Análise de sentimentos em textos curtos provenientes de redes sociais

Nadia Felix Felipe da Silva

Nadia Felix Felipe da Silva

Análise de sentimentos em textos curtos provenientes de redes sociais

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, como parte dos requisitos para obtenção do título de Doutora em Ciências - Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientador: Prof. Dr. Eduardo Raul Hruschka

USP – São Carlos
Março de 2016

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados fornecidos pelo(a) autor(a)

F586a

Felix Felipe da Silva, Nadia
Análise de sentimentos em textos curtos provenientes de redes sociais / Nadia Felix Felipe da Silva; orientador Eduardo Raul Hruschka. -- São Carlos, 2016.
112 p.

Tese (Doutorado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2016.

1. Análise de sentimentos em textos curtos. 2. classificação de sentimentos. 3. análise de sentimentos em tweets. 4. classificação semissupervisionada de sentimentos . I. Raul Hruschka, Eduardo, orient. II. Título.

Nadia Felix Felipe da Silva

Sentiment analysis in short texts from social networks

Doctoral dissertation submitted to the Instituto de Ciências Matemáticas e de Computação - ICMC-USP, in partial fulfillment of the requirements for the degree of the Doctorate Program in Computer Science and Computational Mathematics. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. Eduardo Raul Hruschka

USP – São Carlos
March 2016

À minha família.

Agradecimentos

Muito especialmente, desejo agradecer ao meu orientador, Prof. Eduardo Raul Hruschka, por estar sempre presente, com disponibilidade em ajudar, inteligência, dedicação, paciência e muito bom humor, qualidades que o tornam um excelente professor e orientador.

Em particular, gostaria de registrar meu reconhecimento ao Prof. Estevam Rafael Hruschka Junior, que na prática atuou como meu co-orientador, pelo seu esforço em acompanhar todas as fases de desenvolvimento desse trabalho.

Sou grata também aos meus pais, Iva e Dioclemar, e a toda a minha família pelo incentivo e apoio durante os anos de estudo.

Agradeço também meu esposo Márcio, pelo incentivo, compreensão e encorajamento para seguir em frente nos meus objetivos.

Gostaria de deixar meus agradecimentos também aos companheiros de trabalho, que de alguma forma colaboraram com esta tese. Em especial, aos colegas Luíz Fernando Coleta, Jonathan Andrade, Thiago Covões, Luís Paulo Faina, André Luíz Vizine, e tantos outros amigos que por meus caminhos transitaram deixando suas contribuições.

Por fim, agradeço a todos os funcionários do ICMC/USP pela dedicação e competência e à CAPES pelo suporte financeiro[1].

[1]Trabalho realizado com suporte financeiro da CAPES (DS-7253238/D 10/2012 a 02/2016)

“O que vale na vida não é o ponto de partida e sim a caminhada.
Caminhando e semeando, no fim terás o que colher.”
(Cora Coralina)

Resumo

A análise de sentimentos é um campo de estudo com recente popularização devido ao crescimento da *Internet* e do conteúdo que é gerado por seus usuários, principalmente nas redes sociais, nas quais as pessoas publicam suas opiniões em uma linguagem coloquial e em muitos casos utilizando de artifícios gráficos para tornar ainda mais sucintos seus diálogos. Esse cenário é observado no Twitter, uma ferramenta de comunicação que pode facilmente ser usada como fonte de informação para várias ferramentas automáticas de inferência de sentimentos. Esforços de pesquisas têm sido direcionados para tratar o problema de análise de sentimentos em redes sociais sob o ponto de vista de um problema de *classificação*, com pouco consenso sobre qual é o classificador com melhor poder preditivo, bem como qual é a configuração fornecida pela engenharia de atributos que melhor representa os textos. Outro problema é que em um cenário supervisionado, para a etapa de treinamento do modelo de classificação, é imprescindível se dispor de exemplos rotulados, uma tarefa árdua e que demanda esforço humano em grande parte das aplicações. Esta tese tem por objetivo investigar o uso de agregadores de classificadores (*classifier ensembles*), explorando a diversidade e a potencialidade de várias abordagens supervisionadas quando estas atuam em conjunto, além de um estudo detalhado da fase que antecede a escolha do classificador, a qual é conhecida como engenharia de atributos. Além destes aspectos, um estudo mostrando que o aprendizado não supervisionado pode fornecer restrições complementares úteis para melhorar a capacidade de generalização de classificadores de sentimento é realizado, fornecendo evidências de que ganhos já observados em outras áreas do conhecimento também podem ser obtidos no domínio em questão. A partir dos promissores resultados experimentais obtidos no cenário de aprendizado supervisionado, alavancados pelo uso de técnicas não supervisionadas, um algoritmo existente, denominado de C^3E (*Consensus between Classification and Clustering Ensembles*) foi adaptado e estendido para o cenário semissupervisionado. Este algoritmo refina a classificação de sentimentos a partir de informações adicionais providas pelo agrupamento em um procedimento de autotreinamento (*self-training*). Tal abordagem apresenta resultados promissores e competitivos com abordagens que representam o estado da arte em outros domínios.

Palavras-chave: Análise de sentimentos em textos curtos, classificação de sentimentos, análise de sentimentos em *tweets*, classificação semissupervisionada de sentimentos.

Abstract

Sentiment analysis is a field of study that shows recent popularization due to the growth of Internet and the content that is generated by its users. More recently, social networks have emerged, where people post their opinions in colloquial and compact language. This is what happens in Twitter, a communication tool that can easily be used as a source of information for various automatic tools of sentiment inference. Research efforts have been directed to deal with the problem of sentiment analysis in social networks from the point of view of a classification problem, where there is no consensus about what is the best classifier, and what is the best configuration provided by the feature engineering process. Another problem is that in a supervised setting, for the training stage of the classification model, we need labeled examples, which are hard to get in the most of applications. The objective of this thesis is to investigate the use of classifier ensembles, exploring the diversity and the potential of various supervised approaches when these work together, as well as to provide a study about the phase that precedes the choice of the classifier, which is known as feature engineering. In addition to these aspects, a study showing that unsupervised learning techniques can provide useful and additional constraints to improve the ability of generalization of the classifiers is also carried out. Based on the promising results got in supervised learning settings, an existing algorithm called C^3E (Consensus between Classification and Clustering Ensembles) was adapted and extended for the semi-supervised setting. This algorithm refines the sentiment classification from additional information provided by clusters of data, in a self-training procedure. This approach shows promising results when compared with state of the art algorithms.

Key-words: Sentiment analysis of short informal texts, sentiment classification, tweet sentiment analysis, semi-supervised sentiment analysis.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

PLN	Processamento de Linguagem Natural
OM	*Opinion Mining*
SA	*Sentiment Analysis*
SMS	*Short Message Service*
TSA	*Twitter Sentiment Analysis*
BoW	*Bag of Words*
NB	*Naive Bayes*
$MaxEnt$	*Máxima Entropia*
SVM	*Support Vector Machines*
LR	*Regressão Logística*
MNB	*Multinomial Naive Bayes*
CRF	*Conditional Random Field*
TF	Frequência de palavras nos documentos
KNN	*k nearest neighbors*
FH	*Feature hashing*
Lex	Léxico de sentimentos
$\mathcal{X}^T$	Conjunto de treinamento
$\mathcal{X}$	Conjunto alvo (de teste)
M	Número de atributos dos objetos em $\mathcal{X}^T$
N	Número de objetos em $\mathcal{X}^T$
n	Número de objetos em $\mathcal{X}$
c	Número de classes
k	Número de grupos
$\mathbf{D}$	Matriz de distâncias
$\mathbf{S}$	Matriz de similaridades
$\boldsymbol{\pi}$	Vetor de estimativas de distribuição de probabilidades de classes
$\mathbf{y}$	Vetor de estimativas refinadas de distribuição de probabilidades de classes *a posteriori*

α	Parâmetro do algoritmo C^{3}E para controle da importância de classificadores e agrupadores
I	Número de iterações do algoritmo C^{3}E
P	Número de indivíduos da população usada pelas meta-heurísticas
F	Taxa de escalonamento do algoritmo *Differential Evolution*
Cr	Taxa de *crossover* do algoritmo *Differential Evolution*

I

Capítulo

1

Introdução

Apesar de as áreas de Linguística e Processamento de Linguagem Natural (PLN) estarem consolidadas, com uma história de sucesso que se iniciou em meados da década de 40 (Jones, 1994), pouca pesquisa foi feita sobre a opinião e sentimentos das pessoas antes do ano 2000 (Liu, 2012). A tentativa de compreender o poder de persuasão que uma opinião alheia pode exercer sobre outras pessoas fez emergir uma importante área de pesquisa conhecida como **análise de sentimentos**, que é objeto de estudo desta tese de doutorado.

1.1 Análise de sentimentos

As atividades de pesquisa nas áreas de tratamento de opiniões, sentimentos e emoções em textos vêm ganhando destaque, em parte por conta da enorme quantidade de texto disponível na *Web Social* nas formas de notícias, comentários, *blogs*, redes sociais e bate-papos. Nestes ambientes, os usuários são fortemente induzidos a dar opiniões sobre temas polêmicos e fortemente debatidos, ou mesmo sobre eventos, produtos e serviços, e situações do cotidiano (Liu, 2012). Segundo Liu (2012), em Ciência da Computação, o conjunto de técnicas, algoritmos e modelos concebidos responsáveis por realizar o tratamento de opiniões é abordado pela área de **Mineração de Opiniões** (*Opinion Mining* – OM) ou **Análise de Sentimentos** (*Sentiment Analysis* – SA). Trata-se de um campo emergente multidisciplinar que mescla conceitos de mineração de dados, aprendizado de máquina, linguística, processamento de linguagem natural e análise textual, e cujo objetivo é analisar fragmentos textuais e determinar a atitude, emoção, opinião, avaliação ou sentimento do escritor com relação a algum tópico ou entidade (Pang & Lee, 2008; Liu, 2012; Taboada *et al.*, 2011; Pang *et al.*, 2002; Turney, 2002). OM e SA são áreas que têm atraído um grande interesse de pesquisa e que visam ajudar os usuários a encontrar informações opinativas e detectar a polaridade da opinião, ou seja, detectar se um dado texto possui conotação positiva ou negativa — há autores que também admitem a classe neutra (ausência de

sentimento) neste cenário como, por exemplo, Pang *et al.* (2002) e Turney (2002).

OM e SA são comumente usadas como sinônimos para expressar o mesmo significado (Liu, 2012). No entanto, alguns pesquisadores afirmam que essas duas áreas têm por objectivo abordar dois problemas ligeiramente diferentes. De acordo com Tsytsarau & Palpanas (2012), a OM está relacionada à forma de determinar se um texto contém uma opinião (um problema que também é conhecido como Análise de Subjetividade), enquanto que SA está relacionada à detecção da polaridade de sentimento (em que é atribuída um sentimento positivo ou negativo para o texto examinado). De acordo com a definição dada por Liu & Zhang (2012), "uma opinião é uma declaração subjetiva, com uma visão pessoal, que expressa atitude, emoção ou apreciação sobre uma entidade ou um aspecto de uma entidade, um parecer". Neste mesmo estudo, os autores definem que "uma entidade é um objeto concreto ou abstrato, como um produto, uma pessoa, um evento, uma organização, e que pode ser representada como uma hierarquia de componentes, sub-componentes, e seus atributos".

Existem diferentes níveis de análise textual permeando as atividades de pesquisa de análise de sentimentos. Em geral com relação à granularidade da análise podemos distinguir três categorias (Liu, 2012):

1. **Granularidade de documento**: A tarefa consiste em classificar se a opinião implícita no documento expressa um sentimento positivo, negativo ou neutro (Pang *et al.*, 2002; Turney, 2002). Por exemplo, dado um comentário sobre um produto, uma ferramenta de análise de sentimentos determina se tal comentário expressa uma opinião positiva ou negativa. Esta tarefa é geralmente conhecida como *Classificação de Sentimentos do Documento*. Tal análise presume que cada documento expressa opiniões sobre uma única entidade (e.g., um único produto, pessoa etc).

2. **Granularidade de sentença**: Nesta tarefa, o texto é subdividido em sentenças e a análise é feita sobre tais unidades textuais com o objetivo de definir se as mesmas expressam individualmente um sentimento positivo, negativo ou neutro.

3. **Granularidade de entidade e de aspecto**: Na granularidade de documento ou sentenças, a análise não relaciona a entidade ao seu respectivo sentimento, ou seja não se descobre o que exatamente a pessoa gosta ou não. Na granularidade de aspectos, a análise se torna mais específica, sendo anteriormente conhecida como mineração de opiniões baseada em características (*features*). Por exemplo, a sentença "*Apesar do serviço não ser o melhor, eu ainda amo vir a este restaurante.*", se analisada em termos gerais, com granularidade de documento ou de sentença, poderia ser classificada em sua totalidade com conotação negativa, embora esta não seja negativa em sua totalidade. Isto se dá porque em relação à entidade "restaurante" o autor tem uma opinião positiva, mas em relação ao "serviço" o autor tem uma opinião negativa.

Há ainda desafios de pesquisa estabelecidos quanto ao tipo de fonte textual a ser tratada, uma vez que resultados de análises de sentimentos mostram que diferentes tipos de textos requerem métodos especializados de análise como, por exemplo, sentimentos não são expressados da

mesma maneira em textos jornalísticos, blogs, comentários, fóruns, mensagens em redes sociais ou outros tipos de conteúdos gerados por usuários (Balahur *et al.*, 2010).

Tradicionalmente, os sistemas de análise de sentimentos são treinados e utilizados em textos mais longos, como comentários e avaliações de filmes ou produtos. Recomenda-se o trabalho de Pang & Lee (2008), e mais recentemente, o livro sobre análise de sentimento e mineração de opiniões escrito por Liu (2012) àqueles interessados em uma revisão sistemática desta área de pesquisa, sob a ótica tradicional. Esta tese se concentra no problema de análise de sentimento em *microblogs*, cujo tipo de fonte textual possui suas próprias peculiaridades e portanto introduz grandes desafios à análise de sentimentos.

1.2 Análise de sentimentos em *microblogs* e *tweets*

Sítios de redes sociais como *Facebook*[1], *Twitter*[2], *Google+*[3], *Tumblr*[4], *FourSquare*[5] e *LinkedIn*[6] se tornaram aplicações populares, fazendo parte do cotidiano das pessoas e favorecendo o intercâmbio cultural, além do compartilhamento de informações entre seus usuários. Diante deste crescimento na utilização, bem como a facilidade em disponibilizar cada vez mais dados em redes sociais, continua sendo um desafio organizar e extrair informações úteis a partir deste conteúdo.

O conteúdo gerado pelo usuário varia muito, desde um simples *like* ("curtir") em atualizações de *status* no Facebook, a longas publicações em *blogs*. Toda esta informação produzida pode se tornar valiosa para as empresas interessadas em saber a reputação de seus serviços ou produtos, podendo ter efeito prático na qualidade dos mesmos. Além disso, é possível que o governo use estas informações para entender a visão do público sobre diferentes questões sociais, podendo, consequentemente, agir de forma mais eficaz.

