte)
SVM-L	*Support-Vector Machines* linear
SVM-R	*Support-Vector Machines* com base radial

Lista de símbolos

Σ	Concentrador do comitê ou Somatório
$\hat{y}$	Classe ou rótulo
$\vec{x}$	Vetor de atributos que representa uma amostra
χ^2	Distribuição chi-quadrado

Sumário

Prefácio

A difusão das redes sociais e a inclusão digital possibilitaram que as pessoas manifestassem suas opiniões através da *Web*. Tal fenômeno impactou o modo como as empresas operam o atendimento ao cliente e a maneira como os consumidores expressam suas opiniões, abrindo novas oportunidades. Segundo publicação da ComScore[1], as análises disponíveis *online* têm impacto significativo na decisão final de compra dos usuários que as leem. Consequentemente, as empresas notaram a importância de identificar quais usuários são formadores de opinião e quão importante é a tarefa de analisar grande volume de mensagens de forma rápida para descobrir tendências.

Trabalhos da literatura indicam ser promissora a aplicação de métodos de aprendizado de máquina para analisar opiniões em grande quantidade, predizer polaridade de mensagens não vistas e inferir sobre a opinião da maioria (DENECKE, 2008; PANG; LEE; VAITHYANATHAN, 2002). No entanto, as opiniões compartilhadas em redes sociais costumam ter características peculiares.

O tempo gasto nas redes sociais é, geralmente, compartilhado com tarefas relevantes, como trabalho e estudo. Por causa do tempo curto e menos prioritário, as mensagens costumam ser resumidas e repletas de palavras grafadas incorretamente, além de terem em sua composição muitos símbolos, gírias e abreviações, usualmente empregados para compactar a mensagem original e transmitir a informação rapidamente. Consequentemente, analisar o sentimento expresso nessas amostras não é uma tarefa trivial, pois podem haver diversas peculiaridades a respeito da estrutura textual.

Além da gramática inadequada, há problemas bem definidos na literatura relacionados à extração de polaridade de mensagens de texto, tais como identificar frases sarcásticas ou irônicas, discernir significados de palavras ambíguas em um determinado contexto (polissemia), mesclar diferentes palavras com o mesmo significado (sinonímia), encontrar o melhor substituto para cada termo digitado incorretamente e descobrir significados para símbolos como, por exemplo, *smilefaces* (PANG; LEE, 2008; MOSTAFA, 2013). Há um consenso de que tratar estes casos pode aumentar significativamente a eficiência e a acurácia dos métodos preditivos (MOSQUERA; MOREDA, 2013).

No tratamento de problemas como sinonímia e polissemia, destaca-se o emprego de dicionários semânticos para realizar a desambiguação de palavras (NAVIGLI; PONZETTO, 2012; TAIEB; AOUICHA; HAMADOU, 2013). Tais dicionários traçam a relação de significado entre os termos, permitindo encontrar termos com significados próximos ou

[1] *Online Consumer-Generated Reviews Have Significant Impact on Offline Purchase Behavior*. Disponível em <http://goo.gl/PRIHmS>. Acessado em 22/10/2015.

polissêmicos. No entanto, lidar com sinonímia e polissemia nem sempre é suficiente para obter resultados satisfatórios. Muitas vezes os termos resultantes não são suficientes para discernir a polaridade da mensagem, sendo necessário utilizar ontologias para explorar novos termos que podem ser inseridos na amostra original. (KONTOPOULOS et al., 2013; NASTASE; STRUBE, 2013).

Nesse cenário, no qual há diversas técnicas envolvidas e cada uma endereça uma série de dificuldades e as trata com eficiência, a utilização de comitê de máquinas é uma opção viável para aproveitar o melhor de cada técnica, sejam elas de processamento de linguagem natural ou métodos de classificação (DIETTERICH, 2000).

É de suma importância para o cenário atual que sejam exploradas novas técnicas para tratar eficientemente a análise de sentimento voltada para mensagens curtas. Através das facilidades proporcionadas pela Internet, as pessoas têm compartilhado um número cada vez maior de opiniões nas redes sociais e, consequentemente, têm influenciado cada vez mais pessoas dentro do círculo social onde tais opiniões são publicadas.

Objetivos e contribuições

O principal objetivo desta dissertação é oferecer um sistema de comitê de classificadores e técnicas de processamento de linguagem natural para detectar o sentimento expresso em mensagens curtas e ofuscadas extraídas de redes sociais, através da melhoria na qualidade dos atributos das amostras.

Dentre as contribuições oferecidas neste trabalho, destacam-se:

1. Introdução a análise de sentimento no contexto de mensagens curtas e ruidosas;
2. Estudo relacionado à expansão de termos com dicionários semânticos e associação de sentimento com dicionários léxicos;
3. Análise de diferentes abordagens de comitê de máquinas de classificação; e
4. Criação de novas bases de dados rotuladas com mensagens curtas para futuras pesquisas.

Organização

Este manuscrito apresenta a seguinte estrutura:

- No Capítulo 1, é apresentada a área de análise de sentimentos e os principais trabalhos encontrados na literatura.

- No Capítulo 2, são introduzidos conceitos de processamento de linguagem natural, tais como normalização e indexação semântica.
- No Capítulo 3, são oferecidos os métodos de classificação envolvidos neste trabalho.
- No Capítulo 4, é abordado o sistema *online* Sentminer resultante das pesquisas realizadas ao longo deste trabalho.
- No Capítulo 5, são discutidos a realização dos experimentos e os resultados obtidos nesta pesquisa.
- Finalmente, no Capítulo 6, são oferecidas as conclusões e direcionamentos para trabalhos futuros.

1 Análise de sentimento

Obter opiniões acerca de um assunto de interesse, e sumarizá-las, era um desafio nos dias que antecederam o crescimento da Internet. Tal desafio se apresentava para consumidores que queriam obter opiniões a respeito de um produto a ser comprado; para empresas que desejavam medir a aceitação de seus produtos no mercado; para partidos políticos que queriam antecipar a aceitação de seus candidatos pelo eleitorado; para os acionistas que queriam entender melhor a especulação acerca de certos papéis da bolsa de valores; entre muitos outros exemplos. Este desafio levou aos estudos do que hoje é conhecido como detecção de opinião.

A detecção de opinião, atualmente mais conhecida como análise de sentimento ou mineração de opinião, procura sumarizar a opinião de um grupo de pessoas sobre um determinado assunto. Esta é uma tarefa difícil até mesmo para o julgamento humano, variando de pessoa para pessoa o que se entende por ser positivo ou negativo, e a intensidade expressa em uma determinada opinião. Estudos nesta área derivam do processamento de linguagem natural e aprendizado de máquina, os quais podem ser generalizados como áreas da inteligência artificial.

Segundo o dicionário Michaellis, a palavra opinião pode ser descrita como “maneira de opinar; modo de ver pessoal; parecer, voto emitido ou manifestado sobre certo assunto; asserção sem fundamento; presunção; conceito; reputação; juízo ou sentimento que se manifesta em assunto sujeito a deliberação”. A palavra sentimento, que inclusive aparece como uma das definições para opinião, é descrita como “sensação psíquica, tal como as paixões, o pesar, a mágoa, o desgosto; atitude mental a respeito de alguém ou de alguma coisa; opinião; faculdade intuitiva de perceber ou apreciar as qualidades ou méritos de uma coisa”, entre outros que não convém citar. Assim, pode-se observar que ambas as palavras estão relacionadas no que cerne o sentimento ou opinião a respeito de um objeto de interesse.

Antigamente, o desafio imposto por esta tarefa era obter opiniões suficientes que representassem as comunidades interessadas. Hoje, o desafio é processar o número expressivo de opiniões facilmente encontradas em redes sociais, e sumarizá-las.

Para compreender melhor o desafio que a análise de sentimento impõe hoje, é necessário entender o surgimento da detecção de opinião e a evolução da Internet como catalisador para que as redes sociais se tornassem atrativas para a disseminação de opiniões. Desta forma, este capítulo apresenta a seguinte estrutura: a Seção 1.1 revisita o surgimento da detecção de opinião. A Seção 1.2 apresenta, resumidamente, o impacto da evolução da Internet no crescimento das redes sociais. A necessidade de analisar opiniões

em redes sociais é abordada na Seção 1.3. Na Seção 1.4, é discutida a análise de sentimento em mensagens postadas nas redes sociais.

1.1 As opiniões antes da Internet

O primeiro exemplo de pesquisa de opinião, que se tem conhecimento, foi conduzido pelo jornal *The Harrisburg Pennsylvanian* em 1824[1], popularizando a pesquisa de opinião depois de predizer corretamente que o candidato à presidência, Andrew Jackson, venceria as eleições.

Entre 1916 e 1932, a revista semanal *Literary Digest* usou a pesquisa de opinião para predizer corretamente os candidatos eleitos na corrida à presidência. Em 1936, um erro de amostragem levou esta mesma revista a veicular que Alfred Landon seria eleito, mas Roosevelt ganhou as eleições. Como a amostragem era baseada na lista telefônica e registro de veículos, nem todo o eleitorado fora considerado, uma vez que nem todos tinham acesso ao telefone (GRAY, 2014).

O erro do *Literary Digest* despertou o interesse de acadêmicos por pesquisas de opinião, como Elmo Roper[2] e George Gallup[3]. Ambos haviam acertado a predição da reeleição de Roosevelt em 1936 e continuaram a contribuir com metodologia para as pesquisas e métodos estatísticos para amostragem.

Embora ter telefone fosse sinal de riqueza e *status* em 1940, trinta anos mais tarde praticamente todos os lares americanos já teriam um telefone instalado, abrindo caminho para as pesquisas de opinião por telefone. Com custos mais baixos, as pesquisas de opinião por telefone foram exploradas até 1990, quando entraram em declínio e surgiram as primeiras tentativas de pesquisa de opinião realizadas pela Internet.

1.2 Evolução da Internet

A evolução da Internet trouxe maneiras não-ortodoxas de realizar tarefas rotineiras e afetou os costumes dos usuários que aderiram a essa tecnologia. Nos primeiros passos da Internet, o grupo de usuários que a utilizava era restrito; o acesso era difícil, o conteúdo bastante limitado e, geralmente, sua utilização se dava no contexto bélico e, ou, científico. Devido ao rápido alastramento e adoção dos computadores, aliada à necessidade de melhor comunicação, a Internet virou um dos principais meios de comunicação e cresceu pelo mundo.

[1] *The First Politic Poll*. Disponível em <http://goo.gl/PVwGky>. Acessado em 15/10/2015.
[2] Elmo Roper - *Biography*. Disponível em <http://goo.gl/V6bFFd>. Acessado em 22/10/2015.
[3] George Gallup. Disponível em <http://goo.gl/KxogOg>. Acessado em 22/10/2015.

O grande crescimento da Internet se deve à inclusão digital, propiciada pelo barateamento da tecnologia e computadores em geral. Com a inclusão digital, a Internet deixou de ser algo privilegiado e restritivo, permitindo que mais pessoas tivessem a oportunidade de experimentar a tecnologia. Conforme mais usuários aderiam ao uso da Internet, mais se descobriu sobre o potencial dessa tecnologia (BENEVENUTO et al., 2012).

Pela primeira vez, os usuários tinham como produzir conteúdo e publicá-los de forma relativamente simples, em comparação aos outros meios de comunicação existentes na época. Isso porque a produção de conteúdo geralmente associada aos usuários de Internet raramente passa por um processo de curadoria e não se faz necessário tê-lo. Tal característica facilita a publicação, torna abundante a quantidade de conteúdo e motiva os próprios usuários a se tornarem consumidores deste conteúdo.

Com o avanço na criação de conteúdo e a troca de mensagens de forma fácil e barata pela Internet, surgiu a sociabilidade virtual. Alguns empreendedores vislumbraram a necessidade de ferramentas que facilitassem a comunicação e viabilizassem a prática da sociabilidade virtual. Comunicadores de mensagem instantânea, *blogs*, comunidades e fóruns, foram os precursores das ferramentas que causariam impacto real neste contexto: as redes sociais.

As redes sociais, em sua grande maioria, podem ser representadas na forma de grafo cujos vértices são usuários e cujas arestas indicam a proximidade ou conhecimento recíproco dos vértices em suas extremidades. A interação nas redes sociais se dá através do compartilhamento de algum tipo de conteúdo e este conteúdo pode ser visto, exclusivamente ou não, pelas pessoas conhecidas do autor. O tipo de conteúdo e as restrições associadas ao mesmo dependem do nicho da rede social, os quais são inúmeros, podendo ser citados alguns exemplos como: *microblog* para compartilhamento de mensagens curtas (Twitter), multivariado (Facebook), fotos (Instagram) e vídeos (YouTube).

A partir do momento que as pessoas começaram a dividir tempo de tarefas importantes, como estudar e trabalhar, com a atualização de seus perfis nas redes sociais, as atividades do dia a dia migraram para o ambiente virtual. Entre essas atividades estão: contato com amigos e familiares, ler notícias, compartilhar novidades a respeito de si, expressar opiniões, entre outras. O poder das redes sociais ficou conhecido rapidamente no mundo inteiro e, conforme o número de usuários aumentou, tal ambiente se consagrou como promissor no mundo da propaganda no século XXI.

1.3 Opiniões nas redes sociais

O poder da propaganda é inegável, principalmente para áreas como comércio e política. O interesse de tais áreas no crescimento das redes sociais como ambiente para propagandas se deve ao alcance maior de suas atividades de *marketing* e à facilidade na

obtenção de *feedback*.

Na área da política, os partidos e candidatos foram atraídos pela facilidade que as redes sociais oferecem para fazer rápidas análises da opinião e a intenção de voto do eleitorado em suas regiões. Na área comercial, tornou-se viável divulgar produtos e serviços, além de analisar a opinião do grupo de pessoas atingido por tal ação. Segundo pesquisa recentemente divulgada pela ComScore[4], as análises e opiniões publicadas na Internet sobre produtos têm impacto significativo na decisão de compra.

Nos diferentes nichos de redes sociais que surgiram, observou-se maneiras diferentes de redigir críticas, propiciadas pelas características das aplicações. Sites específicos, como especializados em críticas de filmes, permitem que usuários escrevam textos relativamente longos. Os *microblogs*, por outro lado, impõem limites na quantidade de caracteres das mensagens e não são ambientes exclusivamente destinados para publicação de críticas (Figura 1). No processo de descoberta e pesquisa que prosseguiu nas redes sociais, surgiu a necessidade de expressar opiniões de forma mais direta.

Figura 1 – Exemplo de opinião redigida

Fonte: Twitter - Plataforma de *microblog*.

O modelo mais clássico de expressar opinião sem ser diretamente por texto é o binário, indicando sentimento positivo ou negativo a respeito de algo. A Figura 2 contém informações a respeito de um vídeo publicado no YouTube, tais como título e autor (canto superior esquerdo), além do número de visualizações. Também é possível observar que o YouTube usa uma contagem binária (canto inferior direito) para contabilizar a quantidade de pessoas que indicaram se gostaram ou não do conteúdo.

Com o tempo, surgiram formas diferentes da binária. Os humanos, por exemplo, podem ter opinião neutra acerca de algo ou podem expressar sentimentos com diferentes níveis de intensidade, conforme ilustra a Figura 3. Neste comentário encontrado em uma página de venda de um livro na Amazon, além de opinião expressa por texto, a plataforma também permite que o usuário coloque seu sentimento em uma escala: quanto mais estrelas, maior é a satisfação do usuário.

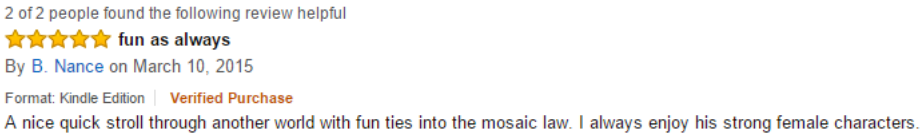
Dificuldades começaram a surgir a partir do momento em que as redes sociais foram inundadas por críticas que salientavam aspectos negativos e os tornavam populares.

[4] *Online Consumer-Generated Reviews Have Significant Impact on Offline Purchase Behavior*. Disponível em <http://goo.gl/PRIHmS>. Acessado em 22/10/2015.

Figura 2 – Exemplo de opinião expressa de forma binária

Fonte: YouTube - Plataforma de compartilhamento de vídeos.

Figura 3 – Exemplo de opinião com intensidade



Fonte: Amazon - Comércio eletrônico.

Tais críticas ficavam acessíveis a um número cada vez maior de indivíduos, sem limites demográficos ou geográficos. Assim, empresas e partidos políticos, dentre muitas outras partes interessadas, foram pressionados a investirem na análise das opiniões publicadas em redes sociais a fim de terem tempo hábil para realizar tomadas de decisões.

Dado o volume muito grande de opiniões que pode ser gerado num pequeno intervalo de tempo na Internet, a utilização de recursos humanos para analisar opinião por opinião não é viável. Com o emprego de recursos computacionais, é possível utilizar técnicas de aprendizado de máquina, os quais simulam a decisão humana baseando-se em conhecimento prévio daquilo que se propõem a analisar. Na tarefa de detecção de opinião automática, as soluções baseadas em métodos de classificação se tornaram recorrentes na literatura (DENECKE, 2008; PANG; LEE; VAITHYANATHAN, 2002).

Os estudos voltados para aplicar processamento de linguagem natural e aprendizado de máquina na análise de sentimentos vêm se tornando abundantes. No entanto, ainda é uma tarefa desafiadora identificar a polaridade de mensagens extraídas das redes sociais. Isso se deve ao fato das amostras normalmente serem curtas e ruidosas, repletas de gírias, abreviações e símbolos.

1.4 A influência dos *smartphones* na comunicação escrita

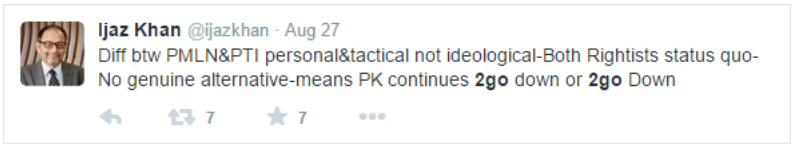
O conceito de *smartphone*, dado como telefonia móvel aliada a computação, foi delineado pela empresa Tesla 1909 em 1971 e patenteado em 1974. No entanto, foi a

partir de 1999 que esta tecnologia se popularizou, aliada à Internet móvel. Com recursos de *hardware* cada vez mais poderosos e constante evolução na Internet móvel, os *smartphones* consolidaram-se como principal meio de acesso às redes sociais nos dias atuais[5].

O intuito de tentar escrever depressa nos *smartphones* leva as pessoas a cometerem erros frequentes e a esquecerem a grafia correta das palavras ou até mesmo a gramática do idioma (CINGEL; SUNDAR, 2012). O uso dos *smartphones* e a popularidade das redes sociais mudou drasticamente a forma como as pessoas escrevem suas mensagens. Ao dividir o tempo do trabalho e dos estudos para atualizar seus perfis em redes sociais, as pessoas se adaptaram a resumir suas mensagens. Além disso, também surgiram redes sociais que impõem limites no tamanho das publicações, como o microblog Twitter.

A necessidade de escrever rapidamente e produzir mensagens curtas levou as pessoas a adotarem artifícios para compactar suas mensagens, tais como abreviações, gírias e símbolos. Não raramente, são introduzidas palavras grafadas incorretamente em meio ao texto devido à pressa, resultando em mensagens curtas e ruidosas. A Figura 4 ilustra um exemplo deste tipo de mensagem, com abreviações como “diff”, “btw” e “2go”, as quais podem ser transcritas como “difference”,“by the way” e “to go”, respectivamente.

Figura 4 – Exemplo de mensagem curta e ruidosa



Fonte: Twitter - Plataforma de *microblog*.

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

Além das mensagens ruidosas, há outros problemas bem conhecidos na literatura, relacionados à detecção de polaridade de mensagens de texto, como frases sarcásticas ou irônicas, discernimento de palavras ambíguas em um determinado contexto (polissemia) ou ocorrência de diferentes palavras com o mesmo significado (sinonímia). Há um

[5] *Socially Mobile: Does the Smartphone Rule Social Media?*. Disponível em <http://goo.gl/eD9z4C>. Acessado em 15/10/2015.

consenso de que tratar estes casos pode conduzir a melhores resultados na detecção de opinião (PANG; LEE, 2008; MOSTAFA, 2013). Os exemplos a seguir refletem algumas dessas dificuldades.

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

Problemas como sinonímia e polissemia podem ser minimizados através de indexação semântica, empregada para realizar a desambiguação de palavras. Tal técnica traça a relação de significados dos termos, permitindo encontrar termos com significados próximos ou polissêmicos no contexto da mensagem. Em geral, o emprego eficiente dos dicionários é condicionado à qualidade da extração dos termos das amostras (NAVIGLI; PONZETTO, 2012; TAIEB; AOUICHA; HAMADOU, 2013). Os métodos de indexação semântica e desambiguação de palavras são discutidos com mais detalhes no capítulo seguinte.

2 Representação computacional

Conforme descrito por Mitchell (MITCHELL, 1997), o campo de pesquisa do aprendizado de máquina é focado em desenvolver algoritmos que aprendam com base em experiência. Este mesmo autor oferece uma definição para programas de computador baseados em aprendizado de máquina:

> "Um programa de computador aprende a partir de uma experiência E com relação a classes de tarefa T e medida de desempenho P, se o desempenho na tarefa T, medido por P, melhora com a experiência E."

Para o problema tratado neste trabalho, a experiência E é dada pelas mensagens, que podem ser curtas e ruidosas. A análise de sentimento a ser realizada sobre cada mensagem é a tarefa T, cujo desempenho pode ser medido por P, onde P é o sentimento esperado para as mensagens analisadas. Assim, dado um algoritmo de aprendizado de máquina, espera-se que ele aprenda a partir de várias amostras de treinamento que possuem opiniões associadas (rótulos), para gerar um modelo genérico e eficaz que seja capaz de predizer satisfatoriamente qual sentimento está associado às mensagens ainda não vistas.

A experiência E também pode ser interpretada como a representação que uma amostra tem baseada nas características relevantes para a tarefa T. Quanto melhor a representação de um problema, maior a probabilidade de um algoritmo de aprendizado de máquina gerar hipóteses satisfatórias (HALEVY; NORVIG; PEREIRA, 2009).

Conhecendo as características intrínsecas de um determinado problema, é possível projetar um sistema que leve em conta somente as características mais relevantes do domínio, sem inserir características que sejam correlatas a outras existentes ou que não ajudam a resolver o problema proposto.

Um exemplo clássico de características relevantes é a base de dados Íris (FISHER, 1936), cujas amostras são divididas em três categorias (íris virgínica, íris setosa e íris versicolor) e possuem quatro características (comprimento e largura da pétala, comprimento e largura da sépala). Essas quatro características, comumente chamadas de atributo em aprendizagem de máquina, são relevantes para o problema de classificar esse gênero de flor íris entre as três espécies disponíveis na base de dados.

Em alguns domínios, é bastante trivial abstrair potenciais características para descrever um problema, embora nem sempre seja fácil de obtê-las ou representá-las. Por exemplo, no processamento de imagens seria comum pensar em classificar objetos de uma imagem com base na forma, na textura e na cor. No entanto, supondo que todas as imagens a serem analisadas tenham a mesma dimensão em quantidade de pixels, contar

o número de pixels de cada amostra não constitui uma característica relevante para o problema.

Outra dificuldade comumente encontrada é transferir a representação baseada na abstração do mundo real para uma representação computacional. No exemplo da base de dados Íris, o comprimento e a largura da pétala e sépala podem ser representados como quatro pontos flutuantes em termos computacionais, um para cada atributo, de forma trivial. No entanto, representar quão côncava é a pétala é mais abstrato, portanto mais difícil de representar computacionalmente.

Encontrar formas de representar texto é um problema que tem sido explorado há muitos anos, e há inúmeras maneiras de realizar esta tarefa embora ainda não exista consenso de qual seja melhor. Algumas das técnicas exploram o significado do texto através das palavras isoladas ou permitem a análise da estrutura morfológica e sintática. Tais técnicas também diferem entre si com relação à porção da amostra de texto a ser analisada, seja por sentença, parágrafos ou todo o conjunto (HARRIS, 1954; AGGARWAL; ZHAI, 2012).

A representação de texto e a sua implicação nos métodos de aprendizado de máquina ainda são desafios em aberto. Para explorar esta questão e prover conhecimento necessário para se compreender o restante deste trabalho, este capítulo é dividido em duas seções. A Seção 2.1 aborda o conceito da representação computacional de texto e a Seção 2.2 detalha as diferentes técnicas de processamento de linguagem natural empregadas para lidar com mensagens curtas e ruidosas.

2.1 Representação computacional de texto

A definição de programa de computador que aprende a partir de experiências sugerida por Mitchell, demanda uma representação computacional eficiente. Neste trabalho, é utilizado o modelo de representação mais tradicional, conhecido como *bag-of-words*. Principalmente, aplicado no contexto de recuperação de informações e processamento de linguagem natural, tal modelo de representação traça a relação entre termos de um vocabulário conhecido e a presença deles nas amostras (FRAKES; BAEZA-YATES, 1992).

Os termos podem ser considerados como a menor unidade de uma amostra de texto. A obtenção dos termos se dá pela aplicação de regras de segmentação, as quais separam a amostra original em um grupo de termos, cujos termos podem repetir dentro da mesma amostra. Os termos resultantes da segmentação podem ser vocábulos idiomáticos, números, símbolos ou uma mistura deles, tais como “ele”, “2015”, “:)” e “Jan/2015”. A regra mais simples e comum de segmentação é utilizar o espaço em branco para dividir a amostra original. No entanto, esta regra pode variar conforme a especificidade do idioma e aplicação. Por exemplo, os idiomas chinês e japonês não usam espaços em branco na

construção de frases (GASPERIN; LIMA, 2001).

A regra de segmentação ("tokenização") pode ser aplicada com diferentes valores de granularidade (*n-gram*), aumentando ou diminuindo a menor unidade possível de uma amostra. Por exemplo, para granularidade 2 em uma determinada regra de segmentação, o menor elemento de um vocabulário seria composto por dois termos. Define-se como vocabulário, o conjunto de todos possíveis termos segmentados, respeitando a granularidade definida. Na Tabela 1 são apresentados dois vocabulários obtidos com granularidades diferentes para uma mesma amostra.

Tabela 1 – Diferentes representações de vocabulários para a amostra "*I bought a crappy gift*".

Representação	Vocabulário
1-gram	{*I*, *bought*, *a*, *crappy*, *gift*}
2-gram	{*I bought*, *bought a*, *a crappy*, *crappy gift*}

Para o valor de granularidade igual a um *(1-gram)*, cada termo é considerado como um elemento diferente no vocabulário. No entanto, para granularidades maiores, a combinação de termos próximos caracteriza um novo elemento do conjunto. No exemplo dado, para granularidade *2-gram*, cada elemento é composto por dois termos "*I bought*", "*bought a*", "*a crappy*" e "*crappy gift*".

É possível notar que o vocabulário computado com valor de granularidade superior a um é capaz de capturar o significado de termos próximos. Por exemplo, a negação antes de um termo seria levada em consideração, evitando que "*not good*" fosse separado em "*not*" e "*good*". No entanto, isso produz matrizes mais esparsas e não apresenta grande diferença de desempenho em análise de sentimento (GO; BHAYANI; HUANG, 2009).

Uma representação bastante conhecida para texto é a *bag-of-words*. Tal representação se baseia na construção do vocabulário e na segmentação das mensagens em termos. Considerando que cada mensagem m é composta por um conjunto de termos $m = t_1, \ldots, t_n$, sendo que cada termo t_k corresponde a um dos elementos de um vocabulário conhecido, pode-se representar cada mensagem como um vetor $\vec{x} = \langle x_1, \ldots, x_n \rangle$, onde $x_1, \ldots, x_n$ são valores dos atributos $X_1, \ldots, X_n$ associados aos termos $t_1, \ldots, t_n$. Neste trabalho, os atributos são valores binários que representam se determinado termo ocorre ou não na mensagem.

Um exemplo de representação *bag-of-words* é dado a seguir. Sejam as amostras:

(A) "*I bought a crappy gift.*"

(B) "*I purchased a good gift.*"

O vocabulário computado com valor de granularidade igual a um, para estas amostras, é dado por {*I*, *bought*, *purchased*, *crappy*, *good*, *a*, *gift*}.

Em sequência, a representação é feita no formato de matriz documento-termo (Tabela 2), onde o número de linhas m corresponde à quantidade de amostras e número de colunas (n) corresponde ao tamanho do vocabulário, ou quantidade de atributos. O valor dos atributos indica se determinado termo do vocabulário ocorre em determinada amostra.

Tabela 2 – Exemplo de representação com *bag-of-words*.

	I	*bought*	*purchased*	*crappy*	*good*	*a*	*gift*
(A)	1	1	0	1	0	1	1
(B)	1	0	1	0	1	1	1

No exemplo dado, os vocábulos "*bought*" e "*purchased*" têm o mesmo significado, mas são representados em colunas distintas. Este fenômeno é conhecido como sinonímia e é bastante característico ao lidar com texto.

Outro fenômeno bastante comum é a polissemia, o qual é caracterizado por vocábulos iguais que apresentam significados diferentes dependendo do contexto. Desta forma, dois termos com significados diferentes são representados na mesma coluna. Por exemplo, nas frases *"My head hurts."* e *"You head the news office."*, o termo *head* tem significados distintos: substantivo cabeça e verbo dirigir, respectivamente.

Tal dificuldade em lidar com sinonímia e polissemia para encontrar boas representações de texto são ainda mais agravantes em mensagens curtas e ruidosas. Ao abreviar palavras, como "*great*" para "*gr8*", as pessoas geram termos que afetam a representação de texto da mesma forma que a sinonímia; enquanto que palavras grafadas incorretamente, como "*mispelled*" e "*misspelled*", têm o mesmo efeito que a polissemia. Neste sentido, há diferentes técnicas de processamento de linguagem natural que auxiliam na redução da ocorrência desses fenômenos.

2.2 Técnicas de processamento de linguagem natural

Em categorização de texto, o emprego direto da *bag-of-words* para representar mensagens curtas e ruidosas, repletas de gírias, símbolos e abreviações, geralmente conduz a resultados insatisfatórios (GABRILOVICH; MARKOVITCH, 2005; GABRILOVICH; MARKOVITCH, 2007).

As mensagens curtas e ruidosas são decorrentes do limite imposto pelo canal de comunicação, como a rede social Twitter, e devido aos novos hábitos de comunicação escrita desenvolvido pelas pessoas. Estas mensagens podem ser tão pequenas ao ponto de seu conteúdo conter apenas um único símbolo, como ":)" ou ":(". O ruído pode se caracte-

rizar por palavras grafadas incorretamente ou abreviadas, siglas, expressões idiomáticas, vocábulos estrangeiros ou símbolos, como os já citados.