Outra aplicação diz respeito aos usuários que são potenciais clientes de um produto, os quais podem usar as informações opinativas de outros usuários para decidir se irão adquirir ou não o produto. Até recentemente, as principais fontes de informação opinativas eram fornecidas por amigos e sites especializados. Agora, os consumidores podem consultar as experiências passadas e opiniões publicadas por outros usuários antes de comprar um produto específico. Um estudo recente[7] descobriu que quase 80% dos consumidores se mostram mais interessados em uma empresa por causa da presença de sua marca em redes sociais. Alguns dados levantados nesta pesquisa são relevantes para esta tese e apresentados a seguir:

- Os consumidores que buscam por um serviço ou produto em redes sociais têm em média 39 anos de idade e 84% são do sexo feminino. Isto é um demonstrativo de que as

[1] http://www.facebook.com/
[2] http://twitter.com/
[3] https://plus.google.com
[4] http://tumblr.com/
[5] http://foursquare.com/
[6] http://linkedin.com/
[7] http://www.iabuk.net/blog/unlocking-the-power-of-social-media

rede sociais têm atraído muitos adultos, um nicho diferente dos adolescentes, comumente citados nestes ambientes.

- O estudo também revelou que 86% dos consumidores acessam as redes sociais em casa através de seus *laptops* (59%), seguidos pelos *smartphones* (31%), *desktops* (30%) e finalmente os *tablets* (12%).

De acordo com outros dois estudos desenvolvidos com mais de 2.000 adultos americanos cada (Pang & Lee, 2008), observou-se ainda que:

- 81% das pessoas entrevistadas já fizeram alguma pesquisa *online* sobre um produto pelo menos uma vez na vida;
- Em média, 87% dos leitores de comentários *online* relataram que tais comentários tiveram uma influência significativa sobre a sua compra;
- Consumidores relataram estar dispostos a pagar de 20% a 99% a mais por um produto bem avaliado nestes ambientes;
- 32% disseram já ter sugerido uma classificação de um produto, serviço ou pessoa através de um sistema de classificação *on-line*;
- 30% dos entrevistados disseram já ter postado um comentário *online* ou crítica a respeito de um produto ou serviço.

O Twitter é atualmente o *microblog* mais popular na *Internet* e possui como característica principal permitir aos seus usuários divulgar o que estão fazendo em tempo real, para todos os usuários ligados à sua rede (seus seguidores), com um limite máximo de 140 caracteres[8], via vários mecanismos de comunicação, incluindo clientes *web*, clientes móveis (*mobile clients* – *smartphones*, tablets e similares) e SMS. Essas *micro-mensagens* são denominadas *tweets*. A popularização do Twitter trouxe consigo o crescimento do tráfego de informações e consequentemente a dificuldade de análise e extração de conteúdo útil nestes ambientes. Informações levantadas pela própria plataforma e sites especializados, demostram que:

- Em média, 500 milhões de *tweets* são enviados por dia (Twitter, 2014);
- Há 288 milhões de usuários ativos (Twitter, 2014);
- 80% dos usuários ativos do Twitter provêm de aparelhos móveis e/ou portáteis (Twitter, 2014);
- Até 2012, 56% dos *tweets* de clientes ainda são ignorados pelas empresas e companhias potencialmente interessadas nesse tipo de informação (Basch, 2012);

[8] http://www.140characters.com/2009/01/30/how-twitter-was-born/

- Os três países com maior número de usuários no Twitter são: os EUA, com 107 milhões, o Brasil, com 33 milhões e o Japão, com cerca de 30 milhões (Bullas, 2012). 77% das contas estão fora dos EUA (Twitter, 2014);

- Em média, usuários seguem (ou são seguidos por) 51 pessoas (Basch, 2012);

- 32% de todos usuários da *Internet* são também usuários do Twitter (McGee, 2012);

- 69% dos seguidores no Twitter são sugeridos por amigos (Kane, 2012).

Além dos fatores relacionados ao número de usuários ativos e do envio médio de mensagens realizado por tais usuários, a aplicação de técnicas de análise de sentimentos em *tweets* é também justificada e facilitada pela existência de uma interface de desenvolvimento[9] fornecida pelos desenvolvedores do Twitter, que provê serviços como *download de tweets*.

A seguir é apresentado uma lista de termos específicos da plataforma Twitter, que será usada no decorrer desta tese:

- *Tweet:* Uma mensagem ou post na plataforma Twitter recebe o nome de *tweet*. Um *tweet* é restrito ao limite de 140 caracteres, como já mencionado anteriormente.

- *User ou Username:* Para poder postar *tweets*, o usuário deve se registar no Twitter com um pseudônimo ou *username*. Para interagir com outros usuários do Twitter, os autores podem mencionar outros pseudônimos de usuários usando o símbolo `@` (por exemplo `@maria`).

- *Hashtag:* Usando o símbolo #, os usuários podem atribuir *tags* a seus *tweets*, indicando por exemplo a relevância de um certo tópico, por exemplo, $\#ProtestoBR$.

- *Follower:* Usuários no Twitter podem se conectar uns aos outros, significando que serão notificados sobre novos *tweets* postados por suas conexões. A relação de seguir alguém no Twitter não é bidirecional, tal como no Facebook por exemplo.

- *Retweet: Tweets* podem ser redistribuídos através da funcionalidade chamada *retweeting*, a qual é usada para propagar uma mensagem de outro usuário para seus seguidores. Neste caso, o *tweet* é então não modificado, ou somente marcado com a abreviação `RT` para indicar o *retweet*.

A Figura 1.1 mostra um exemplo de *tweet*. O mesmo foi extraído do perfil do usuário NetflixBrasil. É possível observar que ele contém alguns dos termos inerentes à plataforma Twitter mencionados anteriormente, como por exemplo, o *username* (NetflixBrasil) e a citação a outro usuário (@CarlaPernambuco). Além do uso de *hashtags* (#ChefsTable).

[9] https://dev.twitter.com/overview/api

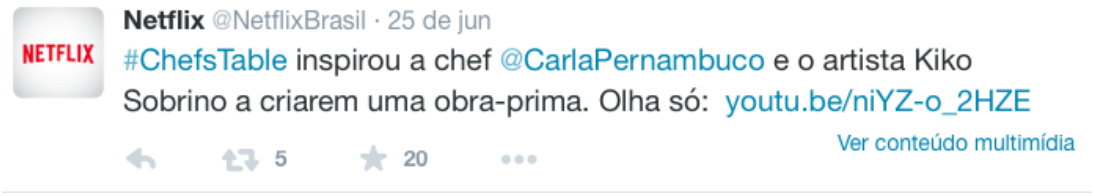


Figura 1.1: Um exemplo de *tweet* postado pelo usuário `@NetFlixBrasil`

1.3 Desafios em tratar *tweets*

Ambientes que favorecem a geração de textos curtos e informais são cenários de pesquisa desafiadores para a análise de sentimento, nos quais o principal objetivo é extrair avaliações de produtos, categorização por sentimentos, agrupamento e extração de padrões comportamentais. O Twitter é um excelente representante desde cenário, com uma área de pesquisa específica para esta plataforma conhecida como *Twitter Sentiment Analysis* (TSA) — (Go *et al.*, 2009). A TSA aborda o problema de analisar o texto de *tweets* em termos dos sentimentos que são expressados no mesmo. Além dos desafios que os sistemas de análise de sentimento que lidam com textos tradicionais enfrentam, a análise de sentimento no Twitter tem que lidar com dificuldades adicionais: tamanho do texto, variação na ortografia, esparsidade dos dados, definição de contexto e negação, são alguns dos muitos desafios elencados nesta seção. Desta forma, a seguir são listadas tais dificuldades, justificando as razões pelas quais não se deve aplicar métodos de análise de sentimentos de propósito geral em *tweets* sem que antes haja um estudo dos prós e contras de seu uso com esta fonte de dados.

Tamanho do texto: Mensagens de *microblogs* são geralmente muito curtas. As mensagens oriundas do Twitter têm como limitante superior 140 caracteres, entretanto um estudo[10] realizado em 2012 relatou que os seus usuários em geral são bem mais sucintos que isso, os quais têm por hábito publicar mensagens com 28 caracteres em média. Enquanto isso pode parecer uma vantagem, uma vez que os autores tendem a ir direto ao ponto, um desafio apresentado neste cenário é que a opinião expressa pode ser dependente de poucas palavras. Tais palavras podem não estar disponíveis no recurso léxico (um dicionário de palavras pré-compilado já anotado com o sentimento em questão) ou podem não ter ocorrido nos dados de treinamento. Uma discussão sobre este fenômeno é apresentada por Bermingham & Smeaton (2010). Devido à limitação do comprimento, a maioria dos *tweets* contém uma única sentença. Portanto, para a tarefa de análise de sentimentos em *tweets* não existe diferença fundamental quanto à análise feita em granularidade de documento ou de sentença. No caso de *tweets*, a SA pode ser aplicada em dois níveis de granularidade: documento/sentença e entidade/aspectos. Esta tese se concentra na granularidade de documento, i.e. presumindo que um *tweet* contém apenas uma sentença.

Variação na ortografia: Devido à espontaneidade, o contexto informal e restrições de com-

[10] http://thenextweb.com/twitter/2012/01/07/interesting-fact-most-tweets-posted-are-approximately-30-characters-long/

primento da mensagem, a ortografia em *tweets* tende a ter uma maior variabilidade do que em outros gêneros textuais (*web, blogs,* jornais etc). Diversos fenômenos são identificados devido à variação ortográfica, incluindo erros ortográficos, abreviações, uso de maiúsculas e repetições de letras e sílabas para atribuir ênfase. Isso induz a muito mais esparsidade na entrada, o que é um desafio especial para a utilização de recursos lexicais. Brody & Diakopoulos (2011) apresentam um estudo detalhado sobre alongamento enfático e seu impacto em análise de sentimentos em *tweets*. De acordo com este estudo, o alongamento enfático (e.g. *Eu estou felllllllizzzzzzz!!!!!*) é muito frequente no Twitter, ocorrendo em cerca de um a cada seis *tweets*.

Esparsidade dos dados: *Tweets* contêm uma grande quantidade de "ruído" devido ao uso extensivo de erros de ortografia. Este fenômeno causa uma esparsidade dos dados e tem um impacto sobre o desempenho global da análise de sentimento. A principal razão para a esparsidade de dados é o fato de que uma grande porcentagem dos termos que aparecem nos *tweets* ocorrem menos de 10 vezes (Saif *et al.*, 2012a) em todo o conjunto de dados (*corpus*). Um estudo focado na redução da esparsidade dos dados nos *tweets* é apresentado por Saif *et al.* (2012a), que propuseram uma suavização semântica para reduzir a esparsidade.

Negação: A presença de palavras de negação desempenha um papel importante na detecção da polaridade do sentimento de uma mensagem. A detecção e o tratamento adequado das negações não é trivial e é um desafio para análise de sentimento. Detectar negações é importante porque podem causar a inversão da polaridade de uma mensagem (uma mensagem positiva torna-se negativa ou vice-versa). Vários pesquisadores adotam uma técnica simples, invertendo a polaridade das mensagens, quando uma palavra de negação é detectada. Uma abordagem mais avançada para o tratamento de negação em análise de sentimentos em *tweets* é apresentada por Svetlana Kiritchenko & Mohammad (2014). Eles desenvolveram dois léxicos separados, um com termos que geralmente aparecem em contextos com negações e um com termos que aparecem em contextos sem negações. Em seu estudo, eles mostraram que a negação de termos positivos tendem a implicar em sentimentos negativos enquanto que a negação de sentimentos negativos tendem a fazer com o que o sentimento permaneça negativo.

***Stop Words*:** *Stop Words* são palavras que têm baixo poder de discriminação (por exemplo, "a", "é", "que" etc.) e elas geralmente são filtradas antes de processar o texto. Tipicamente as *stop words* são listas de palavras pré-compiladas e não são adequadas para o Twitter, podendo influenciar o desempenho da análise de sentimentos nesses ambientes. Por exemplo, a palavra do inglês "*like*", é geralmente considerada como uma palavra membro das listas de *stop words*. No entanto, tal palavra tem um importante poder de discriminação de sentimento. Para solucionar este tipo de problema, há trabalhos focados na construção de listas de *stop word* para o Twitter. Saif *et al.* (2014) apresentaram um estudo no qual eles avaliam o impacto da remoção de *stop words* em análise de sentimentos em *tweets*

efetivamente. Foram usados seis conjuntos de dados diferentes nos quais foram aplicados seis diferentes métodos de identificação de *stop words* com o objetivo de avaliar o quanto eles influenciam no desempenho de dois métodos de análise de sentimentos. Os autores levantaram observações quanto ao tamanho do conjunto de atributos e performance na classificação.

Símbolos especiais (Special tokens): Emoticons e URLs são alguns exemplos de tokens não encontrados em outros gêneros textuais. Tais símbolos podem levar a dificuldades ao tentar usar ferramentas de processamento de linguagem natural, como *part-of-speech taggers* e *parsers* sintáticos. *Parses* sintáticos, por sua vez, são treinados em textos jornalísticos, que são consideravelmente diferentes de mensagens de *microblogs*.

Variação de Tópicos: Os temas discutidos no Twitter não são limitados, tendo portanto uma variedade muito grande. Isso pode causar problemas para análise de sentimento, por exemplo, quando palavras expressam um sentimento diferente em diferentes contextos. A maior parte do trabalho que é feito em TSA tem como objetivo classificar a orientação do sentimento de um *tweet* sem considerar a relevância tópica. Para capturar a relevância do tema de um *tweet*, muitos pesquisadores simplesmente consideram a presença de uma palavra como um indicador do tópico. Além disso, outros estudos (por exemplo Davidov *et al.* (2010)) consideram o símbolo *hashtag* como um forte indicador de tópico.

Quantidade de Dados: Enquanto as mensagens são frequentemente muito pequenas (com um número de caracteres relativamente pequeno), a quantidade de textos postados pode ser extremamente grande. Em 2014 o Twitter anunciou que seu serviço de postagens acumulou 24,9 milhões de *tweets* (381.605 *tweets* por minuto) postados durante o *Super Bowl*, a final da liga profissional de futebol americano dos Estados Unidos. Em um ambiente desta proporção, uma informação opinativa sobre um determinado tópico está completamente escondida e, portanto, é praticamente impossível que um usuário comum extraia, através das diferentes fontes, conteúdo útil.

Estilo de linguagem: Devido à grande diversidade de usuários do Twitter, o estilo de escrita e linguagem também é muito variável. As mensagens variam de estilos de escrita estritamente formais, como por exemplo de textos jornalísticos, a textos completamente informais com gírias e expressões idiomáticas. Além disso, o vocabulário usado pode mudar rapidamente. Tudo isso pode conduzir a problemas na análise de sentimentos através de recursos léxicos ou em lidar com dados de treinamento que foram anotados com a classe previamente.