Para lidar com as dificuldades impostas pelas mensagens curtas e ruidosas, um sistema de expansão[1] com etapas bem definidas foi proposto em (SILVA et al., 2014). Os autores identificaram dois problemas-chaves ao lidar com este contexto e propuseram soluções: as mensagens ofuscadas prejudicam o processamento de texto e a quantidade pequena de informação contida nelas é insuficiente para treinar os métodos de classificação.

Tal sistema de expansão é utilizado neste trabalho como alicerce para a proposta a ser discutida adiante. Nele, a mensagem de entrada é processada em etapas, sendo que em cada uma delas é gerada uma saída diferente, as quais são combinadas posteriormente para formar uma nova amostra. Dentre as etapas disponíveis estão a normalização léxica, a geração de conceitos por indexação semântica e a desambiguação de conceitos. Estas etapas são detalhadas nas subseções a seguir.

2.2.1 Normalização léxica

A tarefa de traduzir variantes léxicas de palavras e expressões normalmente ofuscadas para sua forma canônica se chama normalização léxica. As variantes léxicas e expressões podem ser encontradas de diferentes formas, tais como letras repetidas, como *"goood"* que pode ser traduzida para *"good"*; abreviações como "b4", que significa *"before"*; siglas como *"afk"* que pode ser traduzida como *"away from keyboard"*; ou até mesmo reduzir as palavras para radicais, removendo termos muito semelhantes como plurais ou superlativos, tais como *"stores"* para *"store"* e *"largest"* para *"large"*, respectivamente.

A importância da aplicação da normalização léxica é devida justamente ao impacto que a mesma causa na representação *bag-of-words*, ao aproximar termos equivalentes e reduzir o número de elementos do conjunto vocabulário, além de auxiliar outras técnicas de processamento de linguagem natural que dependem de palavras grafadas corretamente (SILVA et al., 2014). Assim, ao pré-processar amostras desconhecidas, aumenta-se a chance de algum termo semelhante ser substituído por um elemento contido no vocabulário.

Nesta etapa, foram empregados dois dicionários para normalizar e traduzir termos conhecidos como *Lingo*, nome dado ao conjunto de gírias e abreviações comumente utilizados na Internet, por palavras da língua inglesa. O primeiro é um dicionário padrão de inglês[2], usado para consultar se uma determinada palavra existe no idioma inglês ou não, e reduzir para seu radical quando possível. O segundo dicionário é o de Lingo[3], que

[1] *TextExpansion*. Disponível em <http://lasid.sor.ufscar.br/expansion/>. Acessado em 22/10/2015.
[2] *Freeling English dictionary*. Disponível em <http://devel.cpl.upc.edu/freeling/>. Acessado em 22/10/2015.
[3] *NoSlang Lingo dictionary*. Disponível em <http://www.noslang.com/dictionary/full/>. Acessado em

é utilizado para traduzir um termo Lingo para sua forma canônica em inglês. Caso não haja uma tradução para o termo de entrada, o termo original é mantido.

2.2.2 Geração de conceitos por indexação semântica

A tarefa de analisar termos isolados de uma mensagem e encontrar sinônimos para eles a partir de uma base de conhecimento é chamada de geração de conceitos por indexação semântica. Neste trabalho, os conceitos são provenientes do repositório BabelNet (NAVIGLI; LAPATA, 2010), um moderno e grande dicionário semântico da língua inglesa, composto por conceitos extraídos da WordNet e Wikipedia. Como esta etapa demanda entradas em inglês, a *normalização léxica* é aplicada para garantir que os termos estão, de fato, em inglês.

Nos estágios iniciais do experimento deste trabalho, a ser discutido posteriormente, a base de conhecimento Wikipedia trouxe ruídos em excesso para as amostras ao ser empregada a técnica de geração de conceitos. Portanto, no contexto das futuras discussões abordadas neste trabalho, o repositório BabelNet foi configurado para utilizar apenas a base de conhecimento WordNet[4].

WordNet é um sistema *online* de referência léxica, projetado a partir de teorias psicolinguísticas acerca da memória léxica humana. A organização neste sistema se dá pelo significado das palavras e não pela forma delas. Sua principal estrutura é representada por conjuntos de sinônimos contendo substantivos, verbos, adjetivos e advérbios. A versão 3.0 contém 155 mil palavras e 117 mil conjuntos de sinônimos (NAVIGLI; LAPATA, 2010).

O sistema de conjunto de sinônimos pode ser melhor compreendido com um exemplo. Considere o vocábulo *"drink"* e três possíveis diferentes significados para ele: (a) "consumir líquidos", (b) "consumir álcool" e (c) "brinde". Para cada uma destas possíveis interpretações, suponha que o sistema WordNet traz um conjunto diferente de sinônimos:

(a) $\{drink_v^1, imbibe_v^3\}$

(b) $\{drink_v^2, booze_v^1, fuddle_v^2\}$

(c) $\{toast_v^2, drink_v^3, pledge_v^2, salute_v^1, wassail_v^2\}$

Cada palavra do conjunto de sinônimos é anotada com o seu tipo gramatical (n para substantivo, v para verbo, a para adjetivo e r para advérbio), e o número atribuído corresponde a ordem com que este termo e significado ocorrem no *corpus* do sistema WordNet. No exemplo, o número 1 está atribuído à palavra *"drink"* do primeiro conjunto de sinônimo, pois o termo *"drink"* é mais frequente neste *corpus* com o significado de consumir líquidos.

22/10/2015.

[4] *WordNet*. Disponível em <http://wordnet.princeton.edu/>. Acessado em 22/10/2015.

O funcionamento esperado na etapa de geração de conceitos é a obtenção de uma lista de sinônimos para cada termo de uma mensagem, combinando todos os possíveis significados deste termo em diferentes conjunto de sinônimos. Assim, ao pesquisar a palavra *"drink"*, os termos resultantes seriam a união dos conjuntos a, b e c descritos anteriormente.

A geração de conceitos deve ser usada com cautela, pois pode agregar muito ruído à mensagem, prejudicando o desempenho da representação de texto, em vez de melhorá-lo. A desambiguação de conceitos é geralmente aplicada em combinação com a geração de conceitos para remover ruídos e tornar a amostra mais significativa.

2.2.3 Desambiguação de conceitos

A quantidade de conceitos gerados para cada termo pode ser excessiva, deixando a amostra original com bastante ruídos. Uma técnica de desambiguação não-supervisionada foi proposta na literatura para selecionar o conceito mais relevante para cada palavra, de acordo com o seu contexto na mensagem original (NAVIGLI; PONZETTO, 2012).

Tal técnica explora a conectividade do grafo montado pelas relações anotadas no sistema WordNet. Cada vértice deste grafo é um conjunto de sinônimos e as arestas são relações léxicas e, ou, semânticas entre estes conjuntos de sinônimos. A partir da mensagem original, são separados os termos a serem procurados no grafo e, para cada termo, é realizada uma busca em profundidade. Os menores caminhos encontrados para cada busca, a partir de cada termo procurado, formam um novo grafo, e os melhores conceitos são selecionados a partir de medidas de adjacência calculadas neste novo grafo.

Para exemplificar, suponha a seguinte frase *"She drunk some milk"*. Uma lista de palavras comuns (*stopwords*) pode ser empregada para excluir os termos *"she"* e *"some"*, a fim de evitar buscas exaustivas e economizar recursos computacionais. O verbo "drunk" também pode ser reduzido para o infinitivo a fim de generalizar o termo ("drink" sem a partícula "to"). Assim, a geração de conceitos e a desambiguação destes são feitos a partir das palavras *"drink"* e *"milk"*. O sistema WordNet é consultado para identificar quantos são os conjuntos de sinônimos para cada uma destas palavras e, em seguida, a busca em profundidade é executada a partir destes conjuntos de sinônimos diferentes.

Com a busca em profundidade, é encontrado o menor caminho entre os elementos do conjunto de sinônimos de *"drink"* e outro elemento dos conjuntos de sinônimos do mesmo termo em outra ordem (um caminho de $drink^1$ até $drink^2$, por exemplo) ou um elemento do conjunto de sinônimos dos outros termos da mesma mensagem (um caminho de $drink^1$ até $milk^1$, por exemplo). Finalmente, a decisão de quais conceitos são retornados é tomada com base na distância entre esses termos e seus sinônimos.

As técnicas de processamento de linguagem natural descritas até aqui podem ser combinadas de diferentes maneiras para gerarem amostras expandidas. Uma amostra pode ser expandida utilizando uma única técnica ou a combinação de mais de uma técnica. Para cada combinação de técnicas, um conjunto diferente de amostras expandidas é gerado.

As possíveis regras de combinação podem ser pensadas como análise combinatória das quatro possíveis etapas de expansão: "original", "normalização", "geração de conceitos" e "desambiguação". No entanto, nem todas devem ser computadas, como a regra de combinação ["Geração de conceitos" + "Desambiguação"] que tem resultado idêntico à regra ["Geração de conceitos"], uma vez que a etapa de "desambiguação" resulta em um subconjunto dos conceitos obtidos na etapa "geração de conceitos". Nos experimentos discutidos posteriormente, são empregadas apenas as regras de combinação listadas na Tabela 3.

No código de quatro letras associado às regras de combinação, cada posição corresponde a uma etapa ou técnica de normalização e indexação semântica. A letra em determinada posição indica se os termos de uma determinada etapa são incluídos "Y", ou não "N". Assim, a regra de combinação "YNNN" indica que os termos originais são utilizados, mas não são utilizados os termos obtidos nas outras etapas.

Tabela 3 – Regras de combinação de técnicas de normalização e indexação semântica.

	Técnicas de normalização e indexação semântica			
Regra	Original	Normalização	Geração de conceitos	Desambiguação
E_1 - YNNN	Sim	Não	Não	Não
E_2 - YYNN	Sim	Sim	Não	Não
E_3 - YYYN	Sim	Sim	Sim	Não
E_4 - YYNY	Sim	Sim	Não	Sim
E_5 - YNYN	Sim	Não	Sim	Não
E_6 - YNNY	Sim	Não	Não	Sim
E_7 - NYNN	Não	Sim	Não	Não
E_8 - NYYN	Não	Sim	Sim	Não
E_9 - NYNY	Não	Sim	Não	Sim
E_{10} - NNYN	Não	Não	Sim	Não
E_{11} - NNNY	Não	Não	Não	Sim

Para exemplificar como funciona a regra de combinação, suponha que a frase *"plz lemme noe when u get der"* fosse processada pela regra *NYYN* - ["Normalização" + "Geração de conceitos"]. O resultado seria idêntico ao demonstrado na Tabela 4, onde cada linha representa a saída de cada etapa. De acordo com a regra de combinação escolhida, a amostra expandida deverá conter os termos resultantes da etapa de normalização acrescidos aos termos da geração de conceitos, sem incluir elementos da amostra original e desambiguação.

Supondo que a regra de combinação fosse *NYNY* - ["Normalização" + "Desambi-

Tabela 4 – Exemplo de amostra produzido pela regra de combinação [“Normalização” + “Geração de conceitos”] na mensagem *“plz lemme noe when u get der”*.

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

guação”] para a mesma frase de entrada, a amostra final seria *“please let lease me know cognition when you get there”*, pois seriam incluídos apenas os termos resultantes das etapas de normalização e desambiguação.

Após utilizar essas técnicas de normalização e indexação semântica, espera-se que a amostra resultante seja mais informativa e adequada para os métodos de aprendizado de máquina. Nota-se que neste contexto, a geração de conceitos pode preencher a amostra com muito ruído. Isso pode ser decorrente dos termos pesquisados ou do domínio no qual a mensagem está inserida. Não há consenso de que exista uma regra de combinação única que seja a melhor para todos os cenários, variando conforme o domínio ao qual são aplicadas e a quais métodos de aprendizado de máquina tais representações são submetidas.

3 Métodos de classificação

A tarefa da análise de sentimento nos dias atuais é normalmente realizada através de métodos de aprendizado de máquina. Esses métodos podem ser divididos pelo tipo de sinal ou retorno que os sistemas utilizam para aprender, sendo geralmente divididos em três grandes grupos: aprendizagem por reforço, supervisionada e não-supervisionada (RUSSELL; NORVIG, 2003).

- Aprendizagem por reforço: um programa de computador interage com um ambiente dinâmico e aprende conforme explora esse ambiente;
- Aprendizagem supervisionada: um programa de computador recebe amostras (entradas) e respostas esperadas (saídas) para as mesmas, e gera uma hipótese genérica capaz de mapear as entradas para as saídas corretas; e
- Aprendizagem não-supervisionada: um programa de computador recebe amostras sem saídas esperadas e deve aprender diferentes estruturas para os dados recebidos.

Dentro destes grupos há subdivisões, dependendo do tipo de dado analisado, mas no contexto deste trabalho o interesse está voltado para a aprendizagem supervisionada. Neste grupo, há dois tipos diferentes de possíveis saídas: discretas e contínuas. As saídas discretas, também chamadas de rótulos ou classes, são características nos métodos de classificação; enquanto as saídas contínuas são características nos métodos de regressão.

A análise de sentimento é uma tarefa tipicamente de aprendizagem supervisionada, especificamente de classificação. Para cada mensagem de entrada é esperada uma saída discreta, como categorias bem definidas, tais como “sentimento positivo”, “negativo” ou “neutro”; ou um intervalo discreto, supondo que o sentimento seja denotado por um intervalo de inteiros $[1, 5]$, onde 1 corresponde ao sentimento mais negativo e 5 ao sentimento mais positivo. Neste trabalho, as possíveis classes esperadas ao analisar o sentimento associado a uma mensagem são: sentimento positivo e sentimento negativo.

Na literatura, há várias propostas de métodos de classificação empregados nas mais diversas tarefas, os quais são normalmente agrupados pelo tipo de estratégia de seleção de hipótese empregada, tais como probabilidade, otimização, distância e árvores. Em 2008, um importante estudo elencou os dez melhores métodos de classificação para contextos diversos devido à alta capacidade de generalização (WU et al., 2008), os quais voltaram a aparecer em outro estudo semelhante, porém mais recente (FERNANDEZ-DELGADO et al., 2014).

Consagrados como estado da arte, as seções a seguir oferecem uma sucinta introdução acerca de cada um dos seguintes métodos: k-NN, regressão logística, máquinas de vetores de suporte, métodos baseados em árvores (C4.5 e florestas aleatórias), *Naïve* Bayes e comitês de máquina de classificação.

3.1 k-Vizinhos Próximos

O método k-Vizinhos Próximos (do inglês, *k-Nearest Neighbors* - k-NN) é da família dos métodos de classificação baseados em proximidade, portanto é capaz de mapear padrões entre dados de entrada e saídas discretas baseando-se em medidas de dissimilaridade. O princípio de funcionamento destes algoritmos é que as amostras de uma mesma classe são próximas no espaço n-dimensional e a medida de dissimilaridade habitualmente utilizada é a distância euclidiana (SALTON; MCGILL, 1986).

A etapa de classificação neste método envolve calcular a dissimilaridade de uma nova amostra para todas as amostras conhecidas e classificá-la com a categoria mais frequente entre as "k" amostras mais próximas a ela. O algoritmo k-NN não gera um modelo a partir das amostras, mas consulta a todas para cada nova classificação. Além disso, o algoritmo k-NN é bastante penalizado quando uma das classes do problema apresenta muito mais amostras que as demais classes.

3.2 Regressão logística

A regressão logística é uma adaptação da regressão linear, capaz de gerar hipóteses com saídas discretas, a qual mapeia a relação entre uma variável dependente categórica com uma ou mais variáveis independentes (WALKER; DUNCAN, 1967).

O funcionamento deste método é focado em encontrar os melhores coeficientes para cada atributo que conduzem à melhor função para dividir as classes. No plano 2D, por exemplo, a função ótima seria aquela que projeta uma reta capaz de separar as amostras de classes diferentes. Tais coeficientes são obtidos por meio de otimização, semelhante aos modelos lineares genéricos.

O processo de classificação baseia-se em passar uma nova amostra pela função ótima, multiplicando os coeficientes encontrados com os valores da amostra, e utilizar este resultado como entrada da função sigmoidal para que esteja no intervalo $[0, 1]$. Aos valores maiores que $0,5$, tal amostra é classificada como classe positiva, caso contrário como classe negativa.

3.3 Máquinas de vetores de suporte

A técnica máquinas de vetores de suporte (do inglês, *Support-vector Machines* - SVM) foi proposta em (CORTES; VAPNIK, 1995) como um classificador linear e tem sido reconhecida como um dos métodos de classificação mais promissores dentre os disponíveis. O funcionamento do método é semelhante à regressão logística, que busca encontrar um hiperplano capaz de separar as classes analisadas. A diferença que torna as máquinas de vetores de suporte tão poderosa, é que este método busca o hiperplano ótimo, ou seja, aquele com a maior margem de separação das classes dentre os hiperplanos possíveis. O hiperplano ótimo pode ser visto como a hipótese mais genérica também.

Classificadores não-lineares baseados em SVM foram propostos utilizando diferentes funções de *kernel* (BOSER; GUYON; VAPNIK, 1992). O *kernel* é uma função não-linear que substitui o produto escalar usado no classificador linear para transformar o espaço dos atributos, de modo que possa existir um espaço n-dimensional cujas classes são separáveis por um hiperplano. Neste trabalho, além do SVM linear, são utilizadas outras duas varições deste método: SVM com *kernel* polinomial e SVM com *kernel* radial.

3.4 Métodos baseados em árvores

Os métodos baseados em árvores de decisão são populares na literatura. Basicamente, uma árvore de decisão é a decomposição hierárquica do espaço de dados, obtida ao encontrar os melhores atributos para separar as classes de um determinado problema, em passos sucessivos. Nas folhas da árvore estão as classes, e nos diferentes níveis acima estão os atributos capazes de prover a melhor separação, com o atributo mais relevante no topo (BREIMAN et al., 1984).

A ordem que os atributos são escolhidos na estrutura da árvore é feita com base em algumas medidas, como entropia, informação mútua, índice gini, χ^2, entre outros (AGGARWAL; ZHAI, 2012).

Quanto mais decisões tiver em uma árvore de decisão e menos registros tiver em cada folha, mais propícia a árvore fica em se tornar especialista nos dados de treinamento, reduzindo seu potencial de generalizar o problema. Uma das propostas encontradas na literatura para tornar as árvores de decisão mais genéricas é o método de florestas aleatórias.

3.4.1 Florestas aleatórias

O método de classificação conhecido como florestas aleatórias consiste em gerar várias árvores de decisão e eleger a classe final com base no voto de cada uma das árvores pertencentes à floresta. São iniciadas n árvores baseadas em subconjuntos aleatórios do

conjunto de treinamento original, as quais compõem a floresta aleatória (BREIMAN, 2001).

O processo de classificação é feito de forma semelhante ao da árvore de decisão. Uma nova amostra é classificada em cada árvore da floresta aleatória, e a classe mais comum como resposta é eleita a classe final da nova amostra.

3.5 *Naïve* Bayes

O método *naïve* Bayes é da família dos métodos de classificação probabilísticos baseado no teorema de Bayes, e assume que o valor dos atributos contribuem de forma independente da probabilidade de uma determinada amostra pertencer a uma classe (LANGLEY; IBA AND; THOMPSON, 1992).

Seja uma amostra representada pelo vetor $\vec{x} = \langle x_1, \ldots, x_n \rangle$, com n atributos, a probabilidade desta amostra pertencer a uma determinada classe C_k é dada por $p(C_k|\mathbf{x})$, a qual pode ser decomposta com o teorema de Bayes em

$$p(C_k|\mathbf{x}) = \frac{p(C_k)\ p(\mathbf{x}|C_k)}{p(\mathbf{x})},$$

onde $P(C_k)$ é a probabilidade a *priori* associeda à ocorrência da classe C_k, $P(x)$ é a probabilidade a *priori* associeda à ocorrência da amostra x e $P(x|C_k)$ é a probabilidade associeda à ocorrência da amostra x dada a classe C_k.

Como este método assume que os atributos são independentes, essa premissa permite que o cálculo de probabilidade de um evento para uma classe seja simplificado como

$$p(C_k|x_1, \ldots, x_n) = p(C_k) \prod_{i=1}^{n} p(x_i|C_k),$$

o qual leva a um classificador probabílistico que oferece uma classe $\hat{y}$, conhecendo-se as probabilidades de cada valor de cada atributo ocorrer para cada classe, descrito como

$$\hat{y} = \underset{k \in \{1,\ldots,K\}}{\operatorname{argmax}}\, p(C_k|x_1, \ldots, x_n).$$

Este grupo de métodos baseado no teorema de Bayes consegue generalizar hipóteses em torno de problemas multiclasses, trabalha bem com atributos categóricos, além de obter bons resultados com base de dados pequenas. No entanto, este método é sensível ao pré-processamento dos dados, tal como valores contínuos que precisam ser discretizados para que as probabilidades sejam calculadas (HARRINGTON, 2012). Além disso, um outro problema conhecido dos métodos probabilísticos é o baixo desempenho ao lidar com espaços de atributos de alta dimensionalidade. Essa característica é conhecida como maldição da dimensionalidade e muitos trabalhos na literatura recomendam aplicar seleção de atributos para evitá-la (ALMEIDA; ALMEIDA; YAMAKAMI, 2011).

O método *naïve* Bayes apresenta variações devido à modelagem de evento no cálculo da probabilidade. Nos experimentos deste trabalho foram avaliadas três técnicas diferentes: *naïve* Bayes Gauss, *naïve* Bayes Bernoulli e *naïve* Bayes Multinomial.

3.5.1 *Naïve* Bayes Gauss

O modelo gaussiano aplicado ao método *naïve* Bayes surgiu da necessidade de utilizar atributos contínuos. Seja μ_c a média dos valores dos atributos em x associado a uma classe C_k, e σ_c^2 a variância destes valores, a distribuição da probabilidade de uma determinada amostra para uma classe passa a ser dada por

$$p(\mathbf{x}|C_k = c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} e^{-\frac{(x-\mu_c)^2}{2\sigma_c^2}},$$

onde a distribuição normal é parametrizada com μ_c e σ_c^2.

3.5.2 *Naïve* Bayes Multinomial

O modelo multinomial é normalmente empregado em classificação de texto, onde os eventos são descritos como a ocorrência de um termo em uma determinada amostra. Desta forma, cada amostra é representada como um histograma, onde os atributos armazenam a contagem de um determinado evento x_i para uma amostra particular (RENNIE et al., 2003). Assim, a probabilidade de um evento com relação a uma classe é dada por

$$p(\mathbf{x}|C_k) = \frac{(\sum_i x_i)!}{\prod_i x_i!} \prod_i {p_{ki}}^{x_i},$$

onde $p_k i$ é a distribuição multinomial dos possíveis valores que um evento i pode assumir em uma determinada classe k.

3.5.3 *Naïve* Bayes Bernoulli

Diferente do modelo multinomial, o modelo de Bernoulli também considera a ausência de um determinado evento no cálculo da probabilidade de uma determinada amostra pertencer a uma classe (MCCALLUM; NIGAM, 1998). Este modelo também tem como entrada um vetor de atributos binários. O cálculo da probabilidade de um evento pertencer a uma classe, neste método, é dado por

$$p(\mathbf{x}|C_k) = \prod_{i=1}^{n} p_{ki}^{x_i}(1 - p_{ki})^{(1-x_i)}.$$

Ainda há outra vertente de estudos na literatura que busca combinar diferentes métodos de classificação para obter hipóteses ainda mais genéricas. Tais sistemas de combinação são conhecidos como comitê de máquinas de classificação ou *ensemble*. Há diversas

estratégias envolvidas na combinação de classificadores e maiores detalhes são oferecidos na seção a seguir.

3.6 Comitê de máquinas de classificação

A técnica do comitê de máquinas de classificação, ou *ensemble*, foi desenvolvida para obter hipóteses mais genéricas através da combinação de diferentes classificadores. Em um comitê, cada classificador é um membro com poder de voto e a decisão final depende dos votos de todos os membros, podendo cada um ter voto com peso diferente. Essa técnica é aplicada normalmente para minimizar as deficiências de cada método de classificação, evitando o super-ajustamento e a maldição da dimensionalidade. Embora hajam diferentes tipos de comitês de classificadores, comumente eles possuem maior poder preditivo que métodos tradicionais isolados (DIETTERICH, 2000; WANG et al., 2014).

Em comitês de máquinas de classificação, a decisão da classe final e a relação desta com os votos de cada membro também é uma área de estudo. Os comitês de máquinas de classificação com votação majoritária ponderada são encontrados com bastante frequência na literatura. A eficácia desses métodos geralmente está associada ao peso que cada classificador tem em seu voto, de forma que, em um dado cenário, um classificador com menor poder preditivo possui voto com peso inferior ao voto de um classificador com maior capacidade de predição (XIA; ZONG; LI, 2011).

A estratégia sobre a qual um comitê de máquinas é montado pode variar, dependendo de como os classificadores são criados. Há estratégias que envolvem treinar classificadores com subconjuntos diferentes de atributos, subconjuntos diferentes de amostras, utilizar vários métodos de classificação, entre outras. As duas estratégias mais conhecidas são *bagging* e *boosting*.

3.6.1 *Bagging*

A estratégia *bagging* de classificadores consiste em treinar diferentes métodos de classificação em subconjuntos diferentes de dados. Estes classificadores podem ser baseados em um mesmo método de classificação ou podem ser de métodos diferentes (BREIMAN, 1996). Por exemplo, a floresta aleatória é intrinsecamente um comitê de classificadores baseado em *bagging*, pois gera diferentes árvores de decisão para subconjuntos aleatórios das amostras originais.

Essa estratégia é capaz de tornar o modelo genérico, pois cada classificador fica especializado em porções diferentes de dados, evitando ficar especialista no conjunto completo original.

3.6.2 *Boosting*

A estratégia *boosting* de classificadores consiste em treinar diferentes métodos de classificação, ou modelos, incrementalmente a partir de erros dos classificadores anteriores. As amostras rotuladas incorretamente são reintroduzidas na fase de treino e um novo modelo gerado. O algoritmo mais famoso a adotar essa estratégia é o *Adaboost* (SCHAPIRE, 1999).

A abundância de técnicas de processamento de texto e métodos de classificação existentes demanda a escolha de alguma estratégia para combiná-los de tal forma a obter hipóteses robustas e genéricas. Como as mensagens de texto podem ser processadas por diversas técnicas de NLP que, por sua vez, podem ser combinadas com vários métodos de classificação recomendados para a detecção de polaridade, pressupõem-se que um sofisticado sistema de comitê de máquinas de classificação seja capaz de gerar hipóteses mais genéricas e, portanto, com maior capacidade preditiva. A proposta deste trabalho é oferecer tal sistema, o qual é discutido no capítulo seguinte.

4 Sentminer

Sentminer foi o nome escolhido para o sistema que derivou a partir das pesquisas deste trabalho. Tal sistema combina diferentes métodos de classificação com técnicas de processamento de linguagem natural a fim de oferecer modelos genéricos, porém robustos, capazes de identificar a polaridade associada a uma mensagem, sendo esta positiva ou negativa. Este sistema é disponibilizado como uma ferramenta *online* para consulta e uso público, permitindo que os usuários insiram uma ou mais mensagens e obtenham qual o sentimento associado.

Neste capítulo são oferecidas descrições detalhadas sobre as etapas de funcionamento do sistema proposto nas Seções 4.1 (seleção de modelo) e 4.2 (classificação), enquanto a Seção 4.3 descreve o sistema *online* disponível para o público.

4.1 Etapa de seleção do modelo

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

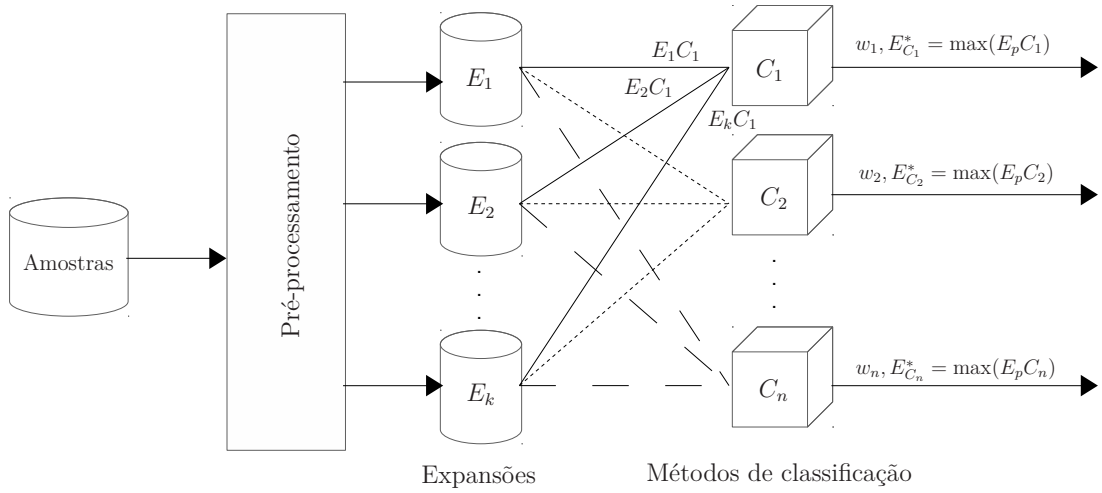
Na etapa de *seleção de modelo*, inicialmente, os parâmetros de cada método de classificação que compõem o sistema são ajustados através de uma busca em grade (do inglês, *grid search*). Este processo envolve combinar todas as variações de valores para cada parâmetro de um determinado método, a fim de encontrar a melhor configuração que maximiza o desempenho do método.

Como a busca em grade pode consumir bastante tempo, apenas um subconjunto estratificado e aleatório de amostras é utilizado nesta etapa. A partir disso, as amostras de entrada são normalizadas e expandidas com as técnicas de normalização e expansão de texto ($E_1, \ldots, E_k$). Todas as possíveis regras de combinação são utilizadas, sendo que cada uma delas produz um conjunto de saídas diferente, que são posteriormente usadas para treinar e avaliar os métodos de classificação ($C_1, \ldots, C_n$). Em seguida, o desempenho alcançado por cada possível combinação expansor-classificador ($E_p \rightarrow C_j$) é avaliado para definir qual regra de combinação é a mais adequada para cada método de classificação do sistema ($E^*_{c_j} = \max(E_p, C_j)\ \forall\ p \in \{1, \ldots, k\},\ j \in \{1, \ldots, n\}$). Finalmente, um peso w_j (grau de confiança) é calculado para cada combinação j, baseado na sua acurácia individual comparada com aquela que obteve melhor acurácia no sistema (Eq. 4.1). A

Figura 5 ilustra essa etapa.

$$w_j = \frac{1}{\left|\log_2\left(\frac{Acc_j}{max(Acc_j)+T}\right)\right|}, \quad 1 \leq j \leq n. \tag{4.1}$$

Figura 5 – Na etapa de seleção de modelo, o conjunto de dados original é processado pelas técnicas de normalização e expansão de texto $(E_1, \ldots, E_k)$. Em seguida, cada base resultante da expansão é usada no treinamento de cada método de classificação $(C_1, \ldots, C_n)$. Para cada método de classificação, é selecionada a melhor combinação expansor-classificador, $E^*_{c_j} = \max(E_p, C_j) \, \forall \, p \in \{1, \ldots, k\}$, $j \in \{1, \ldots, n\}$ e calculado um peso w_j, correspondente ao grau de confiança de tal combinação.