Contexto Multilingual: Enquanto os jornais e *blogs on-line* tendem a ser escritos em um único idioma, usuários de plataformas de *microblogging* podem usar uma ampla variedade de línguas, às vezes até na mesma mensagem ou frase. Um estudo desenvolvido por Narr *et al.* (2012) apresentou um classificador de sentimentos independente de linguagem. Tal classificador foi avaliado em *tweets* de quatro línguas: Inglês, Alemão, Francês e

Português. Eles mostraram que o método proposto neste estudo é aplicável nas quatro línguas avaliadas sem a necessidade de processamentos adicionais.

***Tokenização*:** Outro desafio relacionado à análise de sentimentos é a tokenização das sentenças. Em vez da divisão por espaços em branco, comumente empregada em tarefas de processamento de textos para a tokenização de sentenças, Owoputi *et al.* (2013a) propuseram um tokenizador específico para o Twitter.

1.4 Objetivos

O objetivo geral dessa tese é desenvolver métodos de análise de sentimentos para aplicações reais. Sob o ponto de vista prático, além de lidar com todos os desafios mencionados anteriormente, faz-se importante ressaltar que as pesquisas em análise de sentimento em *tweets* têm avançado na direção de métodos de classificação baseados no uso de léxicos (Turney, 2002; Taboada *et al.*, 2011; Thelwall *et al.*, 2010; Ortega Bueno *et al.*, 2013; Hu *et al.*, 2013a) bem como em métodos de classificação desenvolvidos segundo o paradigma supervisionado (Go *et al.*, 2009; Pak & Paroubek, 2010; Barbosa & Feng, 2010; Bakliwal *et al.*, 2012; Mohammad *et al.*, 2013; Kiritchenko *et al.*, 2014; Hamdan *et al.*, 2013). Tais abordagens presumem a necessidade de dois recursos de alto custo a serem construídos previamente à análise:

1. Métodos baseados em léxicos obrigatoriamente necessitam de uma lista de palavras previamente computada. Nesta lista, atribui-se a cada palavra presente um sentimento – positivo ou negativo. Tal atribuição é feita por um especialista de domínio, e é completamente dependente do contexto. O uso de um léxico desenvolvido para outra fonte de dados (diferente de *tweets*, podendo ser comentário de filmes, livros ou *blogs*, por exemplo) ou outro domínio diferente do domínio em questão, quando usada indiscriminadamente ou sem nenhuma adaptação tende a impactar diretamente na acurácia do método.

2. Para métodos de análise de sentimentos definidos sob o paradigma supervisionado, é imprescindível a definição prévia de um conjunto de treinamento, em que um especialista de domínio atribui a cada *tweet* analisado um rótulo de classe de sentimento. Nestes casos, a acurácia do método de análise posteriormente empregado será diretamente proporcional ao tamanho e à qualidade do conjunto de treinamento.

Tendo em vista os inconvenientes dos métodos de classificação baseados em léxicos e em conjuntos de tweets representativos e apropriadamente rotulados, são objetivos específicos dessa tese desenvolver abordagens que forneçam:

- Robustez e desempenho superiores (ou pelo menos competitivos) àqueles obtidos por classificadores tradicionais supervisionados. Para atingir esse objetivo, serão propostos *ensembles* baseados em classificadores e *ensembles* baseados em classificadores e agrupadores de dados, fazendo uso de uma grande variedade de fontes de dados (incluindo léxicos).

- A flexibilidade de não necessitar de uma base de dados completamente rotulada na fase de treinamento do modelo de classificação. Nesse sentido, a partir dos promissores resultados experimentais obtidos no cenário de aprendizado supervisionado, alavancados pelo uso de técnicas não supervisionadas, um algoritmo existente, denominado de C^3E (*Consensus between Classification and Clustering Ensembles*), foi adaptado e estendido para o cenário semissupervisionado.

1.5 Hipóteses de pesquisa

Foram desenvolvidas três principais frentes de trabalho. A primeira frente busca entender o cenário de análise de sentimentos sob a ótica supervisionada, com a aplicação de agregadores de classificadores, dado que existe uma extensa literatura em outros domínios mostrando que um agregador obtido a partir de classificadores diversificados é tipicamente mais robusto em comparação à seus componentes individuais (e.g., ver Kuncheva (2004); Dietterich (2000); Rokach (2010); Zhou (2012); Zhang & Ma (2012); Schapire & Freund (2012), e que agregadores têm sido pouco aplicados em análise de sentimentos. Levando isso em consideração, a primeira hipótese desta tese é definida como:

Hipótese 1 *Ensembles* de classificadores diversificados apresentam resultados competitivos ou superiores aos resultados inferidos a partir de classificadores individuais para análise de sentimentos em *tweets*.

A segunda frente de trabalho envolve a agregação de classificadores e agrupadores como estratégia de refinamento do modelo de classificação de sentimentos. Fundamentalmente, resultados de agrupamentos podem ajudar a aumentar a capacidade de generalização de classificadores (Bousquet *et al.*, 2003; Cai *et al.*, 2010, 2009; Gao *et al.*, 2013, 2009), bem como podem ser úteis para identificar mudanças de conceito no conjunto-alvo (Masud *et al.*, 2013, 2010a,b). Partindo dessa premissa, a segunda hipótese desta tese é definida como:

Hipótese 2 O conhecimento inferido no conjunto-alvo ao fazer uso de um agrupador de dados permite refinar o modelo de classificação de sentimentos inferido no conjunto de treinamento, fornecendo resultados competitivos ou superiores aos resultados inferidos a partir de classificadores derivados apenas de fontes de informações supervisionadas.

A terceira frente de trabalho busca flexibilizar a obrigatoriedade de um conjunto de treinamento que deve estar previamente rotulado. Considerar a análise de sentimentos como uma tarefa de classificação supervisionada inviabiliza a sua aplicação em muitos cenários reais com dados do Twitter, uma vez que tal pressuposto exige uma base de treinamento completamente rotulada. Para contornar essa limitação, é amplamente conhecido que se pode usar uma configuração semissupervisionada. Trabalhos nessa linha, entretanto, são raros para análise de sentimento em *tweets*. Uma exceção é o trabalho de Xiang & Zhou (2014), que faz uso de agrupamento de dados para refinar modelos de classificação. Neste contexto, define-se a terceira hipótese de pesquisa da tese:

Hipótese 3 Algoritmos semissupervisionados obtidos por meio da combinação de agregadores de classificadores e agrupadores de dados proporcionam classificadores de sentimentos que requerem menos dados rotulados e que fornecem boa capacidade de generalização.

Ao abordar essas três frentes de pesquisa, a presente tese contribui com a literatura tanto de aprendizado supervisionado quanto de aprendizado semissupervisionado para a análise de sentimentos em textos curtos provenientes de redes sociais, especialmente para *tweets*. As abordagens propostas, bem como os experimentos realizados para validar as hipóteses aqui levantadas, constituem contribuições relevantes para diversas áreas, tais como: aprendizado de máquina, análise de sentimentos e processamento de linguagem natural. Boa parte do material apresentado nessa tese foi publicada durante o trabalho de doutoramento em: da Silva *et al.* (2014c,b,a)

1.6 Organização da Tese

O restante desta tese está organizado da seguinte forma:

Capítulo 2 - Aprendizado Supervisionado: São apresentados os algoritmos supervisionados tradicionalmente empregados em análise de sentimentos em *tweets* e a validação da **Hipótese de pesquisa 1**. As contribuições deste capítulo foram publicadas em periódico e conferência internacionais (da Silva *et al.*, 2014c,b).

Capítulo 3 - Agregadores de Classificadores e Agrupadores: São apresentadas as fundamentações teóricas e práticas que envolvem a combinação de classificadores e agrupadores. O algoritmo C^3E (Acharya *et al.*, 2011) é detalhadamente descrito e validado no cenário de análise de sentimentos em *tweets*. O conteúdo deste capítulo foi publicado em da Silva *et al.* (2014a). Este capítulo trata da **Hipótese de pesquisa 2**.

Capítulo 4 - Aprendizado Semissupervisionado agregando classificadores e agrupadores: Na primeira parte é apresentada uma análise bibliográfica do cenário semissupervisionado de análise de sentimentos em *tweets*. Tal análise ainda não havia sido feita na literatura e foi documentada em da Silva *et al.* (2016a). Na segunda parte do capítulo, apresenta-se um *framework* semissupervisionado que combina as tarefas de agrupamento e classificação. Os resultados apresentados estão relacionados à **Hipótese de pesquisa 3**. Tais resultados foram publicados em da Silva *et al.* (2016b).

Capítulo 5 - Conclusões e trabalhos futuros: Este capítulo sumariza as principais conclusões originadas desta tese, além de apresentar possíveis direções de pesquisa para trabalhos futuros.

Capítulo

2

Aprendizado Supervisionado

Após uma análise criteriosa dos diversos trabalhos sobre análise de sentimentos aplicados em *tweets* (TSA), é possível observar que, em sua maioria, tais métodos podem ser categorizados segundo os paradigmas de classificação: *(i) supervisionada*, com uso de algum algoritmo de aprendizagem de máquina e com a obrigatoriedade de uma base de treinamento previamente rotulada; *(ii) guiada pelo uso de um léxico*, i.e., uma lista de termos positivos e negativos (manualmente ou automaticamente computada) que direcionará o processo de inferência da polaridade; *(iii) supervisionada híbrida*, baseada no uso de algum algoritmo de aprendizagem de máquina em conjunto com um léxico; ou *(iv) baseada em grafos* (*Graph-Based*).

Considerando tais paradigmas de classificação, os três primeiros métodos (baseados em aprendizagem de máquina, baseados em léxicos e os métodos híbridos) podem ser aplicados a qualquer tipo de fonte textual (por exemplo, resenhas de livros, filmes, comentários de produtos, blogs e redes sociais), enquanto a abordagem baseada em grafos explora as propriedades específicas da rede social, como, por exemplo, os relacionamentos entre os seus usuários (Speriosu *et al.*, 2011) e, portanto, necessita de um *corpus* com tais informações. O foco deste capítulo está nos métodos baseados em paradigmas de classificação supervisionada e de classificação híbrida.

Os métodos baseados em paradigmas de classificação supervisionada e de classificação híbrida seguem uma sequência de passos comuns às aplicações de mineração de textos (Liu, 2012) — ver Figura 2.1. O primeiro passo diz respeito à seleção de um conjunto de *tweets* de interesse. Este processo se dá por meio de uma busca por tópicos de pesquisa, *hashtags*, *emoticons*, um período de pesquisa, entre outras formas. Estes *tweets* devem ser rotulados, sendo que na maioria dos casos este processo de anotação se dá manualmente por um especialista de domínio ou de maneira automática, considerando as *hashtags* e *emoticons* utilizadas na pesquisa — nesta última forma de rotulação os *tweets* que possuem *hashtags* e *emoticons* com conotação positiva são considerados positivos, os que possuem *hashtags* e *emoticons* com conotação negativa

são considerados negativos, e os que não possuem *hashtags* e *emoticons* são desconsiderados[1]. Posteriormente, estes *tweets* são pré-processados e transformados em um modelo de representação baseado em *bag-of-words* (BoW). A BoW é fornecida como entrada para o treinamento do algoritmo de classificação, que proverá um modelo de classificação pronto para ser usado em novos *tweets* (diferentes dos *tweets* de treinamento).

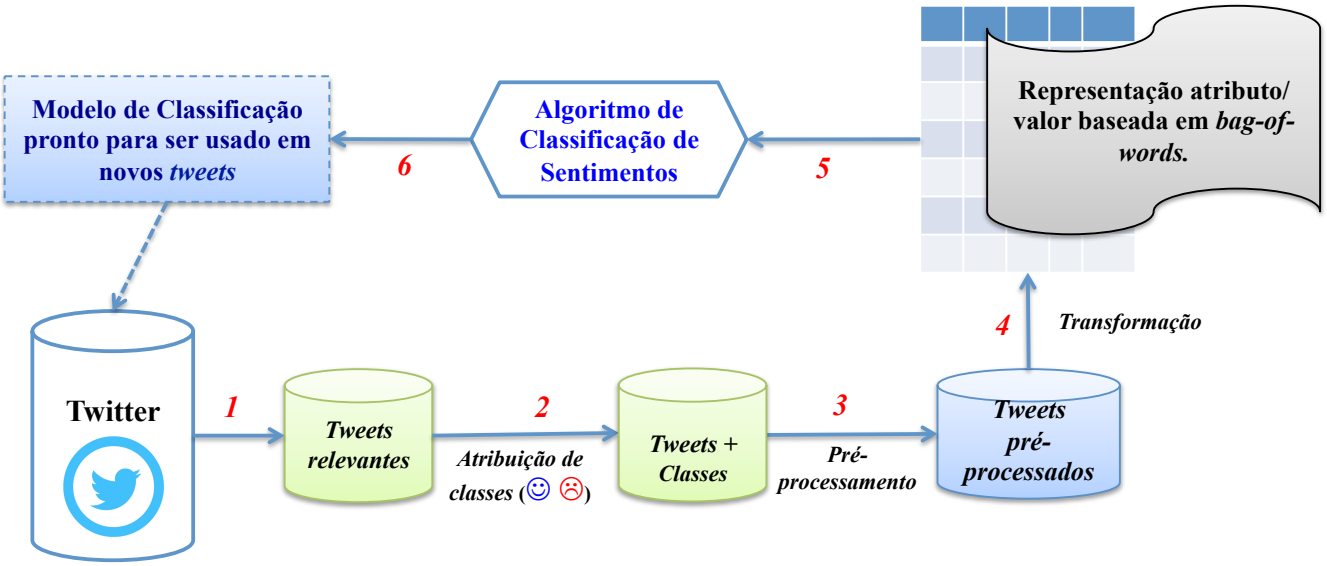


Figura 2.1: Passos para obtenção de um modelo de classificação de sentimentos em *tweets*.

Em uma *bag-of-words*, os *tweets* são representados por meio de uma tabela na qual as colunas representam os termos, palavras, léxicos ou outros tipos de atributos existentes nas mensagens, e os valores associados às colunas são referentes à frequência (ou presença) desses termos (ou atributos) no *tweet*. Dessa forma, uma coleção de *tweets*, após o pré-processamento, pode ser representada como ilustrada na Tabela 2.1.

	t_1	t_2	$\cdots$	t_m
$tweet_1$	a_{11}	a_{12}	$\cdots$	a_{1m}
$tweet_2$	a_{21}	a_{22}	$\cdots$	a_{2m}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$tweet_n$	a_{n1}	a_{n2}	$\cdots$	a_{nm}

Tabela 2.1: Representação de mensagens do Twitter.

Mais especificamente, a Tabela 2.1 representa n *tweets* e m termos[2], sendo cada $tweet_i = (a_{i1}, a_{i2}, \cdots, a_{ij}, \cdots, a_{im})$. O valor a_{ij} refere-se ao valor associado ao $j-$ésimo termo do *tweet* i, ou seja, a_{ij} é o valor do termo t_j no $tweet_i$ e pode ser calculado de diversas formas. Alguns autores utilizam valores binários. Neste caso, $a_{ij} = 1$ significa a presença do termo j na mensagem i e o valor 0 significa a ausência do termo.