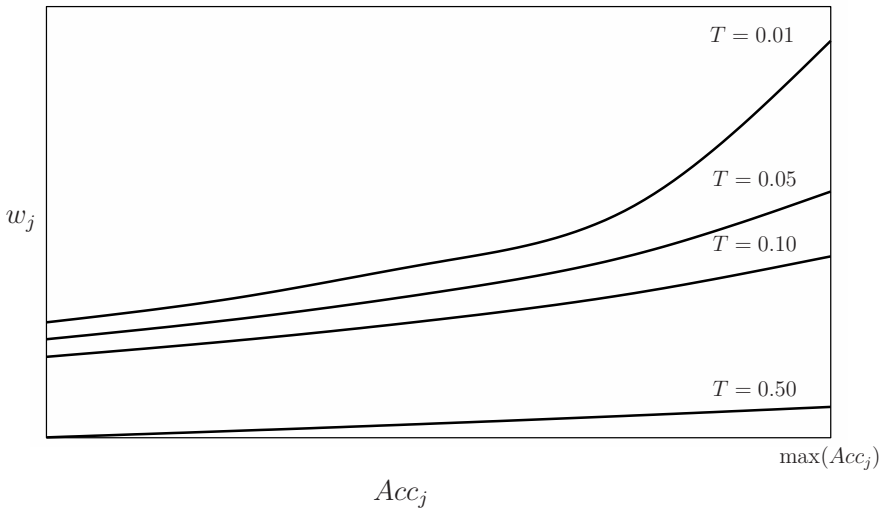


Na Equação 4.1, uma constante $0 < T \leq 1$ foi adicionada para controlar o balanceamento entre os maiores e menores pesos (w_j). Quanto menor o valor de T, maior é a diferença dos pesos entre os classificadores com melhor e pior desempenho, com base nas acurácias obtidas. Assim, conforme T tende a 0, maior é a diferença do peso do voto w_j do classificador com maior acurácia $(Acc_j = max(Acc_j))$ e todos os demais classificadores com $Acc_j < max(Acc_j)$. Por outro lado, conforme T tende a 1, menor será a diferença do peso de voto entre todos os classificadores. A Figura 6 ilustra como a escolha do valor de T pode influenciar na diferença entre os pesos dos votos dos classificadores com diferentes desempenhos obtidos na etapa de seleção de modelo.

4.2 Etapa de classificação

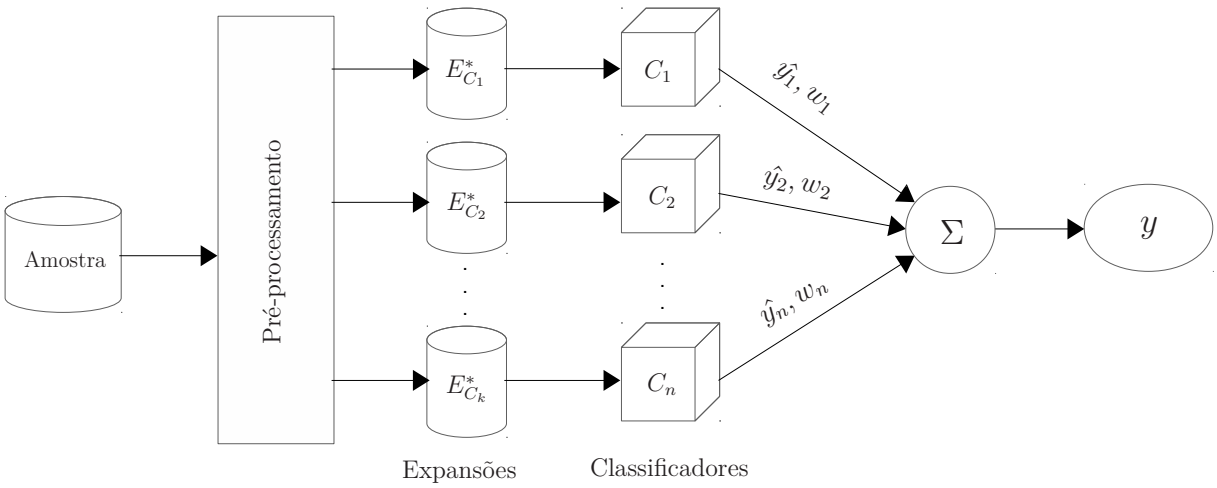
Finalizada a etapa de seleção, o melhor modelo é escolhido e os métodos de classificação são treinados. Em seguida, na *etapa de classificação*, as amostras são pré-processadas pelas técnicas de normalização e indexação semântica selecionadas na etapa anterior. As saídas de cada uma dessas técnicas são mescladas de acordo com a regra de combinação mais adequada para cada método de classificação, gerando a amostra de entrada que será

Figura 6 – A constante T influencia na diferença entre os pesos dos votos dos classificadores com desempenhos diferentes na etapa de seleção de modelo. Quanto menor o valor de T, maior será a diferença de peso entre os classificadores com maior e menor acurácia.



processada por cada classificador que, por sua vez, emite uma predição com um certo grau de confiança (w). O rótulo final é então computado pelo voto majoritário ponderado, conforme ilustrado na Figura 7.

Figura 7 – Após a etapa de seleção de modelo, o sistema associa quais técnicas de normalização e expansão de texto (e regra de combinação) (E_p^*) são mais adequadas para cada método de classificação (C_j), e o treinamento é então realizado. Em seguida, na etapa de classificação, dada uma amostra de entrada, ela é pré-processada e classificada por cada modelo que envia ao concentrador (Σ) sua predição ($\hat{y}_j$) com um grau de confiança (w_j). O rótulo final é então computado com base no voto majoritário ponderado.



Em resumo, neste trabalho foi projetado um comitê de máquinas de classificação adequado para manipular mensagens de texto curtas e ruidosas, como as comumente

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

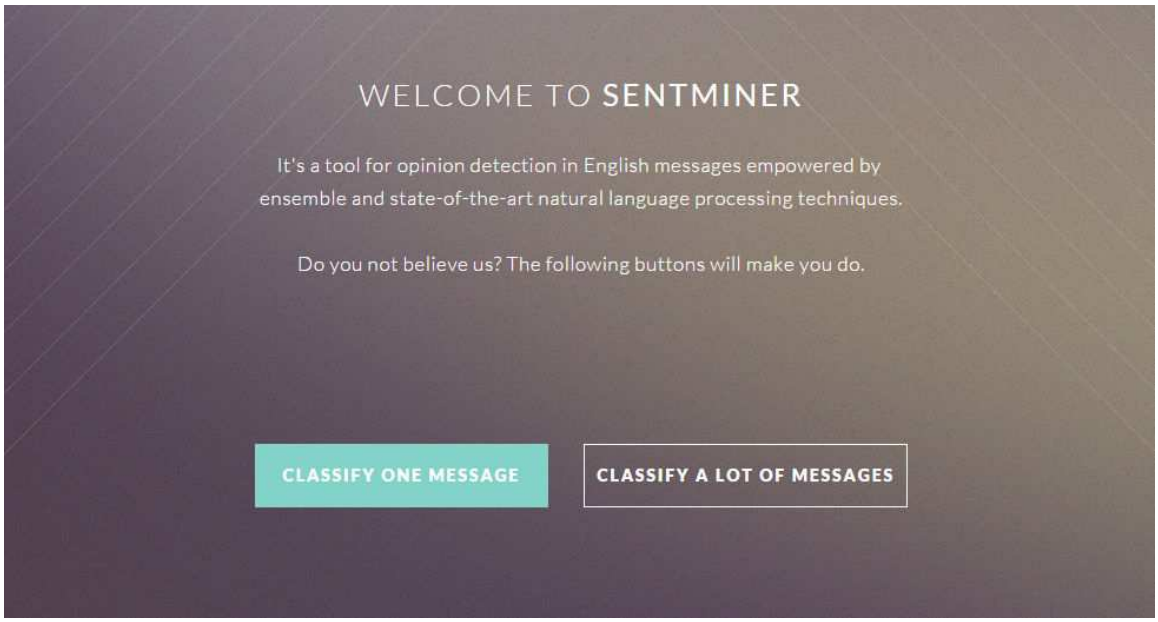
4.3 Ferramenta *online*

O sistema Sentminer está disponível dentro da suíte de ferramentas *ML-Tools*, em <http://lasid.sor.ufscar.br/ml-tools/>, junto com outras potenciais ferramentas para trabalhar com mensagens curtas e analisar mensagens que podem ser extraídas do Twitter.

Conforme descrito previamente, o intuito deste sistema é permitir que o público seja capaz de inserir mensagens e obter respostas do sistema com relação ao sentimento associado à informação inserida pelo usuário. Embora o sistema seja especialista em mensagens curtas do Twitter, a ferramenta é genérica o suficiente para prover resultados satisfatórios para mensagens extraídas de outras redes sociais e com tamanhos diferentes.

A tela inicial do sistema é apresentada na Figura 8 e traz as duas principais opções de funcionamento ao usuário. Estas opções permitem classificar uma mensagem individual ou um lote (*batch*) de mensagens.

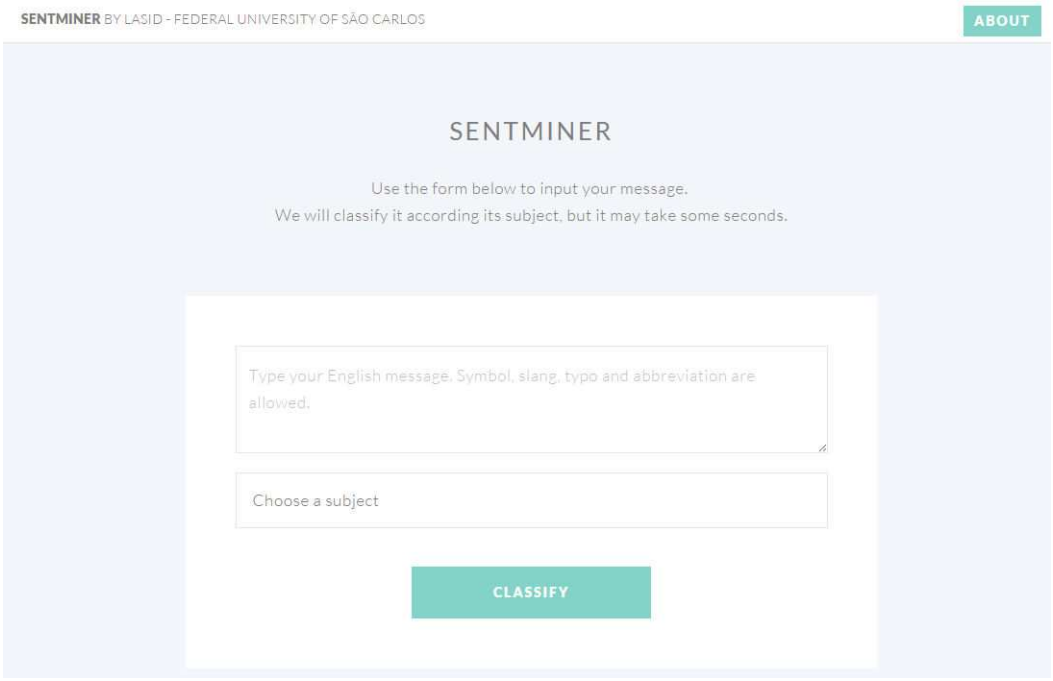
Figura 8 – O sistema abrange duas opções principais: classificar uma mensagem ou um lote (*batch*) de mensagens, oferecidas ao usuário logo no começo da experiência de navegação pelo *site*.



Quando o usuário escolhe a opção de classificar uma única mensagem, o sistema o encaminha para a tela da Figura 9. Nesta tela, o usuário precisa informar qual é a

mensagem a ser classificada e qual o domínio, ou contexto, ao qual aquela mensagem está relacionada. Para a opção de classificar mensagens em lote, o usuário deve enviar um arquivo texto contendo uma mensagem por linha, além de indicar a qual domínio tais mensagens estão associadas, conforme a Figura 10.

Figura 9 – Ao classificar uma mensagem, o usuário deve informar qual é a mensagem a ser classificada e qual o domínio, ou contexto, ao qual aquela mensagem está relacionada.



Após o sistema processar a informação, ou seja, normalizar e expandir a amostra inserida pelo usuário com todas as regras de combinação, as novas amostras são submetidas a um modelo já treinado, o qual define o sentimento associado.

A resposta para a classificação de uma mensagem ou lote de mensagens é exibida para o usuário conforme ilustram as Figuras 11 e 12, respectivamente. Na primeira figura, a mensagem inserida é exibida juntamente com a intensidade do sentimento associado. Quanto mais próxima das extremidades, mais forte é o sentimento negativo (à esquerda) ou positivo (à direita). Quando o indicador permanece no centro, o sentimento associado pode ser considerado como neutro. Na segunda imagem, é apresentado um sumário de quantas mensagens do lote estão em cada classe, a intensidade geral de sentimento associado ao lote e um *link* para que o usuário possa fazer *download* de um arquivo com todos os rótulos gerados pelo sistema.

Figura 10 – Ao classificar um lote de mensagens, o usuário deve enviar um arquivo texto contendo uma mensagem por linha, além de indicar a qual domínio tais mensagens estão associadas.

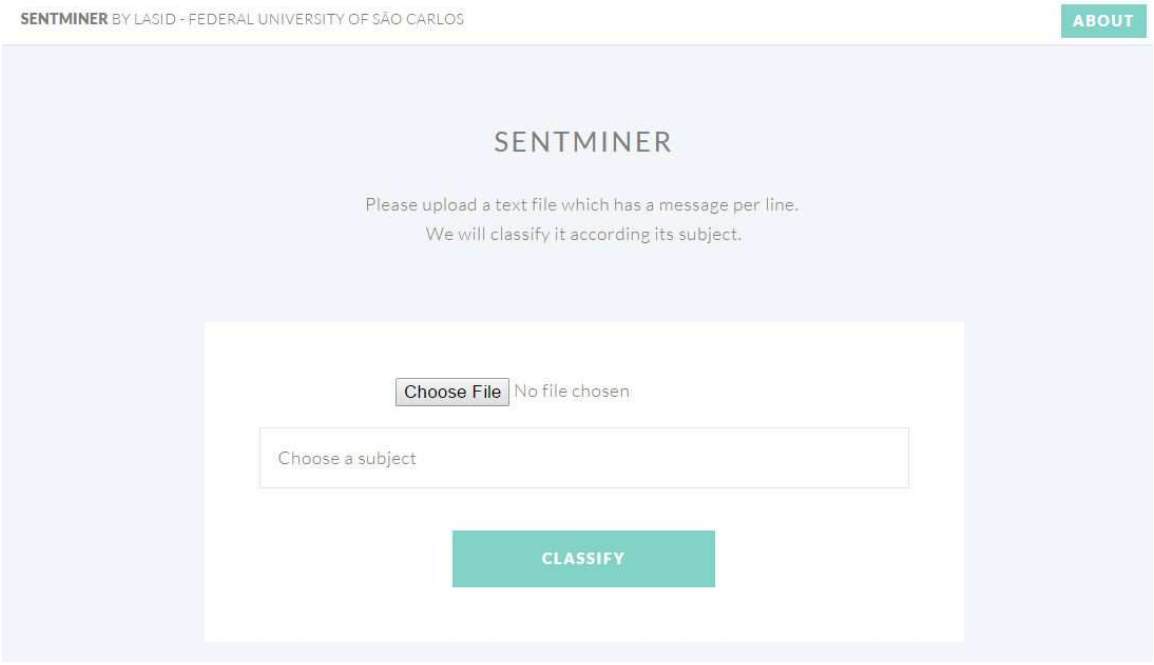


Figura 11 – A mensagem inserida é apresentada junto com a intensidade do sentimento associado. Quanto mais próximo do extremo esquerdo, mais negativo é o sentimento. Mais próximo à direita, mais positivo.

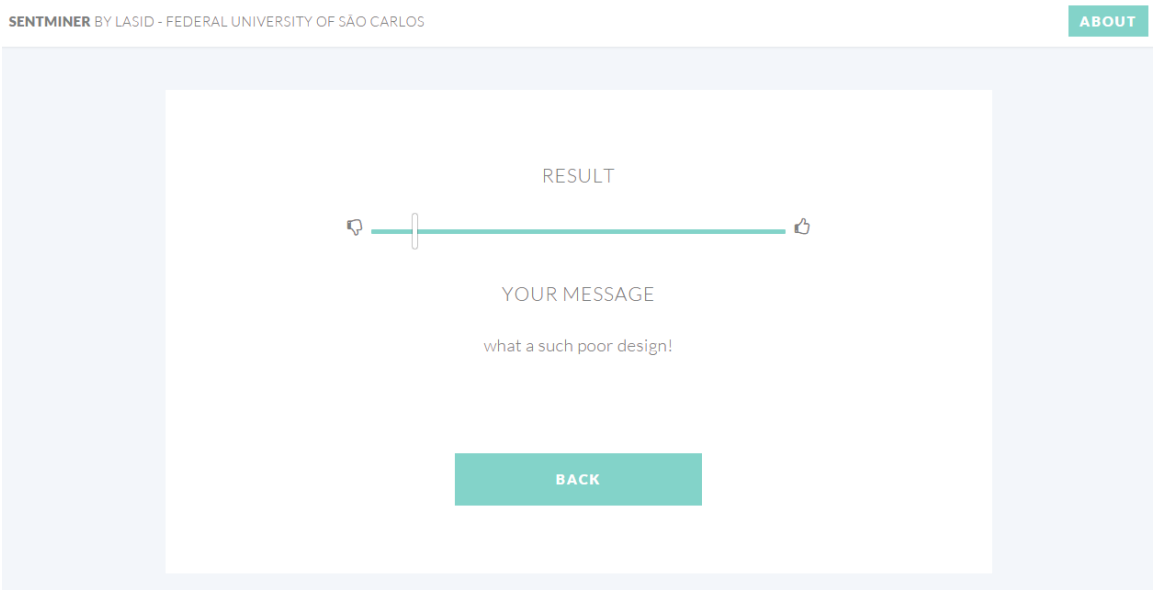
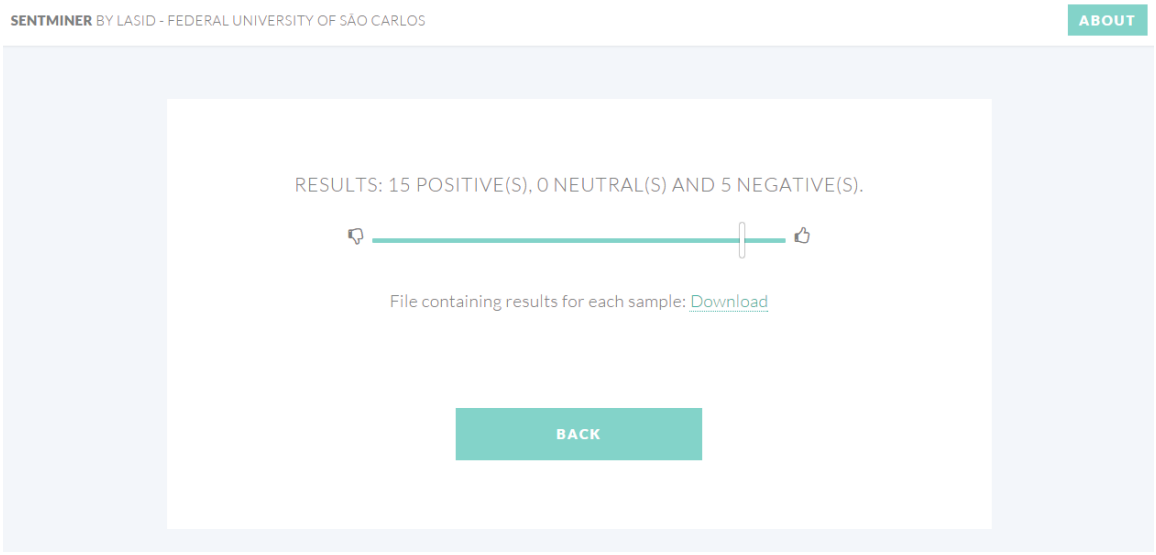


Figura 12 – O resultado da classificação de um lote de mensagens apresenta um sumário de quantas mensagens do lote estão em cada classe de sentimento, a intensidade geral de sentimento associado ao lote e um *link* para que o usuário possa fazer *download* de um arquivo com todos os rótulos gerados pelo sistema.



5 Experimentos e resultados

Neste capítulo são oferecidos detalhes sobre os experimentos realizados na validação do sistema proposto, os resultados obtidos e a discussão destes. A metodologia empregada nos diferentes experimentos é detalhada na Seção 5.1. Os resultados e a discussão de cada experimento são oferecidos na Seção 5.2.

5.1 Metodologia

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

5.1.1 Bases de dados e representação

Foram empregadas nove bases de dados reais, públicas e não-codificadas. Na Tabela 5, é apresentada a quantidade de amostras em cada classe, bem como o tema relacionado ao objeto de interesse de cada base. Cada amostra foi rotulada como positiva ou negativa, de acordo com a opinião que a mensagem expressa com relação ao objeto de interesse.

Tabela 5 – Bases de dados usadas na avaliação do sistema proposto.

Base de dados	# Positivas	# Negativas	Tema
STS-Test	181	177	Geral
HCR	537	886	Medicina
OMD	709	1195	Política
SS-Tweet	1252	1037	Geral
Sanders	519	572	Geral
UMICH	796	669	Filme
iPhone6	371	161	*Smartphone*
Archeage	724	994	Jogo
Hobbit3	354	168	Filme

As seis primeiras bases de dados foram propostas e utilizadas em trabalhos anteriores, STS-Test em (AGARWAL et al., 2011), HCR em (SPERIOSU et al., 2011), OMD em (SHAMMA; KENNEDY; CHURCHILL, 2009), SS-Tweet em (THELWALL; BUCKLEY; PALTOGLOU, 2012), Sanders em (ANALYTICS, 2011) e UMICH em (UMICH, 2011). Para a validação deste trabalho, estas bases foram pré-processadas de acordo com os procedimentos descritos em (SAIF et al., 2013).

As mensagens das últimas três bases de dados listadas na Tabela 5 foram criadas a partir de mensagens coletadas do Twitter e rotuladas pelo grupo de pesquisa do

Laboratório de Sistemas Inteligentes e Distribuídos (LaSID) da Universidade Federal de São Carlos, campus Sorocaba. Esta etapa foi realizada no segundo semestre de 2014, utilizando uma ferramenta de rotulação colaborativa, desenvolvida para essa finalidade, chamada *Labeling*[1]. Todas as bases criadas estão publicamente disponíveis em <http://dcomp.sor.ufscar.br/talmeida/sentcollection/>.

As amostras foram representadas computacionalmente por unigramas gerados a partir do conteúdo das mensagens. Os símbolos comumente utilizados no Twitter foram preservados, tais como números e sinalizadores de *hashtags* e *retweets*. Além disso, dois novos atributos foram criados para indicar o número de termos positivos e negativos presentes na mensagem. Para isso, cada termo da amostra foi procurado em um dicionário léxico[2], e se tal termo estiver associado a um valor positivo, o atributo referente ao contador de termos positivos é incrementado em um. Por outro lado, caso o termo esteja associado a um valor negativo, então o contador de termos negativos é incrementado em um. Segundo Mostafa (MOSTAFA, 2013) e Nastase e Strube (NASTASE; STRUBE, 2013), abordagens semelhantes apresentam bons resultados em tarefas de detecção de opinião.

5.1.2 Métodos de classificação

O sistema proposto é formado por métodos de classificação (Tabela 6) apresentados no Capítulo 3. Tais métodos utilizam representação e estratégia de seleção de hipótese diferentes, tais como probabilidade, otimização, distância e árvores. Com isso, é esperado que o comitê de máquinas de classificação seja flexível e capaz de gerar hipóteses genéricas e robustas, já que o próprio comitê encontra de forma automática a melhor estratégia e modelo para cada base de dados.

Tabela 6 – Métodos de classificação empregados no sistema proposto.

Métodos de classificação
Naïve Bayes Bernoulli (NB-B)
Naïve Bayes Multinomial (NB-M)
Naïve Bayes Gaussiano (NB-G)
Máquinas de vetores de suporte (SVM)
Árvores de decisão (C4.5)
N-vizinhos próximos (k-NN)
Boosted C4.5 (B.C4.5)
Regressão logística (RL)

[1] *Labeling*. Disponível em <http://lasid.sor.ufscar.br/ml-tools/>. Acessado em 22/10/2015.
[2] *Opinion Lexicon English*. Disponível em <http://goo.gl/czIfkd>. Acessado em 22/10/2015.

5.1.3 Avaliação do sistema

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

Na seleção do modelo, a base de dados de entrada foi processada e expandida pelas possíveis técnicas de processamento de texto, normalização e indexação semântica, resultando em um conjunto de bases expandidas (uma para cada possível regra de combinação). Em seguida, para cada base expandida, foram ajustados os valores dos parâmetros dos métodos de classificação através de *grid search*.

Neste trabalho, todos os experimentos e processo de validação foram implementados com a suíte `Scikit-learn` (PEDREGOSA et al., 2011) para Python. Os métodos de classificação e os parâmetros configurados na fase de *grid search* são descritos na Tabela 7. Os nomes dos parâmetros foram mantidos conforme descritos na documentação do `Scikit-learn`.

Tabela 7 – Lista de parâmetros configurados na fase de *grid search*.

Método	Parâmetro	Intervalo
NB-M	*alpha*	$10^{-7} \ldots 10^{7}$, passo 1 na potência
NB-B	*alpha*	$10^{-7} \ldots 10^{7}$, passo 1 na potência
NB-B	*binarize*	$10^{-3} \ldots 10^{3}$, passo 1 na potência
SVM-L	C	$10^{-3} \ldots 10^{3}$, passo 1 na potência
SVM-R	C	$10^{-3} \ldots 10^{3}$, passo 1 na potência
SVM-R	*gamma*	$10^{-3} \ldots 10^{3}$, passo 1 na potência
C4.5	*max features*	sqrt, log_2
k-NN	K	$1 \ldots 20$, passo 1
B.C4.5	*max features*	sqrt, log_2
B.C4.5	*estimators*	[50,100,250,500]
Regressão Logística	C	$10^{-3} \ldots 10^{3}$, passo 1 na potência

A seleção dos melhores parâmetros foi feita com base na F-medida, usando validação cruzada com 5 partições. Posteriormente, após os parâmetros terem sido ajustados, o sistema determinou qual base expandida obteve o melhor desempenho para cada método de classificação. Em outras palavras, nesta etapa foi selecionada a regra de combinação mais adequada para cada método de classificação do sistema. Finalmente, a constante T, usada para ajustar a diferença entre os pesos dos votos dos classificadores (Eq. 4.1), foi empiricamente definida com valor igual a 0,05.

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

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

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

A métrica F-medida (F_1) representa a média harmônica das medidas precisão (P) e revocação (R) encontradas para um determinado conjunto de dados, sendo expressa por

$$F_1 = 2 \cdot \frac{\mathrm{P} \cdot \mathrm{R}}{\mathrm{P} + \mathrm{R}}.$$

5.1.4 Cenários

Neste trabalho foram analisados quatro cenários distintos seguindo a metodologia descrita. Tais cenários são compostos pelas opções de utilizar ou não tanto seleção de atributo quanto a expansão das amostras, listados conforme:

Experimento 1 Usar amostras originais sem aplicar seleção de atributos;

Experimento 2 Usar amostras originais com seleção de atributos;

Experimento 3 Usar amostras expandidas sem aplicar seleção de atributos; e

Experimento 4 Usar amostras expandidas com seleção de atributos.

5.1.5 Seleção de atributos

A seleção de atributos foi realizada através de um processo incremental. Primeiramente, o teste estatístico χ^2 é aplicado sobre as amostras de treino para encontrar uma lista em ordem decrescente de quais atributos são mais relevantes. Em seguida, o conjunto de treino é reamostrado e, então, submetido ao método *Naïve* Bayes Multinomial, o qual foi escolhido pelo baixo custo computacional e bom desempenho. A reamostragem do conjunto de treino é feita de 5% em 5% até que todos os atributos (100%) tenham sido testados. Ao final, é selecionada a quantidade de atributos que apresentou melhor desempenho e os demais atributos são descartados.

Os resultados obtidos no experimento realizado com amostras expandidas e aplicação de seleção de atributos foram os mais promissores, portanto são demonstrados e discutidos com maiores detalhes na seção seguinte. Os resultados obtidos em todos os experimentos são apresentados no Anexo A.

5.2 Resultados

Os resultados discutidos nesta Seção foram obtidos ao seguir criteriosamente a metodologia descrita na seção anterior, através do cenário com amostras expandidas e seleção de atributos. Portanto, são oferecidas discussões acerca da diferença de desempenho entre métodos individuais e o sistema proposto com comitê de máquinas de classificação, embasadas por análise estatística.

A Tabela 8 apresenta os resultados obtidos pelos métodos de classificação para cada base de dados. O desempenho de cada técnica foi avaliado pela média e desvio padrão da F-medida obtida nos dez testes executados. Para cada base, os melhores resultados estão destacados em negrito.

Sob as mesmas condições, é evidente que o sistema de comitê de máquinas de classificação proposto obteve desempenho superior a qualquer um dos métodos de classificação individuais avaliados. No entanto, fica evidente que algumas bases de dados oferecem mais dificuldades aos métodos de classificação do que outras. A base SS-Tweet é uma base de assuntos mistos, o que dificulta a generalização dos métodos, ocasionando resultados de F-medida baixos. Por outro lado, a base UMICH é composta por amostras muito claras quanto ao sentimento expresso, resultando em desempenho alto.

Para garantir que os resultados não foram obtidos por acaso, foi realizado o teste estatístico não-paramétrico de Friedman (FRIEDMAN, 1940), seguindo a metodologia descrita em (JAPKOWICZ; SHAH, 2011).

O teste de Friedman avalia se a hipótese nula, que neste caso afirma não haver diferenças entre os resultados obtidos, pode ser rejeitada com base no ranqueamento computado através do desempenho obtido por cada método de classificação para cada base de dados. Assim, os métodos de classificação foram ranqueados de acordo com a F-medida e são apresentados na Tabela 9. O método com melhor desempenho recebe a posição 1 de *ranking*, enquanto o método com pior desempenho recebe a posição i, onde i é o número total de métodos.

Seja k a quantidade de métodos de classificação, n o número de bases de dados e R a soma dos *rankings* de cada método, a medida χ_F^2 foi calculada de acordo com a

Tabela 8 – Média e desvio padrão da F-medida obtidos para cada método avaliado.