[1] Na rotulação por meio de *hashtags* e *emoticons*, admite-se que situações de sarcasmo e ironia não existem.
[2] Neste trabalho, as palavras "termos", "atributos" e "palavras" são usadas indistintamente.

Os desafios em lidar com *tweets*, mencionados no capítulo 1, inspiram diversos estudos em todos os estágios enumerados na Figura 2.1. Neste capítulo são tratadas as estratégias automáticas – estágios 3 a 6 – para o desenvolvimento de algoritmos de classificação de sentimentos em *tweets*. Com foco na validação da hipótese de pesquisa 1 desta tese, inicialmente, na seção 2.1, são apresentados os trabalhos da literatura que empregam métodos de classificação supervisionada de sentimentos em *tweets* e que utilizam apenas um único algoritmo de classificação (*single classifier*) — por conveniência chamados aqui de algoritmos tradicionais de classificação. Na seção 2.2 são apresentados os métodos de análise de sentimentos de *tweets* que fazem uso da agregação (*ensembles*) de vários modelos de classificação.

O leitor interessado em métodos de análise de sentimentos em *tweets* baseados em léxicos (*Lexicon-Based*) pode recorrer aos trabalhos de Turney (2002); Taboada *et al.* (2011); Thelwall *et al.* (2010); Ortega Bueno *et al.* (2013) e Hu *et al.* (2013a). Para métodos baseados em grafos (*Graph-Based*), sugere-se os trabalhos de Speriosu *et al.* (2011); Cui *et al.* (2011); Wang *et al.* (2011) e Tan *et al.* (2011).

2.1 Algoritmos tradicionais de classificação

A maioria dos métodos propostos para lidar com análise de sentimentos em *tweets* (TSA) empregam um único classificador, sendo que os mais utilizados são: Naive Bayes (NB), Máxima Entropia (MaxEnt), *Support Vector Machines* (SVM), Multinomial Naive Bayes (MNB), Regressão Logística (LR) e *Conditional Random Field* (CRF).

Um dos primeiros estudos propostos com o objetivo de realizar análise de sentimentos automática em *tweets* foi o de Go *et al.* (2009), que abordou a TSA como uma classificação binária, categorizando os *tweets* em positivos ou negativos. *Emoticons* (para positivo "`:)`" e para negativo ":(") foram usadas para coletar dados de treinamento com a API do Twitter e rotular os *tweets*. Os dados de treinamento foram pré-processados antes de serem usados para treinar o classificador. O pré-processamento incluiu remoção de letras repetidas (e.g. *huuuuuuuuungry* foi transformado em *hungry*) e remoção dos *emoticons* para não inserir viés (*bias*) nos dados de treinamento. Foram usados unigramas e frequência das palavras nos documentos (TF) como estratégias de representação dos *tweets* em uma *bag-of-words*. Ainda em relação ao trabalho de Go *et al.* (2009), foram feitos experimentos com três técnicas tradicionais de classificação: Naive Bayes, Entropia Máxima e Máquina de Suporte Vetorial (*Support Vector Machines*– SVM). Os melhores resultados foram obtidos com o Naive Bayes. Também foram feitos experimentos com bigramas[3], entretanto a acurácia observada foi menor. Os autores atribuem a queda da acurácia com bigramas à esparsidade dos dados. Além das técnicas citadas, os autores compararam o uso de unigramas com a mesma *bag-of-words* acrescentando técnicas mais sofisticadas de processamento de língua natural (*Natural Language Processing* - NLP) tais como *Part of Speech- POS*[4] para tratar desambiguação de termos, e técnicas para tratar mensagens

[3]Termos compostos por duas palavras. Também conhecidas como 2-grams.

[4]classe sintática das palavras –e.g. *"A menina é bela"*, neste caso o *POS* para cada termo desta frase é *A = artigo*, *menina=substantivo*, *é = verbo* e *bela=adjetivo*

negativas implícitas em expressões como sarcasmo e ironia. Entretanto, tais técnicas de processamento de linguagem não melhoraram os resultados. Em um primeiro momento, os autores não manipularam *tweets* neutros. Como uma segunda parte dos experimentos, foram coletados *tweets* que não continham *emoticons* com o objetivo de serem considerados neutros. Para os testes, foram manualmente anotados 33 *tweets* como neutros. Na sequência, foi treinado um classificador com três classes, e os testes revelaram uma queda na acurácia dos resultados.

Semelhantemente a Go *et al.* (2009), Pak & Paroubek (2010) também utilizaram *emoticons* como rótulos de classe para anotar cerca de 300.000 *tweets*. Pak & Paroubek (2010) abordaram o problema como uma tarefa de classificação multi-classe e cada *tweet* foi classificado como positivo, negativo ou neutro. Os *tweets* neutros foram coletados a partir de contas de jornais e revistas, tais como *New York Times*[5] e *Washington Post*[6]. Os autores também apresentaram a distribuição dos termos e a sua respectiva freqüência depois de uma análise linguística. Eles compararam o desempenho dos algoritmos SVM, NB e CRF utilizando diferentes tipos de atributos, incluindo unigramas, bigramas, n-gramas e a posição dos n-gramas em relação ao texto. Os seus resultados mostraram que a melhor combinação foi Naive Bayes em conjunto com n-gramas e POS. Eles também observaram que o desempenho aumentou com mais *tweets* sendo inseridos no conjunto de treinamento.

Barbosa & Feng (2010) abordaram o problema de TSA com um classificador baseado em duas etapas. A primeira etapa teve como objetivo determinar se a mensagem era opinativa ou não, enquanto que a segunda etapa tinha como meta classificar se o *tweet* com opinião possuía uma conotação positiva ou negativa. Barbosa & Feng (2010) usaram informações de três diferentes ferramentas de detecção de sentimentos desenvolvidas por outros cientistas para anotar uma coleção de *tweets*. *Tweets* classificados com diferentes sentimentos a partir das ferramentas foram removidos, resultando em um conjunto de dados de treinamento de 200.000 *tweets*. Os autores utilizaram POS, o léxico MPQA (Wiebe *et al.*, 2005) e atributos específicos relacionados à rede social obtidos a partir do Twitter, tais como *retweets, hashtags, URLs, emoticons,* etc. Os melhores resultados ocorreram com o classificador SVM, obtendo uma acurácia de $81{,}9\%$ para a detecção de subjectividade e $81{,}3\%$ para a detecção de polaridade. Uma conclusão interessante apresentada pelos autores foi que as características relacionadas à sintaxe providas pelo POS foram importantes para a detecção de subjetividade enquanto que as características relacionadas aos *tweets* foram mais importantes para a detecção de polaridade.

Davidov *et al.* (2010) também apresentaram uma abordagem supervisionada para TSA. Além das *emoticons* usadas para coleta de *tweets* de sentimento para o conjunto de treinamento, eles também usaram *hashtags*. Marcas de pontuação, padrões frequentes, além de n-grams também foram utilizados como atributos e alcançaram um *F-score* de $86{,}0\%$ para classificação binária com um classificador KNN (*K-Nearest Neighbor*).

Um dos classificadores mais utilizados para TSA é o SVM. Bakliwal *et al.* (2012) empregou tal classificador com diferentes técnicas de pré-processamento, uma por uma, a fim de medir a sua eficácia. Normalização de *tweets (Spelling correction*), *stemming* e remoção de *stop-words*

[5]http://www.nytimes.com
[6]https://www.washingtonpost.com

foram aplicados e melhoraram a acurácia do classificador. Os conjuntos de dados de Stanford (Go *et al.*, 2009) e o Mejaj (Bora, 2012) foram usados para avaliar a sua abordagem.

Da mesma forma, Mohammad *et al.* (2013) buscaram resolver o problema de análise de sentimentos em *tweets* utilizando o SVM. Mohammad *et al.* (2013) usaram o conjunto de dados provido pelo SemEval 2013 (Nakov *et al.*, 2013). Eles representavam cada *tweet* como um vetor de atributos incluindo, *n-gramas* baseados em caracteres e baseados em palavras, POS, palavras com todos os caracteres em maiúsculo (como marcas de sentimentos), *hashtags*, léxicos, padrões de pontuação e repetições dos mesmos, *emoticons*, alongamento enfático das palavras e tratamento de negação. Eles observaram que o classificador SVM treinado com todos os atributos mencionados anteriormente obteve melhor desempenho do que os classificadores treinados somente com unigramas.

Um método SVM com kernel linear foi proposto por Kiritchenko *et al.* (2014) para TSA. Os autores utilizaram uma variedade de atributos baseados em estilo de escrita, em sentimentos e em semântica. O SVM com kernel linear superou em desempenho o classificador MaxEnt. Os atributos de sentimentos são inicialmente derivados de um léxico específico para *tweets*, o que segundo os autores contribui positivamente para a adequada captura das palavras de sentimento no contextos negativos.

Um *framework* de classificação contendo três passos foi apresentado por Asiaee T. *et al.* (2012). No primeiro passo, eles identificaram os *tweets* do tópico de interesse. No segundo passo são identificados os *tweets* que possuem sentimento, enquanto, no último passo, os *tweets* foram anotados com polaridade de sentimento. Eles estudaram o desempenho a partir de uma série de métodos clássicos (KNN e NB) e também propuseram novos métodos incluindo SVM ponderado e o aprendizado baseado em dicionário (*Dictionary Learning*). Um resultado enumerado como relevante pelos autores é aquele em que o desempenho da classificação foi melhorado a partir da redução de esparsidade na *bag-of-words*.

A seleção de atributos é muito importante para a eficácia dos métodos supervisionados, uma vez que está intimamente relacionada ao poder preditivo do classificador, como mencionado anteriormente. Neste cenário, vários estudos foram propostos na literatura com o objetivo de analisar o impacto de diferentes atributos sobre TSA (Agarwal *et al.*, 2011; Aisopos *et al.*, 2011; Kouloumpis *et al.*, 2011; Saif *et al.*, 2012b; Hamdan *et al.*, 2013).

Agarwal *et al.* (2011) realizaram um estudo com os diferentes atributos, examinando o seu desempenho em diferentes algoritmos. Após uma extensa análise dos atributos, eles concluíram que as características mais úteis para TSA foram as que se referem à polaridade do sentimento de um termo (léxico). Os autores realizam estudos considerando somente a classificação com duas classes (positiva e negativa) e também avaliam o impacto com a inserção da classe neutra.

Aisopos *et al.* (2011) sugeriram o uso de grafos baseados em n-gramas para melhorar a acurácia da classificação. Os autores usaram dois algoritmos de classificação: MNB e o baseado em árvores de decisão C4.5 (Quinlan, 1993). Os dois classificadores foram treinados em cerca de 1 milhão de *tweets*. A anotação com a classe de sentimento para os dados de treinamento baseou-se na presença ou ausência de *emoticons*. Os experimentos realizados por eles mostraram que o melhor modelo foi inferido a partir do classificador baseado em árvores de decisão

C4.5, treinado em um grafo de 4-gramas. O melhor modelo obteve acurácia de $66{,}77\%$ para classificação binária e $50{,}67\%$ para a classificação em três classes.

Uma abordagem com foco em semântica foi proposta por Saif *et al.* (2012b) que examinou um conjunto de atributos semânticos. Neste cenário, entende-se por atributos semânticos aqueles relacionados a conceitos semânticos (e.g. pessoa, cidade) que representam entidades extraídas (e.g. Steve Jobs, Londres). Saif *et al.* (2012b) fizeram uso de atributos semânticos em um classificador Naive Bayes. Os resultados mostraram um aumento médio na acurácia para identificar sentimentos negativos e positivos de cerca de $6{,}5\%$ e de $4{,}8\%$ usando unigramas e POS, respectivamente.

Hamdan *et al.* (2013) propuseram o uso de muitos atributos e recursos, com o objectivo de alcançar um bom desempenho em tarefas de TSA. Dos atributos analisados, foram incluídos conceitos a partir da DBPedia[7], grupos verbais e adjetivos da WordNet (Miller, 1995a), e características fornecidas pelo léxico SentiWordNet (Baccianella *et al.*, 2010a). Hamdan *et al.* (2013) também utilizaram um léxico de emoções, abreviaturas e gírias para melhorar a acurácia na TSA. O método proposto permitiu melhorar a acurácia média em 2% e 4%, considerando essas características para o SVM e o Naive Bayes treinados com unigramas, respectivamente.

Em vez de aplicar a análise de sentimentos com granularidade de documento, Jiang *et al.* (2011) usaram uma abordagem de aprendizado de máquina para lidar com a tarefa de TSA baseada em aspecto — a abordagem com granularidade baseada em aspectos foi detalhada no capítulo 1. Eles propuseram um método que combina atributos independentes da classe e dependentes da classe. No método proposto por eles, foram usadas regras definidas manualmente para detectar padrões sintáticos que mostram se um termo foi relacionado à um objeto específico. Eles também empregaram um SVM binário para a classificação de subjetividade e polaridade. Além disso, também foram utilizados atributos específicos do Twitter, como por exemplo *retweets*, respostas e menções, que foram usadas para criar um grafo. Tal grafo refletiu as similaridades entre os *tweets* considerados. Jiang *et al.* (2011) relataram uma acurácia média de 85%, considerando as classes positiva, negativa e neutra.

A Tabela 2.2 resume as abordagens do estado da arte mencionadas anteriormente que empregam um único método supervisionado para TSA. A primeira coluna apresenta a referência ao trabalho em questão. A segunda coluna refere-se ao propósito do artigo, sendo que este pode ser *análise de sentimentos com granularidade de tweet* (TSA), em que o sentimento é obtido considerado a análise do *tweet* em sua totalidade ou *análise de sentimentos com granularidade de entidade e de aspecto em tweets* (entidade-TSA). A terceira coluna fornece uma lista completa dos diferentes algoritmos que foram examinados e adotados no estudo relacionado. Os atributos empregados pelos autores (se relatados no estudo) são apresentados na quarta coluna. A última coluna refere-se ao(s) conjunto(s) de dados utilizado(s) no estudo. No caso de os autores terem criado um conjunto de dados que não está disponível publicamente, isto também é mencionado na tabela em questão.