Método	Archeage	HCR	Hobbit	IPhone6	OMD	Sanders	SS-Tweet	STS-Test	UMICH
B. C4.5	0,785±0,04	0,685±0,04	0,881±0,04	0,677±0,06	0,743±0,03	0,703±0,03	0,562±0,02	0,807±0,06	0,934±0,02
C4.5	0,753±0,02	0,643±0,03	0,864±0,02	0,681±0,05	0,682±0,04	0,640±0,04	0,543±0,03	0,752±0,06	0,920±0,02
Comitê	**0,869±0,02**	**0,733±0,03**	**0,921±0,02**	**0,738±0,04**	**0,811±0,02**	**0,756±0,03**	**0,612±0,02**	**0,863±0,03**	**0,969±0,01**
KNN	0,723±0,03	0,633±0,03	0,844±0,04	0,651±0,07	0,728±0,05	0,675±0,04	0,525±0,06	0,776±0,06	0,953±0,02
Logistic	0,836±0,03	0,679±0,03	0,908±0,03	0,698±0,05	0,775±0,02	0,711±0,03	0,585±0,02	0,804±0,03	0,962±0,02
NB B.	0,856±0,02	0,695±0,03	0,869±0,05	0,735±0,03	0,781±0,02	0,723±0,05	0,598±0,02	0,817±0,05	0,948±0,01
NB G.	0,793±0,02	0,590±0,02	0,772±0,08	0,689±0,04	0,628±0,04	0,651±0,02	0,530±0,03	0,751±0,04	0,760±0,06
NB M.	0,843±0,02	0,683±0,03	0,882±0,03	0,699±0,05	0,781±0,04	0,727±0,03	0,593±0,04	0,821±0,05	0,956±0,01
SVM-L	0,842±0,03	0,686±0,03	0,913±0,03	0,708±0,04	0,771±0,01	0,692±0,04	0,583±0,03	0,813±0,03	0,965±0,01
SVM-R	0,820±0,02	0,695±0,02	0,856±0,07	0,651±0,07	0,736±0,05	0,673±0,09	0,539±0,09	0,756±0,07	0,953±0,02

Tabela 9 – Ranqueamento de F-medida obtida pelos métodos para cada base de dados.

Bases	B. C4.5	C4.5	Comitê	k-NN	RL	NB-B	NB-G	NB-M	SVM-L	SVM-R
Archeage	8	9	1	10	5	2	7	3	4	6
HCR	5	8	1	9	7	2	10	6	4	3
Hobbit	5	7	1	9	3	6	10	4	2	8
Iphone6	8	7	1	10	5	2	6	4	3	9
OMD	6	9	1	8	4	3	10	2	5	7
Sanders	5	10	1	7	4	3	9	2	6	8
SS-Tweet	6	7	1	10	4	2	9	3	5	8
STS-Test	5	9	1	7	6	3	10	2	4	8
UMICH	8	9	1	6	3	7	10	4	2	5
Soma	56	75	9	76	41	30	81	30	35	62

Equação 5.1.

$$\chi_F^2 = \left[\frac{12}{nk(k+1)} \sum_{j=1}^{k} R_j^2 \right] - 3n(k+1). \tag{5.1}$$

A hipótese nula pode ser rejeitada se χ_F^2 estiver em uma distribuição χ^2, com $k-1$ graus de liberdade para $n > 15$ ou $k > 5$. Para valores menores de n e k, a medida χ^2 aproximada é imprecisa, sendo necessário procurar valores em tabelas de χ_F^2 específicas para o teste de Friedman (JAPKOWICZ; SHAH, 2011). Assim, considerando um intervalo de confiança $\alpha = 0,001$, $n = 9$ e $k = 9$, o valor crítico é dado por 27,877. Portanto, como $\chi_F^2 = 63,3515$, conclui-se que há uma diferença significativa entre os desempenhos obtidos pelos métodos de classificação avaliados e, portanto, a hipótese nula pode ser rejeitada.

Em seguida, o teste de Nemenyi (NEMENYI, 1963) foi empregado para comparar o desempenho dos métodos em pares. Nesse cenário, para cada par de métodos, um valor q foi calculado com base na distribuição do *ranking* de desempenho dos métodos de classificação avaliados (Eq. 5.2). Assim, a hipótese nula, que indica não haver diferença no desempenho entre os dois métodos, pode ser rejeitada se q for maior que o valor crítico $q\alpha$, obtido para o grau de confiança α desejado.

$$q = \frac{\overline{R_{j1}} - \overline{R_{j2}}}{\sqrt{\frac{k(k+1)}{6n}}}. \tag{5.2}$$

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

Outra análise importante para o contexto deste trabalho se refere ao número de vezes que uma regra de combinação foi escolhida por um método de classificação. Mais importante, se as regras de combinação propostas são mais úteis para os métodos de classificação que os dados originais sem expansão (regra YNNN).

Os experimentos foram repetidos dez vezes com 9 métodos para 9 bases, totalizando 810 escolhas de melhor regra de combinação para cada método de classificação na etapa de seleção de modelo. Todas essas escolhas são apresentadas na Tabela 10, somando os valores individuais para todas as bases. Os valores destacados indicam a regra de combinação mais utilizada para um determinado método de classificação.

Tabela 10 – Quantidade de vezes que cada regra de combinação foi escolhida pelos métodos de classificação.

Regra	B. C4.5	C4.5	k-NN	RL	NB-B	NB-G	NB-M	SVM-L	SVM-R	Total
E_1	**59**	**60**	**43**	**24**	**21**	6	12	**18**	15	**258**
E_2	20	20	14	5	13	6	9	8	4	99
E_3	10	9	9	21	2	6	3	**18**	**30**	108
E_4	0	0	4	0	3	3	4	0	1	15
E_5	0	0	0	0	2	5	6	1	1	15
E_6	0	0	5	14	0	6	7	14	17	63
E_7	1	1	5	2	13	14	**23**	2	1	62
E_8	0	0	2	7	2	**19**	2	10	5	47
E_9	0	0	3	1	14	5	9	1	0	33
E_{10}	0	0	2	2	17	5	12	1	0	39
E_{11}	0	0	3	14	3	15	3	17	16	71

Em 31,8% dos casos, os dados originais apresentaram melhor resultado quando combinados a algum método de classificação. No entanto, os métodos com maior desempenho (a família de métodos *naïve* Bayes) utilizaram as outras regras de combinação, indicando que a expansão das amostras beneficiou o desempenho dessa classe de métodos.

Outra evidência importante que pode ser extraída desses dados é o fato de que as regras de combinação que envolvem "geração de conceitos" (E_3, E_6, E_8 e E_{11}) representaram 35,6% das escolhas realizadas, fortalecendo a hipótese apontada de que as mensagens originais (curtas e ruidosas) têm pouca informação e que é necessário utilizar técnicas de linguagem natural para enriquecê-las.

Finalmente, na Tabela 11 é apresentada a média do tempo de execução (em segundos) demandado para concluir as etapas seleção de modelo, treinamento e teste de cada método de classificação para cada base de dados. Os menores valores estão destacados em negrito. Tais experimentos foram realizados em um computador com processador Intel Xeon X3430 (quatro núcleos - 2,4GHz) e 8GB de RAM.

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

Tabela 11 – Média e desvio padrão do tempo de execução (em segundos) consumido nas etapas de seleção de modelo, treinamento e teste de cada método de classificação para cada base de dados.

Método	Archeage	HCR	Hobbit	IPhone6	OMD	Sanders	SS-Tweet	STS-Test	UMICH
B. C4.5	2,30±1,85	2,90±2,55	2,50±0,92	1,50±1,43	8,00±4,12	6,50±4,03	16,60±9,81	1,20±1,33	1,40±0,66
C4.5	0,70±0,46	0,40±0,49	0,40±0,49	0,20±0,40	0,40±0,49	0,90±0,30	1,40±0,49	0,20±0,40	**0,20±0,40**
Comitê	93,10±23,74	41,10±5,50	7,60±1,36	7,70±2,19	78,60±17,56	61,70±39,11	386,10±49,31	6,50±0,92	37,60±10,54
k-NN	8,20±2,27	2,70±1,10	0,80±0,60	0,70±0,46	7,90±0,94	4,90±1,87	29,40±7,49	0,70±0,46	3,90±1,37
Logistic	0,50±0,50	0,30±0,46	0,10±0,30	0,10±0,30	**0,30±0,46**	0,90±0,70	1,60±0,49	**0,00±0,00**	0,40±0,49
NB-B	5,40±1,62	2,50±0,50	1,40±0,49	1,40±0,49	3,90±0,94	4,20±1,99	13,00±1,79	1,50±0,50	3,10±1,04
NB-G	**0,40±0,49**	**0,10±0,30**	**0,00±0,00**	**0,00±0,00**	**0,30±0,46**	**0,30±0,46**	1,50±0,81	**0,00±0,00**	0,30±0,64
NB-M	1,00±0,63	0,40±0,49	0,20±0,40	0,60±0,49	0,60±0,49	0,70±0,90	2,50±0,92	0,20±0,40	0,80±0,40
SVM-L	0,70±0,64	0,50±0,50	**0,00±0,00**	0,40±0,49	0,70±0,46	0,50±0,50	**1,30±0,46**	0,20±0,40	0,60±0,49
SVM-R	24,20±6,03	10,60±1,96	0,60±0,66	0,90±0,70	21,90±4,41	14,80±8,32	110,40±16,17	0,80±0,60	10,10±2,98

Conclusões

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

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

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

O sistema proposto foi avaliado com nove bases de dados públicas e os desempenhos obtidos foram comparados com os métodos de classificação individuais. Os resultados da análise estatística indicaram que, sob as mesmas condições e com grau de confiança igual à 99,9%, o comitê de classificadores foi estatisticamente superior a qualquer outro método de classificação avaliado. Também foi constatado que, em 35,8% dos casos analisados, as amostras expandidas contribuíram mais para o desempenho dos métodos de classificação do que a utilização das amostras originais (curtas e ruidosas). Contudo, o comitê demanda maior tempo computacional durante a seleção de modelo e treinamento.

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

Publicações

Durante o período de curso de mestrado, os seguintes trabalhos foram produzidos em colaboração com outros pesquisadores ou de maneira independente:

Revistas

1. ALBERTO, T. C.; LOCHTER, J. V.; ALMEIDA, T. A.. "Post or Block? Advances in Automatically Filtering Undesired Comments." *Journal of Intelligent & Robotic Systems*, v. 1, p. 1-15, 2014.

2. LOCHTER, J. V.; ZANETTI, R. F.; ALMEIDA, T. A.. "Short Text Opinion Detection using Ensemble of Classifiers and Semantic Indexing." *Journal of Knowledge-Based Systems*, 2015. (Em avaliação)

Congressos

1. CAMPANHA, J. M.; LOCHTER, J. V.; ALMEIDA, T. A.. "Detecção Automática de Spammers em Redes Sociais." Anais do XI Encontro Nacional de Inteligência Artificial e Computacional (ENIAC'14), v. 1., p. 1-6, São Carlos - SP, Brasil. 2014.

2. ALBERTO, T. C.; LOCHTER, J. V.; ALMEIDA, T. A.. "TubeSpam: Comment Spam Filtering on YouTube" 15th International Conference on Machine Learning and Applications (ICMLA'15), Miami - FL, USA. 2015. (Aceito para publicação)

3. LOCHTER, J. V.; ZANETTI, R. F.; ALMEIDA, T. A.. "Detecção de Opinião em Mensagens Curtas usando Comitê de Classificadores e Indexação Semântica." Anais do XII Encontro Nacional de Inteligência Artificial e Computacional (ENIAC'15), Natal - RN, Brasil. 2015. (Aceito para publicação)

4. ALBERTO, T. C.; LOCHTER, J. V.; ALMEIDA, T. A.. "Filtragem Automática de Spam nos Comentários do YouTube" Anais do XII Encontro Nacional de Inteligência Artificial e Computacional (ENIAC'15), Natal - RN, Brasil. 2015. (Aceito para publicação)

Referências

AGARWAL, A. et al. Sentiment analysis of twitter data. In: *Proc. of 2011 LSM*. [S.l.]: Association for Computational Linguistics, 2011. (LSM '11), p. 30–38. Citado na página 63.

AGGARWAL, C. C.; ZHAI, C. X. A survey of text classification algorithms. In: *Mining Text Data*. [S.l.]: Springer US, 2012. p. 163–222. Citado 2 vezes nas páginas 38 e 49.

ALMEIDA, T.; ALMEIDA, J.; YAMAKAMI, A. Spam filtering: how the dimensionality reduction affects the accuracy of naive bayes classifiers. *Journal of Internet Services and Applications*, Springer-Verlag, v. 1, n. 3, p. 183–200, 2011. Citado na página 50.

ANALYTICS, S. *Dataset - Twitter Sentiment*. 2011. <http://www.sananalytics.com/lab/twitter-sentiment/>. [Online; acessado em 07/07/2015]. Citado na página 63.

BENEVENUTO, F. et al. Characterizing user navigation and interactions in online social networks. *Information Sciences*, v. 195, p. 1 – 24, 2012. ISSN 0020-0255. Citado na página 31.

BOSER, B. E.; GUYON, I. M.; VAPNIK, V. N. A training algorithm for optimal margin classifiers. In: *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. New York, NY, USA: ACM, 1992. (COLT '92), p. 144–152. Citado na página 49.

BREIMAN, L. Bagging predictors. *Mach. Learn.*, Kluwer Academic Publishers, Hingham, MA, USA, v. 24, n. 2, p. 123–140, ago. 1996. Citado na página 52.

BREIMAN, L. Random forests. *Machine Learning*, Kluwer Academic Publishers, Hingham, MA, USA, v. 45, n. 1, p. 5–32, out. 2001. Citado na página 50.

BREIMAN, L. et al. *Classification and Regression Trees*. Belmont, CA: Wadsworth International Group, 1984. Citado na página 49.

CINGEL, D. P.; SUNDAR, S. S. Texting, techspeak, and tweens: The relationship between text messaging and english grammar skills. *New Media and Society*, v. 14, n. 8, p. 1304–1320, 2012. Citado na página 34.

CORTES, C.; VAPNIK, V. Support-vector networks. *Machine Learning*, v. 20, n. 3, p. 273–297, 1995. Citado na página 49.

DENECKE, K. Using SentiWordNet for multilingual sentiment analysis. In: *Proceedings of the IEEE 24th International Conference on Data Engineering Workshop*. Cancun, Mexico: [s.n.], 2008. (ICDEW '08), p. 507–512. Citado 2 vezes nas páginas 25 e 33.

DIETTERICH, T. G. Ensemble methods in machine learning. In: *Proceedings of the First International Workshop on Multiple Classifier Systems*. Cagliari, Italy: [s.n.], 2000. (MCS '00), p. 1–15. Citado 2 vezes nas páginas 26 e 52.

FERNANDEZ-DELGADO, M. et al. Do we need hundreds of classifiers to solve real world classification problems? *J. Mach. Learn. Res.*, JMLR.org, v. 15, n. 1, p. 3133–3181, jan. 2014. Citado na página 47.

FISHER, R. A. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, v. 7, n. 7, p. 179–188, 1936. Citado na página 37.

FRAKES, W. B.; BAEZA-YATES, R. (Ed.). *Information Retrieval: Data Structures and Algorithms*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1992. Citado na página 38.

FRIEDMAN, M. A comparison of alternative tests of significance for the problem of m rankings. *Ann. Math. Statist.*, The Institute of Mathematical Statistics, v. 11, n. 1, p. 86–92, 03 1940. Citado na página 67.

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 na página 40.

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 na página 40.

GASPERIN, C.; LIMA, V. L. S. de. *Fundamentos do Processamento Estatístico de Linguagem Natural*. Porto Alegre, Rio Grande do Sul, Brazil, 2001. 1–59 p. Citado na página 39.