[7] http://wiki.dbpedia.org

Estudo	Tarefa	Algoritmos	Atributos	Conjunto de Dados
Go *et al.* (2009)	TSA	NB, MaxEnt, SVM	Unigramas, bigramas, POS	Stanford
Pak & Paroubek (2010)	TSA	MNB, SVM, CRF	Unigramas, bigramas, trigramas, POS	Stanford
Barbosa & Feng (2010)	TSA	SVM	POS, léxicos e baseado em *tweets*	privado
Davidov *et al.* (2010)	TSA	KNN	pontuação, n-gramas, padrões, baseado em *tweets*	O'Connor *et al.* (2010)
Bakliwal *et al.* (2012)	TSA	SVM,NB	léxicos, *hashtags, emoticons, URLs*	Stanford e Mejaj Bora (2012)
Mohammad *et al.* (2013)	entidade-TSA	SVM	pontuação, n-gramas, padrões, baseado em *tweets*, POS, negação	SemEval
Kiritchenko *et al.* (2014)	TSA	SVM com kernel linear, MaxEnt	pontuação, n-gramas, padrões, baseado em *tweets*, POS, negação	SemEval
Asiaee T. *et al.* (2012)	TSA	*Dictionary Learning*, SVM, NB, KNN	Não é mencionado	privado
Agarwal *et al.* (2011)	TSA	SVM	POS, unigrams, léxicos, exclamação etc.	privado
Aisopos *et al.* (2011)	TSA	MNB, C4.5	POS, n-grams, léxicos, etc	privado
Saif *et al.* (2012b)	TSA	NB	N-gram, POS, semânticos	Stanford, HCR, OMD
Hamdan *et al.* (2013)	TSA	SVM, NB	unigramas, conceitos (DBPedia), verbos /adjetives (WordNet, SentiWordNet)	SemEval
Jiang *et al.* (2011)	entidade-TSA	SVM	unigramas, *emoticons, hashtags*, pontuação	privado

Tabela 2.2: Resumo dos trabalhos que empregam um único método supervisionado para TSA.

2.2 Agregadores de Classificadores

Em geral, encontrar o algoritmo de classificação cujas propriedades são adequadas para o domínio em questão é fundamental para o sucesso da aplicação. Em análise de sentimentos em *tweets* não há um consenso sobre qual configuração experimental (quais atributos e quais metodologias de classificação) é a que provê melhor desempenho – a Tabela 2.2 reflete este cenário. É prática comum em análise de sentimentos em *tweets* testar diversos algoritmos e escolher o melhor de acordo com algum critério de qualidade do modelo obtido (por exemplo, taxa de erro). Dado que tal prática requer que diversos classificadores sejam construídos, informações valiosas sobre os dados podem ser desprezadas ao se escolher apenas o melhor classificador dentre aqueles obtidos. Agregadores de classificadores, por sua vez, procuram aproveitar tal variedade de modelos, combinando-os de tal forma que soluções melhores, e mais robustas, possam ser obtidas.

Os agregadores de classificadores, métodos também conhecidos como *ensembles*, usam múltiplos classificadores para resolver o mesmo problema, explorando a idéia de que uma coleção de diferentes classificadores, referindo-os individualmente como classificadores *base*, podem oferecer informações complementares com relação aos padrões que serão classificados, melhorando a eficácia de todo o processo de aprendizado (Zhou, 2012). A combinação de múltiplos classificadores para gerar um classificador consolidado tem sido uma área de pesquisa ativa nas últimas décadas[8] (Kuncheva, 2004; Dietterich, 2000; Rokach, 2010; Zhou, 2012; Zhang & Ma, 2012; Schapire & Freund, 2012). Diversos trabalhos têm fornecido resultados teóricos e evidências empíricas sobre a utilidade de tais abordagens para resolver problemas de classificação difíceis — *e.g*, Hansen & Salamon (1990) e Oza & Tumer (2008). Além disso, a extensa literatura sobre o assunto tem mostrado que um agregador obtido a partir de classificadores diversificados é tipicamente melhor do que seus componentes individuais (Kuncheva, 2004; Dietterich, 2000; Hansen & Salamon, 1990).

O princípio, bem como a importância, de se agregar classificadores distintos reside no fato de que, por exemplo, dada a agregação de três classificadores (h_1, h_2 e h_3) e um novo objeto $\mathbf{x}_i$ a ser classificado, se os três classificadores base são idênticos, então quando $h_1(\mathbf{x}_i)$ erra a predição, $h_2(\mathbf{x}_i)$ e $h_3(\mathbf{x}_i)$ também erram. Contudo, se os erros dos três classificadores não estão correlacionados, então, quando $h_1(\mathbf{x}_i)$ está errado, $h_2(\mathbf{x}_i)$ e $h_3(\mathbf{x}_i)$ podem estar certos de tal forma que por meio de *voto majoritário* o novo objeto $\mathbf{x}_i$ poderá ser corretamente classificado (Dietterich, 2000; Hansen & Salamon, 1990). Por ser uma prática simples e normalmente aplicável a qualquer tipo de classificador, muitos trabalhos levam em conta essa estratégia geral (Oza & Tumer, 2008; Bauer & Kohavi, 1999; Breiman, 1996). Diversas estratégias podem ser consideradas para introduzir diversidade nos classificadores base. Como exemplo, *bagging* (Breiman, 1996) faz uso de diversos modelos — cada um induzido por um conjunto de treinamento obtido por meio de *bootstrap* (Efron & Tibshirani, 1993) — para classificar um objeto não rotulado usando-se voto majoritário. Outro conceito, chamado de *boosting* (Freund & Schapire, 1997), atribui pesos na predição de cada modelo levando-se em conta acurácias de classificações passadas — por isso, essa abordagem é considerada iterativa (cada novo modelo é influenciado pelo desempenho daqueles previamente construídos). Outras abordagens conhecidas são *stacking* (Wolpert, 1992), *model averaging* (Perrone & Cooper, 1993) e *forecast combining* (Clemen, 1989).

Dietterich (2000) listou três razões para o uso de *ensembles* de classificadores:

1. *Estatística:* Um algoritmo de aprendizado pode ser visto como um algoritmo de busca no espaço de hipóteses $\mathcal{H}$ com o objetivo de identificar a melhor hipótese. O problema estatístico aparece quando a quantidade de dados disponíveis para treinamento é muito pequena comparada ao tamanho do espaço de hipóteses. Sem dados suficientes, o algoritmo de aprendizado pode encontrar muitas hipóteses diferentes em $\mathcal{H}$. Assumindo que se possa ter um classificador para cada hipótese, se um único classificador é escolhido

[8] O International Workshop on Multiple Classifiers Systems (http://mcs.diee.unica.it) tem acontecido anualmente desde 2003.

dentre os possíveis, existe um risco de que este não possua a melhor capacidade de generalização ou acurácia. Entretanto, com a combinação dos classificadores disponíveis, o risco de seleção de um classificador inadequado é menor (Kuncheva, 2004);

2. *Computacional:* Muitos algoritmos de aprendizado realizam alguma busca local a qual pode encontrar um ótimo local, impedindo que o algoritmo encontre a melhor hipótese. Um *ensemble* construído a partir da execução de um algoritmo de busca local que a cada iteração parte de diferentes pontos, pode fazer com que se encontre uma melhor aproximação do que considerando somente classificadores individuais;

3. *Representacional:* Pode acontecer que o espaço do classificador não contenha o classificador ótimo. Por exemplo, se o classificador ótimo é não-linear, mas o classificador escolhido é linear, este não será capaz de encontrar a solução ótima do problema. No entanto, um conjunto de classificadores lineares pode aproximar qualquer fronteira de decisão. Certos problemas são muitos difíceis para um único classificador. Às vezes, a fronteira de decisão que separa dados de diferentes classes pode ser muito complexa e uma combinação apropriada de classificadores pode fazer possível essa separação.

Algoritmos de aprendizagem tradicionais falham nestas três questões (estatística, computacional e representacional), enquanto métodos de combinação de classificadores podem minimizar tais problemas.

Rokach (2010) enfatiza que estudos experimentais realizados pela comunidade de aprendizagem de máquina mostram que combinando as saídas de vários classificadores é possível reduzir o erro de generalização. Rokach (2010) ainda reitera que os métodos baseados em agregadores de classificadores têm se mostrado eficientes em diversos domínios tais como aplicações financeiras (Lei, 2002), bioinformática (Tan *et al.*, 2003), saúde (Man, 2004), indústria (Maimon & Rokach, 2004), geografia (Bru, 2004), principalmente devido ao fato de vários tipos de classificadores tenderem a ter diferentes *biases* indutivos.

Kuncheva (2004) enumera quatro principais oportunidades de pesquisa relativas à construção de agregadores de "classificadores diversos", ou seja, classificadores que podem apresentar diferentes margens de decisão, com erros e acertos independentes. Na Figura 2.2, são ilustrados tais focos de pesquisa em *ensembles* de classificadores: em (i), o foco é a construção de *ensembles* através da combinação dos classificadores base, que pode ser feita utilizando estratégias como voto majoritário, unanimidade, média das probabilidades das classes inferidas pelos modelos de base, dentre outras; em (ii), o foco são estratégias para a escolha dos classificadores utilizados no *ensemble*; em (iii), o foco é a definição dos atributos que irão compor cada uma das *bag-of-words* de treinamento; e, por fim, em (iv), o foco é a escolha dos exemplos de treinamento que irão ser usados nos classificadores. Em cada um dos quatro objetivos de pesquisa listados por Kuncheva (2004) há a oportunidade de maximizar a diversidade entre os seus classificadores base, com o objetivo de maximizar também o poder preditivo do modelo final.

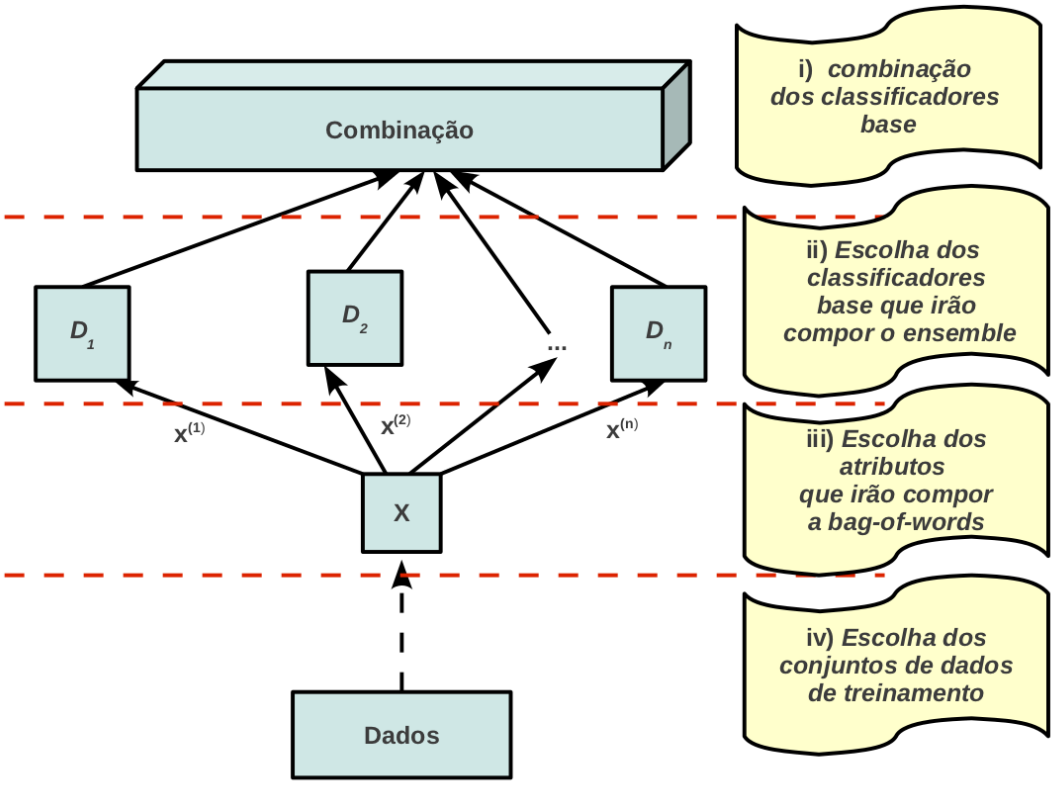


Figura 2.2: Abordagens para construir *ensembles* de classificadores.

É importante enfatizar que não há garantia que a combinação de múltiplos classificadores sempre produzirá melhores resultados que os seus classificadores individuais tidos como base, exceto em alguns casos especiais (Fumera & Roli, 2005). Em outras palavras, a agregação de classificadores não necessariamente terá o melhor desempenho se comparado ao melhor classificador base pertencente ao *ensemble*, entretanto ele certamente reduzirá o risco de realizar a pior seleção dentre os inúmeros classificadores possíveis. A agregação de classificadores é tratada mais detalhadadamente em Oza & Tumer (2008); Bishop (2007) e Kuncheva (2004).

Efetivamente, para o correto uso de *ensembles*, os indivíduos componentes do mesmo devem ter entre si algum nível de diversidade (Tumer & Ghosh, 1996; Krogh & Vedelsby, 1995; Kuncheva, 2004; Kuncheva & Whitaker, 2003b), como já mencionado anteriormente. Classificadores diferentes produzem diferentes fronteiras de decisão. Assim sendo, se a diversidade é efetivamente constatada, erros independentes são produzidos por cada classificador, e ao combiná-los, o erro total será reduzido. A Figura 2.3, adaptada de Polikar (2009), ilustra este conceito para a configuração proposta no domínio de aplicação desta tese: cada classificador (treinado em diferentes subconjuntos do conjunto de treinamento considerado) produz diferentes erros, e a combinação de classificadores pode prover uma fronteira de decisão com maior poder preditivo. De fato, o *erro bayesiano* pode ser estimado a partir do *ensemble* de classificadores (Tumer & Ghosh, 2003). As Figuras 2.4 e 2.5 apresentam exemplos de regras de combinação de classificadores.

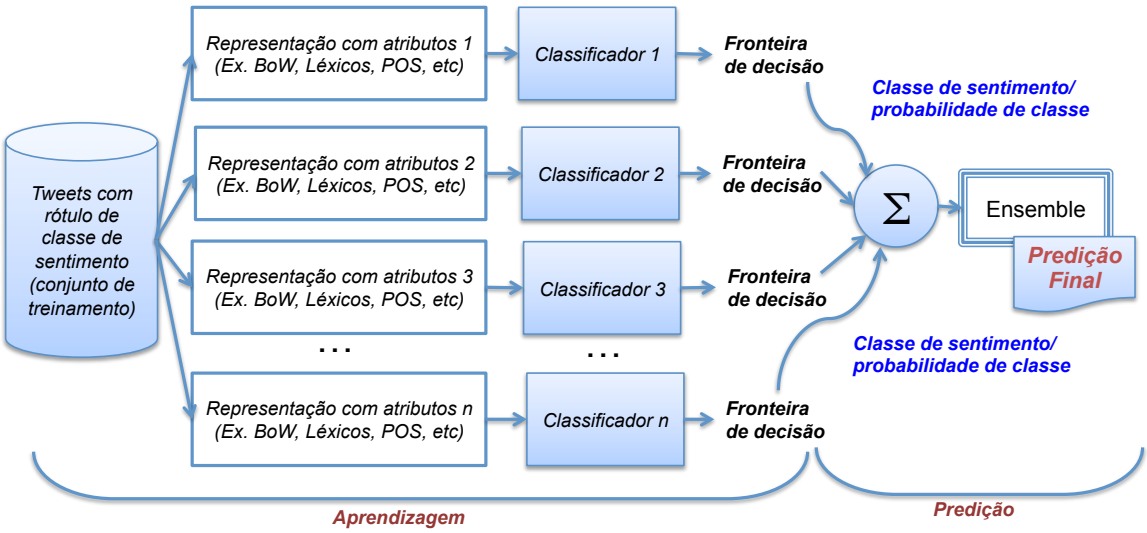


Figura 2.3: *Ensembles* de classificadores para análise de sentimentos em *tweets*. Σ refere-se a regra de combinação (e.g., voto majoritário e média das probabilidades de classe) para os classificadores base.

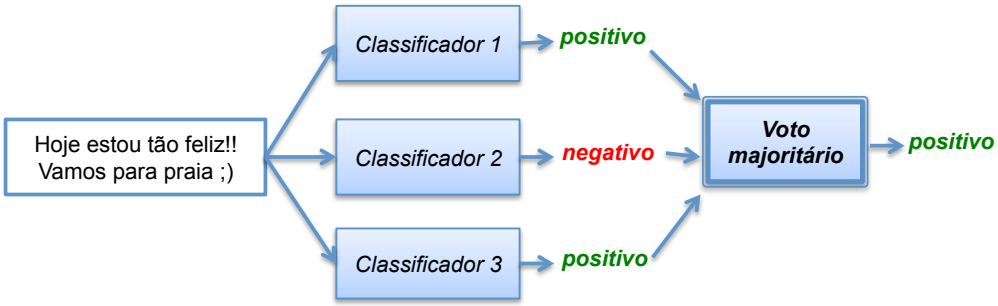


Figura 2.4: Um exemplo de voto majoritário como regra de combinação. Neste caso, a maioria dos classificadores concordam que a classe é positiva.

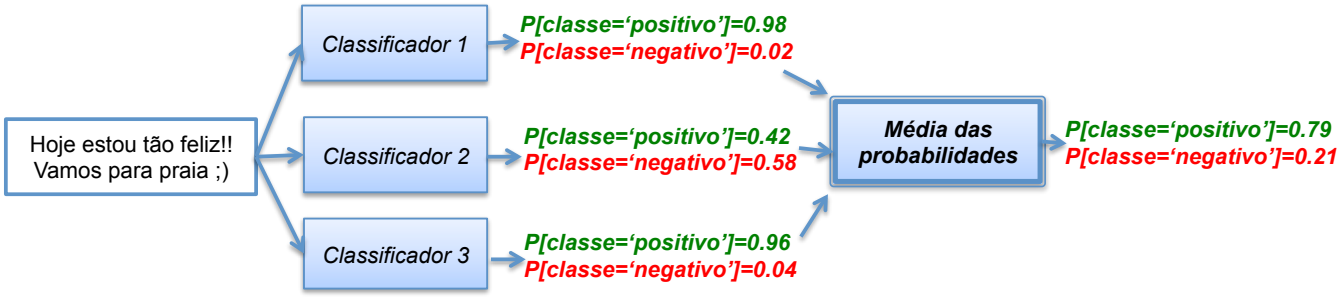


Figura 2.5: Um exemplo de média das probabilidades como regra de combinação de classificadores. Neste caso, a probabilidade P(classe=Positivo$/$*tweet*)>P(class=Negativo$/$*tweet*), então a saída do *ensemble* é positiva.

Ensembles de classificadores para análise de sentimentos em *tweets* têm sido pouco explorados na literatura — algumas poucas exceções são Lin & Kolcz (2012); Clark & Wicentwoski (2013); Rodrıguez-Penagos *et al.* (2013); Hassan *et al.* (2013); Kanakaraj & Guddeti (2015); Hagen *et al.* (2015).

Lin & Kolcz (2012) exploram o uso de agregadores de classificadores em um ambiente integrando com ferramentas de processamento e armazenamento paralelo, tais como *MapReduce* e *Hadoop* (Dean & Ghemawat, 2008). Os dados foram coletados usando *emoticons positivas* e *emoticons negativas* como termos de consulta. Para o treinamento, foram obtidas três bases de dados contendo 1 milhão, 10 milhões e 100 milhões de exemplos de treinamento respectivamente, com igual número de *tweets* positivos e *tweets* negativos. O conjunto de teste também foi coletado usando o mesmo método com *emoticons* e é constituído de 1 milhão de *tweets*, com no mínimo 20 caracteres para cada mensagem. Os dados de treinamento foram particionados e o *ensemble* foi construído combinando diversos modelos de predição inferidos a partir das várias partições do conjunto de treinamento — via *regressão logística*. Os testes foram realizados variando o número de modelos de predição componentes do *ensemble* e o número de exemplos do conjunto de treinamento. Os resultados da abordagem usando *ensembles* foram comparados à abordagem tradicional de classificação utilizando o algoritmo *regressão logística* para inferência de um único modelo de predição com a base de dados de treinamento completa, e a acurácia obtida apresentou uma melhora significativa. O trabalho apresentado por Lin & Kolcz (2012) apresenta resultados iniciais, propondo que a diversidade dos modelos de classificação seja obtida a partir dos modelos treinados em diferentes conjuntos de *tweets*.

Paralelamente a esta tese, Rodrıguez-Penagos *et al.* (2013) e Clark & Wicentwoski (2013) propuseram o uso de *ensembles* para análise de sentimentos com granularidade de aspectos. Nesta perspectiva, o rótulo de classe (positiva, negativa ou neutra) é fornecido para uma frase ou palavra específica dentro do *tweet*, e não necessariamente ao *tweet* como um todo. Clark & Wicentwoski (2013) usam uma combinação de classificadores *Naive Bayes*, em que cada classificador é treinado em um conjunto de atributos diferentes dos demais atributos usados pelos outros classificadores. Os classificadores são combinados em um esquema de voto majoritário. Os classificadores são submetidos a um conjunto de dados diferente do conjunto de treinamento para avaliar sua acurácia. Somente os classificadores avaliados com um valor de acurácia mínimo (definido previamente) são submetidos ao *ensemble*.

Hassan *et al.* (2013) propuseram lidar com o desbalanceamento de classes, com a esparsidade e representatividade dos dados utilizados no treinamento. Os autores realizam um enriquecimento do *corpus* através de múltiplos conjuntos de dados adicionais também relacionados à classificação de sentimentos (todos os conjuntos são rotulados com as classes positivo, negativo e neutro). Os autores utilizam *unigramas, bigramas, part-of-speech,* features semânticas derivadas de WordNet (Miller, 1995b) e léxicos derivados da SentiWordNet 3.0 (Baccianella *et al.*, 2010b). Também são empregadas técnicas de sumarização como Legomena e reconhecimento de entidades nomeadas. Os autores propõem um *ensemble* diversificado, variando os classificadores de base, os atributos de cada classificador de base, os conjuntos de dados utilizados para enriquecimento, bem como a combinação dos classificadores de base feita por voto majoritário, média das probabilidades, probabilidade máxima etc.

Recentemente, Kanakaraj & Guddeti (2015) realizaram um experimento em que avaliaram vários métodos de agregação de classificadores em bases de dados de *tweets* coletados pelos próprios autores, concluindo que o método de agregação baseado em árvores apresentou me-

lhores resultados.

Outro trabalho recente foi o proposto por Hagen *et al.* (2015), em que os autores são participantes de uma competição (Rosenthal *et al.*, 2014). Tal competição tem como objetivo estabelecer comparações quanto a acurácia na análise de sentimentos, uma vez que são fixados vários conjuntos de *tweets* comuns a todos participantes. Hagen *et al.* (2015) propuseram um método baseado em *ensembles*, cujos classificadores de base são três métodos propostos por outros participantes da competição. Os mesmos fizeram análises de erro em cada um dos três classificadores de base escolhidos, e identificaram que tais classificadores obtiveram erros diferentes quando submetidos aos conjuntos de teste. Tais análises de erro são fortes indícios de que os classificadores são independentes entre si, e portanto, que existe diversidade entre eles. Tais classificadores foram combinados entre si por meio da média das probabilidades de classe.

A Tabela 2.3 resume as abordagens que empregam agregadores de classificadores para resolução do problema de análise de sentimentos em *tweets* (TSA).

Estudo	**Ano**	**Atributos**	**Classificador Base**	**Método *Ensemble***	**Banco de Dados**
Lin & Kolcz (2012)	2012	*feature hashing*	regressão logística	voto majoritário	privado
Rodrıguez-Penagos *et al.* (2013)	2013	n-gram, lexicos, POS, atributos baseados na escrita, SentiWordnet	CRF, SVM e método heurístico	voto majoritário, *upper bound*, *ensemble vote*	Nakov *et al.* (2013)
Clark & Wicentwoski (2013)	2013	n-gram, lexicos	Naive Bayes	voto ponderado	Nakov *et al.* (2013)
Hassan *et al.* (2013)	2013	unigrams, bigrams, POS and WordNet Miller (1995b) and SentiWordNet 3.0 Baccianella *et al.* (2010b)	Redes neurais, *Random Tree, REP Tree*, Naive Bayes, Redes bayesianas, Regressão Logística e SVM.	Um modelo que testa várias combinações de características, classificadores e seus parâmetros	*Sanders - Twitter Sentiment Corpus*[3]
Kanakaraj & Guddeti (2015)	2015	unigrams	–	Random Forest, AdaBoost e Bagging	privado
Hagen *et al.* (2015)	2015	Os mesmos atributos de Mohammad *et al.* (2013); Günther & Furrer (2013); Proisl *et al.* (2013)	Classificadores propostos por Mohammad *et al.* (2013); Günther & Furrer (2013); Proisl *et al.* (2013)	Probabilidade média de classe dos três classificadores.	(Nakov *et al.*, 2013; Rosenthal *et al.*, 2014)

[1] http://en.wikipedia.org/wiki/List_of_emoticons

[2] http://www.noslang.com/

[3] http://www.sananalytics.com/lab/twitter-sentiment/

Tabela 2.3: Estudos com agregadores de classificadores.

2.3 Experimentos

Brown *et al.* (2005) sugerem três métodos para criação de *ensembles* de classificadores com diversidade: (i) variar as inicializações para os algoritmos de classificação de base; (ii) variar os conjuntos de treinamentos para seus classificadores de base; e (iii) variar os classificadores de base ou as estratégias de combinação dos classificadores base. Os experimentos realizados nesta tese foram inspirados por (iii), enquanto que Lin & Kolcz (2012) e Clark & Wicentwoski (2013)

abordaram a diversidade segundo (ii). Rodrıguez-Penagos *et al.* (2013) realizaram experimentos praticando a análise de sentimentos com granularidade de aspectos, aplicando a diversidade segundo o conjunto de treinamento para seus classificadores base.

A hipótese 1 de pesquisa desta tese é que, classificadores de sentimentos para *tweets* com alto poder de predição e acurácia podem ser obtidos a partir da combinação de classificadores diversificados (*ensembles*). Esta hipótese se baseia na filosofia de uso de agregadores de classificadores e presume que a engenharia seja realizada de forma apropriada.

A Figura 2.6 apresenta uma visão geral da abordagem utilizada no cenário de análise de sentimentos nesta tese. Os classificadores de base considerados são *Random Forest, Support Vector Machines* (SVM), *Multinomial Naive Bayes* e Regressão Logística. Embora outros classificadores pudessem ter sido escolhidos, os representantes adotados têm sido comumente estudados e amplamente utilizados na prática e, portanto, têm credibilidade para funcionarem como uma prova de conceito.

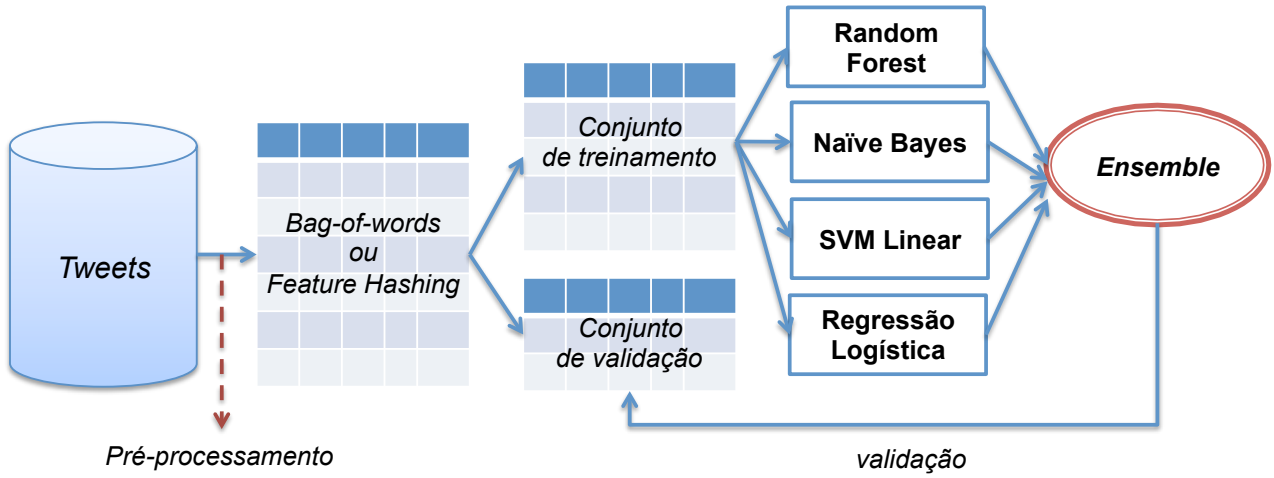


Figura 2.6: Visão geral da abordagem proposta.

Na prática, classificadores são construídos para analisar os dados cuja classe é desconhecida previamente. Este conjunto de dados é comumente conhecido como conjunto alvo ou *target set*. Em um ambiente experimental controlado, como no abordado nesta seção, um conjunto de validação ou conjunto de testes representa o conjunto alvo estabelecido. Nesta tese, e seguindo a boa prática, o conjunto teste/ validação não foi utilizado no processo de construção de *ensembles* de classificadores. Uma vez realizado o processo de treinamento dos classificadores de base, um *ensemble* é formado pela (i) média das probabilidades de classe obtidas por cada classificador de base ou pela (ii) votação por maioria.

2.3.1 Representação e atributos

Como já mencionado anteriormente, em análise de sentimentos em *tweets* não há um consenso sobre quais atributos irão prover o melhor desempenho. Desta forma, com o objetivo de validar a primeira hipótese, foram escolhidas duas técnicas para a representação de atributos: *bag-of-words* com unigramas e *feature hashing*. A segunda técnica foi escolhida por sua ca-

racterística promissora em outros domínios de aplicação (Weinberger *et al.*, 2009; Shi *et al.*, 2009; Ganchev & Dredze, 2008; Forman & Kirshenbaum, 2008; Caragea *et al.*, 2011) e pouca exploração no contexto de análise de sentimentos — apenas o trabalho de Lin & Kolcz (2012) fez uso de *feature hashing* — em *tweets* não disponíveis publicamente. Além disso, nas duas formas de representação foram utilizados um léxico de sentimentos com objetivo de avaliar o poder de enriquecimento promovido pelo mesmo.

Nos experimentos realizados nesta tese, foram avaliados agregadores de classificadores com quatro variações de conjuntos de atributos a saber:

1. Considerando *bag-of-words* com unigramas somente.
2. Considerando *bag-of-words* com unigramas e léxicos provenientes de Liu (2012).
3. Considerando *feature hashing*.
4. Considerando *feature hashing* e léxicos provenientes de Liu (2012).