GO, A.; BHAYANI, R.; HUANG, L. *Twitter Sentiment Classification using Distant Supervision*. [S.l.], 2009. Citado na página 39.

GRAY, D. E. *Doing Research in the Real World*. 3rd. ed. [S.l.]: SAGE, 2014. Citado na página 30.

HALEVY, A.; NORVIG, P.; PEREIRA, F. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, IEEE Educational Activities Department, Piscataway, NJ, USA, v. 24, n. 2, p. 8–12, 3 2009. Citado na página 37.

HARRINGTON, P. *Machine Learning in Action*. Greenwich, CT, USA: Manning Publications Co., 2012. Citado na página 50.

HARRIS, Z. Distributional structure. *Word*, v. 10, n. 23, p. 146–162, 1954. Citado na página 38.

JAPKOWICZ, N.; SHAH, M. *Evaluating Learning Algorithms - A Classification Perspective*. [S.l.]: Cambridge University Press, 2011. Citado 2 vezes nas páginas 67 e 69.

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 26.

LANGLEY, P.; IBA AND, W.; THOMPSON, K. An analysis of bayesian classifiers. In: *Proceedings of the Tenth National Conference on Artificial Intelligence*. [S.l.]: AAAI Press, 1992. (AAAI'92), p. 223–228. Citado na página 50.

MCCALLUM, A.; NIGAM, K. A comparison of event models for naive Bayes text classification. In: *Learning for Text Categorization: Papers from the 1998 AAAI Workshop*. [S.l.: s.n.], 1998. p. 41–48. Citado na página 51.

MITCHELL, T. *Machine Learning*. 1st. ed. [S.l.]: Mc Graw Hill, 1997. Citado na página 37.

MOSQUERA, A.; MOREDA, P. Timproving web 2.0 opinion mining systems using text normalisation techniques. In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing*. Hissar, Bulgaria: [s.n.], 2013. (RANLP'13), p. 491–495. Citado na página 25.

MOSTAFA, M. M. More than words: Social networks' text mining for consumer brand sentiments. *Expert Systems with Applications*, v. 40, n. 10, p. 4241–4251, 2013. Citado 3 vezes nas páginas 25, 35 e 64.

NASTASE, V.; STRUBE, M. Transforming Wikipedia into a large scale multilingual concept network. *Artificial Intelligence*, v. 194, n. 1, p. 62–85, 2013. Citado 2 vezes nas páginas 26 e 64.

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 na página 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 3 vezes nas páginas 25, 35 e 43.

NEMENYI, P. F. *Distribution-free multiple comparisons*. Tese (Doutorado) — Princeton University, 1963. Citado na página 69.

PANG, B.; LEE, L. Opinion Mining and Sentiment Analysis. *Fundations and Trends in Information Retrieval*, v. 2, n. 1–2, p. 1–135, 2008. Citado 2 vezes nas páginas 25 e 35.

PANG, B.; LEE, L.; VAITHYANATHAN, S. Thumbs up? Sentiment classification using machine learning techniques. In: *Proceedings of the Association for Computational Linguistics 2002 conference on Empirical methods in natural language processing*. Philadelphia, USA: [s.n.], 2002. (ACL EMNLP '02), p. 79–86. Citado 2 vezes nas páginas 25 e 33.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. Citado na página 65.

RENNIE, J. D. et al. Tackling the poor assumptions of naive bayes text classifiers. In: FAWCETT, T.; MISHRA, N. (Ed.). *ICML*. [S.l.]: AAAI Press, 2003. p. 616–623. Citado na página 51.

RUSSELL, S. J.; NORVIG, P. *Artificial Intelligence: A Modern Approach*. 2. ed. [S.l.]: Pearson Education, 2003. Citado na página 47.

SAIF, H. et al. Evaluation datasets for twitter sentiment analysis: a survey and a new dataset, the sts-gold. In: *1st Interantional Workshop on Emotion and Sentiment in Social and Expressive Media: Approaches and Perspectives from AI (ESSEM 2013)*. [S.l.: s.n.], 2013. Citado na página 63.

SALTON, G.; MCGILL, M. J. *Introduction to Modern Information Retrieval*. New York, NY, USA: McGraw-Hill, Inc., 1986. Citado na página 48.

SCHAPIRE, R. E. A brief introduction to boosting. In: *Proceedings of the 16th International Joint Conference on Artificial Intelligence - Volume 2*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999. (IJCAI'99), p. 1401–1406. Citado na página 53.

SHAMMA, D. A.; KENNEDY, L.; CHURCHILL, E. F. Tweet the debates: Understanding community annotation of uncollected sources. In: *Proc. of 2009 WSM*. [S.l.]: ACM, 2009. (WSM '09), p. 3–10. Citado na página 63.

SILVA, T. P. et al. Normalização Textual e Indexação Semântica Aplicadas na Filtragem de SMS Spam. In: *Proc. of the 11st ENIAC*. São Carlos, Brazil: [s.n.], 2014. p. 1–6. Citado na página 41.

SPERIOSU, M. et al. Twitter polarity classification with label propagation over lexical links and the follower graph. In: *Proc. of 2011 EMNLP*. [S.l.]: Association for Computational Linguistics, 2011. (EMNLP '11), p. 53–63. Citado na página 63.

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 2 vezes nas páginas 25 e 35.

THELWALL, M.; BUCKLEY, K.; PALTOGLOU, G. Sentiment strength detection for the social web. *J. Am. Soc. Inf. Sci. Technol.*, John Wiley & Sons, Inc., v. 63, n. 1, p. 163–173, jan. 2012. Citado na página 63.

UMICH. *Dataset SI650 - Sentiment Classification*. 2011. <https://goo.gl/Xfr8lI>. [Online; acessado em 07/07/2015]. Citado na página 63.

WALKER, S. H.; DUNCAN, D. B. Estimation of the probability of an event as a function of several independent variables. *Biometrika*, v. 54, n. 1, p. 167–179, jun. 1967. Citado na página 48.

WANG, G. et al. Sentiment classification: The contribution of ensemble learning. *Decision Support Systems*, v. 57, p. 77 – 93, 2014. Citado na página 52.

WU, X. et al. Top 10 algorithms in data mining. *KAIS*, v. 14, n. 1, p. 1–37, 2008. Citado na página 47.

XIA, R.; ZONG, C.; LI, S. Ensemble of feature sets and classification algorithms for sentiment classification. *Information Sciences*, v. 181, n. 6, p. 1138 – 1152, 2011. Citado na página 52.

APÊNDICE A – Resultados

Neste trabalho foram analisados quatro cenários distintos, utilizando as amostras originais, amostras expandidas, todos os atributos e atributos selecionados. A descrição de cada experimento é dada a seguir.

Experimento 1 Usar amostras originais sem aplicar seleção de atributos;

Experimento 2 Usar amostras originais com seleção de atributos;

Experimento 3 Usar amostras expandidas sem aplicar seleção de atributos; e

Experimento 4 Usar amostras expandidas com seleção de atributos.

Os melhores resultados foram obtidos através do Experimento 4 e são discutidos em detalhes no Capítulo 5. O restante deste apêndice apresenta os resultados de cada experimento para eventual consulta.

Tabela 12 – Média e desvio padrão da F-medida obtidos para cada método avaliado no Experimento 1.

Método	Archeage	HCR	Hobbit	IPhone6	OMD	Sanders	SS-Tweet	STS-Test	UMICH
B. C4.5	0,747±0,02	0,643±0,05	0,832±0,04	0,692±0,07	0,692±0,03	0,655±0,04	0,565±0,02	0,741±0,04	0,919±0,03
C4.5	0,746±0,02	0,627±0,03	0,838±0,07	0,687±0,06	0,707±0,03	0,651±0,04	0,573±0,02	0,736±0,05	0,907±0,02
Comitê	**0,870±0,02**	**0,744±0,02**	**0,921±0,02**	**0,744±0,05**	**0,833±0,01**	**0,758±0,03**	**0,618±0,02**	**0,867±0,04**	**0,966±0,01**
KNN	0,688±0,03	0,636±0,04	0,830±0,04	0,645±0,09	0,678±0,04	0,659±0,05	0,590±0,04	0,794±0,04	0,950±0,01
Logistic	0,855±0,02	0,733±0,02	0,910±0,02	0,681±0,10	0,822±0,01	0,738±0,03	0,577±0,02	0,839±0,04	0,958±0,01
NB B.	0,862±0,02	0,710±0,02	0,889±0,04	0,608±0,12	0,787±0,03	0,703±0,12	0,449±0,10	0,796±0,10	0,953±0,01
NB G.	0,798±0,02	0,625±0,02	0,630±0,05	0,612±0,07	0,684±0,03	0,655±0,03	0,521±0,02	0,765±0,06	0,785±0,07
NB M.	0,850±0,02	0,704±0,02	0,865±0,04	0,637±0,08	0,800±0,03	0,713±0,07	0,503±0,07	0,854±0,05	0,941±0,03
SVM-L	0,846±0,03	0,722±0,02	0,906±0,02	0,727±0,05	0,820±0,01	0,739±0,03	0,614±0,02	0,832±0,04	0,961±0,01
SVM-R	0,849±0,02	0,731±0,03	0,909±0,02	0,705±0,10	0,826±0,01	0,740±0,04	0,552±0,10	0,845±0,05	0,963±0,01

Tabela 13 – Média e desvio padrão da F-medida obtidos para cada método avaliado no Experimento 2.

Método	Archeage	HCR	Hobbit	IPhone6	OMD	Sanders	SS-Tweet	STS-Test	UMICH
B. C4.5	0,794±0,04	0,684±0,03	0,889±0,03	0,676±0,05	0,781±0,02	0,710±0,03	0,563±0,03	0,800±0,06	0,927±0,03
C4.5	0,753±0,02	0,643±0,03	0,878±0,02	0,676±0,05	0,707±0,02	0,646±0,05	0,549±0,04	0,762±0,05	0,920±0,02
Comitê	**0,862±0,02**	**0,737±0,02**	**0,928±0,02**	**0,729±0,04**	**0,830±0,02**	**0,761±0,04**	**0,622±0,02**	**0,881±0,03**	**0,969±0,01**
KNN	0,688±0,05	0,632±0,03	0,855±0,04	0,655±0,04	0,754±0,04	0,679±0,03	0,521±0,04	0,799±0,03	0,953±0,02
Logistic	0,837±0,03	0,713±0,03	0,911±0,02	0,665±0,10	0,807±0,03	0,735±0,04	0,594±0,03	0,840±0,04	0,962±0,01
NB B.	0,856±0,02	0,723±0,03	0,909±0,04	0,712±0,04	0,793±0,02	0,735±0,04	0,608±0,03	0,852±0,05	0,949±0,01
NB G.	0,790±0,04	0,595±0,03	0,866±0,04	0,691±0,04	0,662±0,06	0,651±0,03	0,522±0,02	0,769±0,04	0,782±0,04
NB M.	0,840±0,02	0,728±0,03	0,910±0,02	0,703±0,04	0,801±0,03	0,744±0,03	0,572±0,05	0,833±0,05	0,955±0,02
SVM-L	0,841±0,03	0,704±0,04	0,911±0,02	0,674±0,07	0,818±0,02	0,722±0,06	0,605±0,02	0,836±0,05	0,964±0,01
SVM-R	0,840±0,02	0,721±0,03	0,906±0,02	0,691±0,03	0,819±0,02	0,741±0,05	0,584±0,08	0,840±0,05	0,963±0,01

Tabela 14 – Média e desvio padrão da F-medida obtidos para cada método avaliado no Experimento 3.

Método	Archeage	HCR	Hobbit	IPhone6	OMD	Sanders	SS-Tweet	STS-Test	UMICH
B. C4.5	0,758±0,03	0,656±0,05	0,825±0,04	0,686±0,07	0,703±0,02	0,655±0,04	0,573±0,03	0,680±0,08	0,925±0,03
C4.5	0,746±0,02	0,627±0,03	0,838±0,07	0,687±0,06	0,707±0,03	0,651±0,04	0,573±0,02	0,736±0,05	0,907±0,02
Comitê	**0,871±0,02**	**0,746±0,02**	**0,924±0,02**	**0,755±0,05**	**0,826±0,02**	**0,751±0,03**	**0,612±0,02**	**0,870±0,04**	**0,968±0,01**
KNN	0,700±0,03	0,642±0,04	0,828±0,04	0,650±0,10	0,680±0,04	0,667±0,05	0,591±0,03	0,788±0,06	0,948±0,01
Logistic	0,855±0,02	0,733±0,02	0,914±0,02	0,715±0,06	0,806±0,03	0,706±0,03	0,603±0,02	0,807±0,04	0,959±0,01
NB B.	0,860±0,02	0,716±0,02	0,862±0,04	0,646±0,11	0,787±0,03	0,688±0,12	0,516±0,05	0,821±0,04	0,951±0,01
NB G.	0,798±0,02	0,629±0,02	0,592±0,07	0,612±0,07	0,684±0,03	0,634±0,03	0,517±0,03	0,765±0,06	0,778±0,06
NB M.	0,848±0,02	0,705±0,02	0,862±0,05	0,696±0,06	0,790±0,04	0,710±0,05	0,545±0,06	0,833±0,06	0,941±0,03
SVM-L	0,846±0,03	0,722±0,02	0,913±0,02	0,733±0,05	0,802±0,02	0,705±0,04	0,593±0,02	0,818±0,03	0,962±0,01
SVM-R	0,842±0,02	0,732±0,03	0,909±0,02	0,692±0,11	0,781±0,03	0,698±0,04	0,515±0,10	0,845±0,05	0,948±0,03

Tabela 15 – Média e desvio padrão da F-medida obtidos para cada método avaliado no Experimento 4.

Método	Archeage	HCR	Hobbit	IPhone6	OMD	Sanders	SS-Tweet	STS-Test	UMICH
B. C4.5	0,785±0,04	0,685±0,04	0,881±0,04	0,677±0,06	0,743±0,03	0,703±0,03	0,562±0,02	0,807±0,06	0,934±0,02
C4.5	0,753±0,02	0,643±0,03	0,864±0,02	0,681±0,05	0,682±0,04	0,640±0,04	0,543±0,03	0,752±0,06	0,920±0,02
Comitê	**0,869±0,02**	**0,733±0,03**	**0,921±0,02**	**0,738±0,04**	**0,811±0,02**	**0,756±0,03**	**0,612±0,02**	**0,863±0,03**	**0,969±0,01**
KNN	0,723±0,03	0,633±0,03	0,844±0,04	0,651±0,07	0,728±0,05	0,675±0,04	0,525±0,06	0,776±0,06	0,953±0,02
Logistic	0,836±0,03	0,679±0,03	0,908±0,03	0,698±0,05	0,775±0,02	0,711±0,03	0,585±0,02	0,804±0,03	0,962±0,02
NB B.	0,856±0,02	0,695±0,03	0,869±0,05	0,735±0,03	0,781±0,02	0,723±0,05	0,598±0,02	0,817±0,05	0,948±0,01
NB G.	0,793±0,02	0,590±0,02	0,772±0,08	0,689±0,04	0,628±0,04	0,651±0,02	0,530±0,03	0,751±0,04	0,760±0,06
NB M.	0,843±0,02	0,683±0,03	0,882±0,03	0,699±0,05	0,781±0,04	0,727±0,03	0,593±0,04	0,821±0,05	0,956±0,01
SVM-L	0,842±0,03	0,686±0,03	0,913±0,03	0,708±0,04	0,771±0,01	0,692±0,04	0,583±0,03	0,813±0,03	0,965±0,01
SVM-R	0,820±0,02	0,695±0,02	0,856±0,07	0,651±0,07	0,736±0,05	0,673±0,09	0,539±0,09	0,756±0,07	0,953±0,02