Os atributos empregados são apresentados com maiores detalhes a seguir.

2.3.1.1 *Bag-of-words*

O modelo de representação *bag-of-words* (BoW), um dos mais utilizados em diversas aplicações em aprendizagem de máquina (Forman, 2003; Sebastiani, 2002), é também amplamente estudado e praticado em análise de sentimentos (Pang *et al.*, 2002; Dave *et al.*, 2003; Kim & Hovy, 2006; Kouloumpis *et al.*, 2011; Lin & Kolcz, 2012; Ghiassi *et al.*, 2013; Zhu *et al.*, 2014). Esta representação é construída considerando *n-grams*, uma sequência contínua de *n* itens a partir de uma fonte textual ou falada. Em particular, uma BoW considera individualmente as palavras (*grams*) de um *tweet* como atributos, assumindo uma independência entre tais palavras. O texto é representado como uma coleção não ordenada de palavras e convertido em um vetor de características ou atributos. Há duas possíveis formas de mensurar a frequência das palavras (*grams*) em um *tweet*: (i) presença do termo ou *gram* (*tp*), calculada conforme equação 2.1; (ii) frequência do termo ou *gram* (*tf*), calculada conforme equação 2.2. De modo geral, em recuperação de informação ou classificação de textos em relação à tópicos, a frequência de *n-grams tf* tem mostrado melhores resultados, enquanto, em classificação de sentimentos, Pang *et al.* (2002) mostraram que o uso da informação sobre presença de termos, *tp*, proporciona melhores resultados. Esta tese assume o uso da informação sobre presença de termos (Pang *et al.*, 2002) como sendo a solução que provê maior eficácia.

$$tf(palavra,tweet) = |\{palavra \in tweet\}| \tag{2.1}$$

$$tp(palavra,tweet) = \begin{cases} 1 \; se \; palavra \in tweet \\ 0 \; se \; palavra \notin tweet \end{cases} \tag{2.2}$$

Um dos desafios da representação BoW é a escolha das palavras que serão consideradas caraterísticas (atributos). Usando este modelo, o *tweet* "*Eu amo chocolate*" pode ser representado

pelo vetor de características: $\vec{V_0} = \{$ 'eu':1, 'amo':1, 'chocolate':1$\}$. Naturalmente algumas palavras não expressam nenhum sentimento, e por isso não possuem valor semântico para a análise de sentimentos, podendo até mesmo confundir o modelo preditivo. Neste caso, palavras conhecidas como *stop-words* são removidas na etapa de pre-processamento.

2.3.1.2 Léxicos de opiniões

Existem palavras ou mesmo frases que são importantes indicadores de sentimento, sendo conhecidas como *palavras de opinião*. Tais palavras são usadas para indicar sentimentos positivos ou negativos. Por exemplo, as palavras *bom, maravilhoso, esplêndido* são palavras de sentimento positivo, enquanto as palavras *ruim, pobre, terrível* são palavras de sentimento negativo. Além de palavras individuais de sentimento, existem também frases e expressões idiomáticas, por exemplo, a expressão "acabar em pizza", que tem uma conotação negativa. Uma lista destas palavras ou frases são chamadas léxicos de sentimentos, léxicos de opinião ou palavras de polaridade (em inglês *sentiment lexicon* ou *opinion lexicon*).

Embora as palavras e frases de sentimento sejam importantes para análise de sentimento, o uso de somente tais atributos não são suficientes porque:

- Uma palavra de sentimento positivo ou negativo pode ter orientações opostas em diferentes contextos. Por exemplo, a palavra "doce" geralmente indica um sentimento positivo tal como em "Uma doce melodia preenchia o ambiente.", mas também pode implicar em um sentimento negativo, por exemplo, "Este café está muito doce.".

- Uma frase contendo léxicos de opinião pode não expressar nenhum sentimento. Este fenômeno ocorre com freqüência em vários tipos de frases. Sentenças interrogativas e sentenças condicionais são dois tipos importantes. Por exemplo, "Você pode me dizer qual celular Samsung é bom?" e "Se eu encontrar um bom celular na loja, vou comprá-lo.". Ambas as frases contêm a palavra de sentimento "bom", mas não expressam uma opinião positiva ou negativa sobre qualquer celular em específico. No entanto, algumas sentenças condicionais ou frases interrogativas expressam sentimentos, por exemplo, "Alguém sabe como consertar essa impressora horrível?" e "Se você estiver procurando um bom carro, compre um Honda Civic".

- Sentenças com sarcasmo são difíceis de lidar. Por exemplo, "Que celular maravilhoso! Parou de funcionar em dois dias". Expressões com sarcasmo são mais comuns em discussões políticas e certos nichos de usuários.

- Muitas frases sem palavras de sentimento podem conter opiniões.

Os léxicos de opinião podem ser criados de duas formas (Taboada *et al.*, 2011): manual ou automaticamente usando palavras pré definidas como sementes (*seeds*) para expandir a lista de palavras. Alguns exemplos de trabalhos nessas duas categorias são citados brevemente a seguir:

Em Nielsen (2011a), é feita uma classificação de sentimentos em mensagens do Twitter avaliando o uso de um léxico de palavras construído manualmente para textos longos, e

avaliando também o mesmo método de classificação usando um léxico construído especialmente para textos do Twitter.

No trabalho de Mostafa (2013), foi apresentado um estudo com 2.105 *tweets* aleatoriamente coletados a partir de dezesseis companhias aéreas comerciais, com o objetivo de avaliar o sentimento de seus clientes. Foi usado um léxico de palavras de opinião com cerca de 6.800 adjetivos com orientação sentimental conhecida.

Bollen *et al.* (2011) investiga se o sentimento público expressado em larga escala diariamente em mensagens divulgadas pelos usuários do Twitter pode ser usado para predizer quedas ou altas no mercado de ações. Para a tarefa foi utilizado um léxico de 2.718 palavras positivas e 4.912 palavras negativas.

Existem pesquisas focadas em usar sinais emocionais que poderiam estar correlacionados com a polaridade do sentimento para um documento ou para as palavras que formam um documento. Por exemplo, quando a comunicação ocorre no mundo real, é comum as pessoas usarem expressões faciais e gestos para expressarem seus sentimentos. Similarmente em redes sociais, usuários desenvolveram símbolos que estão fortemente relacionados com suas emoções (Liu *et al.*, 2012). Estes símbolos são conhecidos como *emoticons* ou expressões faciais que denotam emoções. O trabalho de Hu *et al.* (2013a) utiliza como sinais emocionais as *emoticons* e o conceito de *teoria da consistência da emoção* que sugere que as palavras que frequentemente co-ocorrem têm a mesma orientação em relação ao sentimento.

Há ainda trabalhos (e.g. Brody & Diakopoulos (2011)) que associam o sentimento ao tamanho da palavra, enquanto outras pesquisas buscaram aproveitar as emoções conhecidas sobre um determinado domínio para facilitar a análise de sentimentos em outro domínio distinto (Li *et al.*, 2010).

Como exemplos de léxicos disponíveis na literatura é possível citar[9]:

- *Opinion Finder lexicon* criado por Wilson *et al.* (2005), possui palavras rotuladas como positivas e negativas;
- *AFINN* apresentado em Nielsen (2011b), possui palavras positivas com valores variando de $+1$ a $+5$ e palavras negativas com valores variando de -1 a -5. O léxico completo possui 2.477 palavras;
- Liu (2012) criou um léxico conhecido como *Bing Liu Lexicon* e o usou em diversas pesquisas. Este léxico contém 2.006 palavras rotuladas como positivas e 4.683 palavras rotuladas como negativas;
- O *NRC Hashtag Sentiment Lexicon* é automaticamente criado usando uma coleção de 775.310 *tweets* com hashtags positivas ou negativas. A partir disso os *tweets* são rotulados como positivos ou negativos, considerando sua respectiva polaridade (baseada na *hashtag*). É atribuído a cada palavra um valor variando de -5 a $+5$. Este léxico foi criado pelo time de Mohammad *et al.* (2013), que ganhou a competição conhecida como *SemEval task* (Nakov *et al.*, 2013) para análise de sentimentos em *tweets*;

[9]Todos os léxicos mencionados foram construídos para a língua inglesa.

- *Sentiment140 Lexicon* foi também proposto por Mohammad *et al.* (2013), o qual usa emoticons positivas e negativas para rotular palavras de sentimentos.

Nesta tese, como prova de conceito, foi utilizado o léxico proveniente de Liu (2012). É importante enfatizar que os algoritmos supervisionados surgem como alternativa para uma maior generalização em relação aos métodos que se baseiam somente em léxicos, buscando encontrar padrões que suavizem os problemas evidenciados com as abordagens baseadas em um conjunto de termos pré-definido e estático.

2.3.1.3 Redução de esparsidade — *Feature Hashing*

Um dos principais problemas de acurácia e eficiência temporal evidenciados na análise de sentimentos em *tweets*, quando realizada a partir da representação baseada em *bag-of-words*, é ocasionada devido à esparsidade. Tal problema é justificado, uma vez que *tweets* são fontes textuais limitadas a 140 caracteres e com uma imensa variabilidade de termos. Evidências em trabalhos cujos focos são a classificação textual baseada em tópicos e em classificação de sequências de proteína demostram que *hashing* (Weinberger *et al.*, 2009; Shi *et al.*, 2009; Ganchev & Dredze, 2008; Forman & Kirshenbaum, 2008; Caragea *et al.*, 2011) é uma estratégia para redução de dimensionalidade não paramétrica eficiente. Apenas o trabalho de Lin & Kolcz (2012) utiliza a técnica de feature hashing como estratégia de redução de dimensionalidade para análise de sentimentos em *tweets*. Desta forma, tal estratégia será melhor investigada nesta tese.

Em particular, para a classificação de sentimentos em *tweets*, *feature hashing* reduz o número de atributos providos como entrada para o algoritmo de aprendizagem. O espaço original com alta dimensão é "reduzido" por *hashing* em um espaço de menor dimensão, i.e., mapeando as *features* em *hash keys*. A Figura 2.7 demonstra como é realizada a aplicação da técnica de *feature hashing* ao *tweet @AnaJulia: "Eu amo chocolate #lacta!"*. No quarto passo, é aplicada a função *hash* da Equação 2.3, a qual recebe um termo s de tamanho l como parâmetro e retorna a soma dos valores ASCII de seus caracteres(c_i).

$$h(s) = \sum_{i}^{l} ASCII(c_i)/10 \tag{2.3}$$

2.3.2 Pré-processamento dos dados

A fim de eliminar ruídos e termos sem qualquer significado semântico para a classificação, bem como realizar uma normalização dos *tweets*, foram realizados algumas etapas de pré-processamento:

- Remoção de menção de pessoas;
- Remoção de pontuação, números e acentuação;
- Remoção de *links e URLs*;

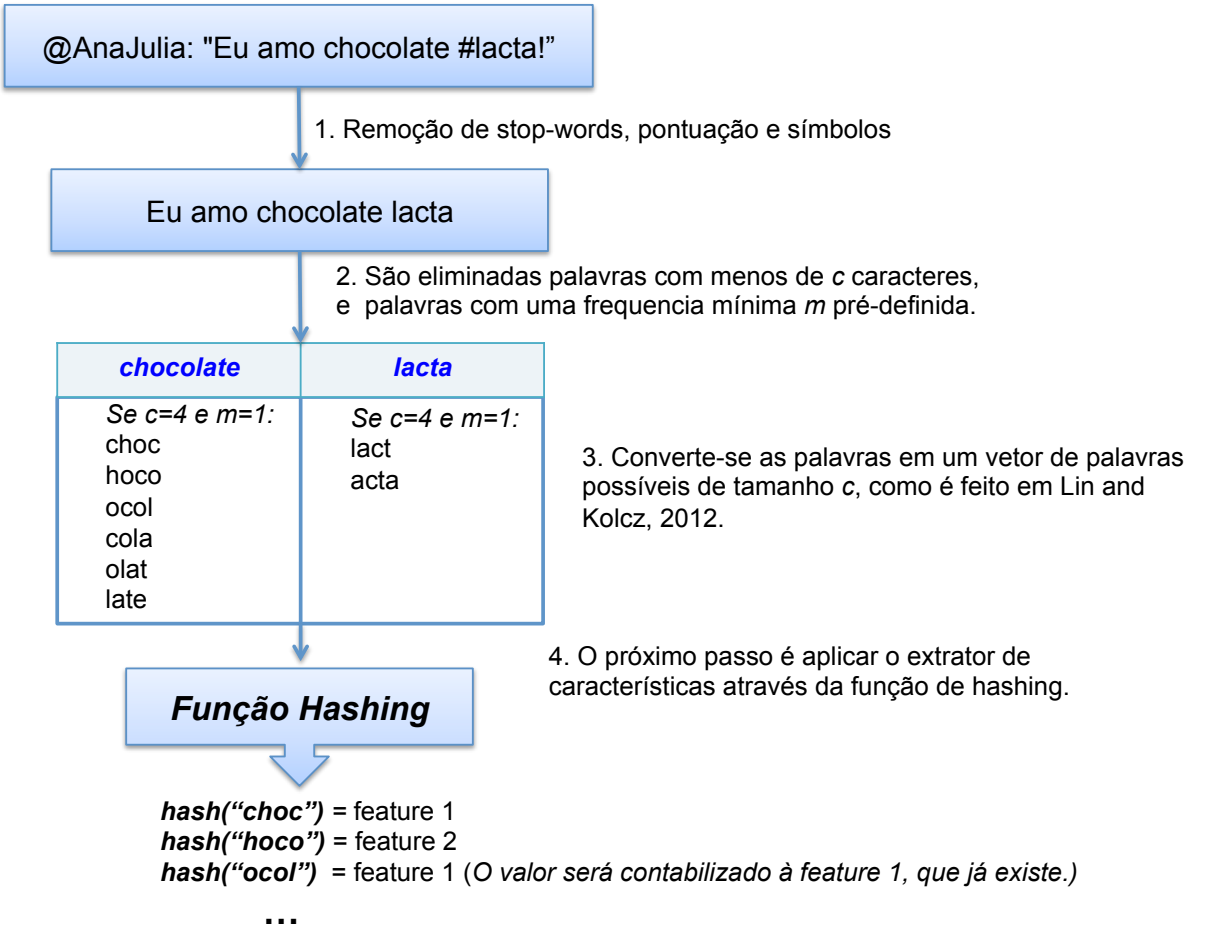


Figura 2.7: Aplicando *feature hashing* em um *tweet*.

- Remoção de *stop words* (ver Figura 2.8) da língua inglesa, uma vez que os conjuntos considerados nos experimentos são dessa natureza;

- Padronização do texto e conversão em minúsculo;

- A *bag-of-words* foi construída com frequência binária, e um termo foi considerado frequente se ele ocorre em mais de um *tweet*;

- Também foi usada a técnica conhecida como *stemming*;

  Stemming Os documentos geralmente contêm diferentes formas de uma palavra, como por exemplo *caminhada, caminhadas, caminhar*. O objetivo do *stemming* é o de reduzir as formas flexionadas de uma palavra para a forma de base comum.

- O léxico proposto por Liu (2012) também foi utilizado, bem como as *emoticons* apresentadas na Tabela 2.4, a fim de trazer novos atributos e enriquecer o modelo de classificação. O número de *emoticons* positivas e o número de *emoticons* negativas existentes em cada *tweet* foram utilizados como atributos, bem como o número de itens léxicos positivos e negativos.

a	*an*	*and*	*are*	*as*	*at*
be	*by*	*for*	*from*	*has*	*he*
in	*is*	*it*	*its*	*on*	*that*
the	*to*	*was*	*were*	*will*	*with*

Figura 2.8: Uma lista com palavras conhecidas como *stop-words*.

Símbolos positivos	Símbolos negativos
:-), :), (-:, (:, B-), ;-), ;),(^_^), (^_~), $:'-), :-], :-d, :>, :d, ;-d, ;>, ;], =)	#-|, (:-, (:-&, (:-(, (t_t), 8-(, 8o|, :$, :'(, :'-(, :(, :-$, :-&, :-(, :-/, :-<, :->, :-@, :-c, :-p, :-s, :-x, :-|, :-||, :/, :<, :@, :[, :`-(, :o, P, :|, =(, =p, >:(, ^o), :((

Tabela 2.4: Símbolos que denotam emoção comumente encontrados em *tweets*.

2.3.3 Conjuntos de dados

Os experimentos realizados neste capítulo utilizaram dos conjuntos de *tweets*: *Sanders, Stanford, Debate Obama-McCain e Reforma da Saúde*. Tais conjuntos são consiederados representativos por serem obtidos a partir de diferentes assuntos (Hassan Saif & Alani, 2013) e estarem disponíveis publicamente. Tais conjuntos de dados foram escolhidos por viabilizar estudos comparativos.

2.3.3.1 Sanders

O conjunto de *tweets Sanders*[10] consiste de mensagens manualmente anotadas por especialistas, coletadas a partir de quatro termos de busca: *@apple, #google, #microsoft, and #twitter*. Cada *tweet* tem um rótulo de sentimentos: positivo, neutro, ou negativo, e irrelevante. A Tabela 2.5 apresenta o número de *tweets* de cada tópico e classe:

Tópico	Classe Positiva	Classe Neutra	Classe Negativa	Termo de busca no Twitter
Apple	191	581	377	$@apple$
Google	218	604	61	$\#google$
Microsoft	93	671	138	$\#microsoft$
Twitter	68	647	78	$\#twitter$

Tabela 2.5: Especificação da base de dados *Sanders*.

[10]Disponível para *download* em http://www.sananalytics.com, com último acesso em 27/08/2015.

2.3.3.2 Stanford

A base de dados *Stanford* (Go *et al.*, 2009) possui 1.600.000 *tweets* de treinamento. Periodicamente, foram feitas consultas à base de dados do Twitter usando símbolos de emoções positivas – :) – e negativas – (: – . Depois de remover os *retweets*, os símbolos de emoção e *tweets* repetidos, restaram 800.000 *tweets* classificados em "positivos", e 800.000 *tweets* classificados em "negativos". Em contraste com o conjunto de treinamento coletado a partir de emoções específicas, o conjunto de teste foi coletado a partir de nomes de produtos, companhias e pessoas e foram manualmente anotados com o rótulo de classe. O conjunto de teste é composto de 177 *tweets* "negativos" e 182 *tweets* "positivos".

2.3.3.3 Debate Obama-McCain – OMD

O banco de dados *Obama-McCain Debate* foi construído a partir de 3.238 *tweets* coletados a partir do primeiro debate presidencial televisionado nos Estados Unidos em 2.008 (Diakopoulos & Shamma, 2010). Os rótulos de classe destes *tweets* foram obtidos a partir de anotadores da empresa *Amazon Mechanical Turk*[11]. Cada *tweet* foi rotulado por um ou mais anotadores como "positivo", "negativo", "neutro", ou "outro". *Tweets* classificados em "outro" são aqueles que não se encaixam em nenhuma das categorias anteriores. Foram mantidas as mensagens rotuladas por no mínimo três profissionais para os sentimentos positivos ou negativos, desta forma resultando em um conjunto de 1.906 *tweets*, sendo 710 da classe positiva e 1.196 da classe negativa.

Em outra configuração deste conjunto de dados, foram considerados somente *tweets* com unanimidade das opiniões. Este foi chamado *Strict Obama-McCain Debate*, o qual tem 916 *tweets* com 347 mensagens da classe positiva e 569 mensagens da classe negativa.

2.3.3.4 Reforma da Saúde - HCR

A base de dados *Health Care Reform* foi construída a partir de *tweets* contendo a *hashtag* #hcr em março de 2010 (Speriosu *et al.*, 2011). Um subconjunto desse *corpus* foi manualmente anotado como sendo da classe "positiva", "negativa" ou "neutra". Este conjunto é disponibilizado já particionado em um conjunto de treinamento e um conjunto destinado aos testes. O conjunto direcionado à fase de treinamento contém 839 *tweets* (215 da classe positiva e 406 da classe negativa). O conjunto de testes possui 154 *tweets* da classe positiva e 511 da classe negativa.

[11] https://www.mturk.com/

2.3.4 Configuração

Os experimentos foram conduzidos usando a plataforma WEKA[12]. Neste ambiente foram executados os algoritmos Multinomial Naive Bayes, Regressão Logística, e *Random Forest*. Foi usado também o pacote para Support Vector Machines (Chang & Lin, 2011) LibSVM[13] para o treinamento de classificadores SVM. Para os conjuntos de dados *Obama-McCain* Debate e *Sanders Twitter Sentiment* foram usados validação cruzada com 10 pastas (*10-fold cross validation*). Para o conjuntos de dado *Health Care Reform*, foram usados os mesmos conjuntos de dados de treinamento e teste disponíveis em Saif *et al.* (2012b). Finalmente, para o conjunto de dados *Stanford Twitter Sentiment Corpus*, foram realizadas amostragens a partir do conjunto de treinamento original, as quais foram também utilizadas no treinamento, de modo que na fase de validação o conjunto de teste considerado foi o mesmo disponível em Go *et al.* (2009).

2.3.5 Resultados

Foram comparados os resultados dos classificadores de base considerados individualmente com os resultados providos pelos *ensembles* propostos. Além disso, foram consideradas diferentes combinações de atributos do tipo *bag-of-words* (BoW), *feature hashing* (FH) e léxicos, avaliando o potential dos *ensembles* segundo a acurácia do seu respectivo modelo de classificação. Os melhores resultados descritos na literatura também são reportados para fins de comparação.

A Tabela 2.6 apresenta os resultados das abordagens baseadas em BoW, enquanto a Tabela 2.7 mostra os resultados das abordagens baseadas em atributos do tipo *feature hashing*. De acordo com as tabelas, os *ensembles* proporcionam ganhos de acurácia em todas as configurações consideradas. Como esperado, agregadores de classificadores proporcionam ganhos de qualidade, mas podem originar custos computacionais adicionais. Com isso, é importante lembrar que na prática é comum se testar diferentes classificadores de sentimento para o mesmo problema e, nesse sentido, custos computacionais adicionais significativos não seriam gerados.

[12] http://www.cs.waikato.ac.nz/ml/weka/
[13] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Conjunto de dados OMD

Método	Acc.(%)	Classe Positiva			Classe Negativa		
		Precision(%)	Recall(%)	F1(%)	Precision(%)	Recall(%)	F1(%)
SVM-BoW	72,25	64,90	55,50	59,80	75,70	82,20	78,80
SVM-BoW+Lex	75,55	68,40	63,90	66,10	79,40	82,40	80,90
RF-BoW	71,04	62,90	54,20	58,20	74,9	81,00	77,80
RF-BoW+Lex	73,82	66,70	59,30	62,80	77,30	82,40	79,80
LR-BoW	70,57	66,90	41,50	51,30	71,70	**87,80**	78,90
LR-BoW+Lex	73,85	66,10	56,90	61,20	76,40	82,70	79,40
MNB-BoW	72,19	64,00	57,90	60,80	76,30	80,70	78,50
MNB-BoW+Lex	75,97	68,80	**65,10**	66,90	**79,90**	82,40	81,20
ENS(LR+RF+MNB+SVM)-BoW	73,14	66,30	56,60	61,10	76,30	82,90	79,50
ENS(LR+RF+MNB)-BoW+Lex	**76,81**	**71,10**	63,70	**67,20**	79,70	84,60	**82,10**
Melhor resultado na literatura: Hu *et al.* (2013b)	*76,30*	-	-	-	-	-	-

Conjunto de dados Strict OMD

Método	Acc.(%)	Classe Positiva			Classe Negativa		
		Precision(%)	Recall(%)	F1(%)	Precision(%)	Recall(%)	F1(%)
SVM-BoW	74,02	67,60	60,20	63,70	77,30	82,40	79,80
SVM-BoW+Lex	78,93	73,80	68,90	71,20	81,80	85,10	83,40
RF-BoW	73,91	65,30	66,60	65,90	79,40	78,40	78,90
RF-BoW+Lex	79,36	70,90	**77,20**	73,90	**85,30**	80,70	82,90
LR-BoW	72,38	70,60	46,40	56,00	73,00	85,20	79,90
LR-BoW+Lex	78,06	**74,50**	64,00	68,80	79,80	**86,60**	83,10
MNB-BoW	75,43	68,70	64,60	66,60	79,20	82,10	80,60
MNB-BoW+Lex	80,13	**74,50**	72,30	73,40	83,40	84,90	**84,10**
ENS(LR+RF+MNB)-BoW	75,55	70,20	59,70	64,50	77,50	84,50	80,80
ENS(LR+RF+MNB)-BoW+Lex	**80,35**	73,50	75,20	**74,40**	84,70	83,50	**84,10**
Melhor resultado na literatura: Hu *et al.* (2013b)	*76,30*	-	-	-	-	-	-

Sanders – Twitter Sentiment Corpus

Método	Acc.(%)	Classe Positiva			Classe Negativa		
		Precision(%)	Recall(%)	F1(%)	Precision(%)	Recall(%)	F1(%)
SVM-BoW	82,43	80,00	83,00	81,50	84,70	82,00	83,30
SVM-BoW+Lex	83,98	81,20	85,40	83,20	86,70	82,70	84,70
RF-BoW	79,24	75,60	81,90	78,60	83,00	76,90	79,80
RF-BoW+Lex	82,35	78,80	84,90	81,80	85,90	80,10	82,90
LR-BoW	77,45	76,40	74,60	75,50	78,30	80,00	79,10
LR-BoW+Lex	79,49	77,20	79,50	78,30	81,60	79,50	80,60
MNB-BoW	79,82	80,10	75,40	77,70	79,60	83,60	81,60
MNB-BoW+Lex	83,41	**82,90**	81,10	82,00	83,80	**85,50**	84,60
ENS(LR+RF+MNB+SVM)-BoW	82,76	80,70	82,80	81,70	84,70	82,70	83,70
ENS(SVM+RF+MNB)-BoW + Lex	**84,89**	82,10	**86,30**	**84,20**	**87,50**	83,60	**85,50**
Melhor resultado na literatura: Ziegelmayer & Schrader (2012)	*84,40*	-	-	-	-	-	-

Melhor amostragem com o conjunto Stanford - Amostragem com 3000 tweets

Método	Acc.(%)	Classe Positiva			Classe Negativa		
		Precision(%)	Recall(%)	F1(%)	Precision(%)	Recall(%)	F1(%)
SVM-BoW	67,41	67,2	69,80	68,50	67,60	65,00	66,30
SVM-BoW+Lex	73,82	72,90	76,90	74,90	74,90	70,60	72,70
RF-BoW	66,57	65,00	73,60	69,10	68,60	59,30	63,60
RF-BoW+Lex	74,37	73,00	78,60	75,70	76,10	70,10	72,90
LR-BoW	64,90	60,00	**92,30**	72,70	82,30	36,70	50,80
LR-BoW+Lex	76,32	73,20	84,10	78,30	80,70	68,40	74,00
MNB-BoW	71,31	72,10	70,90	71,50	70,60	71,80	71,10
MNB-BoW+Lex	79,39	**80,70**	78,00	79,30	78,10	**80,80**	79,40
ENS(LR+RF+MNB)-BoW	72,14	70,50	77,50	73,80	74,20	66,70	70,20
ENS(LR+RF+MNB)-BoW+Lex	81,06	79,70	84,10	**81,80**	**82,60**	78,00	**80,20**
Melhor resultado na literatura: Bakliwal *et al.* (2012)	***87,20***	-	-	-	-	-	-

Conjunto de dados HCR

Método	Acc.(%)	Classe Positiva			Classe Negativa		
		Precision(%)	Recall(%)	F1(%)	Precision(%)	Recall(%)	F1(%)
SVM-BoW	73,99	42,00	32,50	36,60	81,00	86,50	83,60
SVM-BoW+Lex	75,94	47,50	37,00	41,60	82,20	87,70	84,80
RF-BoW	70,83	34,60	29,20	31,70	79,60	83,40	81,50
RF-BoW+Lex	72,93	38,40	27,90	32,30	79,90	86,50	83,10
LR-BoW	73,83	40,00	26,00	31,50	79,80	88,30	83,80
LR-BoW+Lex	74,73	43,00	27,90	33,90	80,40	88,80	84,40
MNB-BoW	72,48	42,80	55,80	48,50	85,30	77,50	81,20
MNB-BoW+Lex	75,33	47,40	**60,40**	**53,10**	**87,00**	79,80	83,30
ENS(LR+RF+MNB)-BoW	75,19	44,70	29,90	35,80	80,80	88,80	84,60
ENS(LR+RF+SVM+MNB)-BoW+Lex	**76,99**	50,50	35,70	41,80	82,20	**89,40**	85,70
Melhor resultado na literatura: Saif *et al.* (2012b)	*71,10*	**53,80**	47,20	50,30	84,50	87,60	**86,00**

Tabela 2.6: Comparação entre os resultados fornecidos para a abordagem com *bag-of-words* (melhores resultados em negrito). LR, RF e MNB referem-se aos algoritmos de classificação regressão logística, *random forest* e *naive bayes* multinominal, respectivamente. ENS indica a uso de *ensembles*, BoW refere-se a *bag-of-words*, lex refere-se ao uso de léxicos e a abreviação SVM-BoW+lex indica um classificador SVM com atributos do tipo *bag-of-words* e provenientes de léxicos. Outras abreviações são Acc. para acurácia F1 para a medida *F-measure*.

Conjunto de dados OMD							
Método	Acc.(%)	Classe Positiva			Classe Negativa		
		Precision(%)	Recall(%)	F1(%)	Precision(%)	Recall(%)	F1(%)
SVM-FH	51,10	37,90	49,20	42,80	63,40	52,30	57,30
SVM-FH+Lex	62,85	50,10	**57,60**	53,60	72,40	66,00	69,00
RF-FH	61,39	47,10	29,60	36,30	65,80	80,30	72,30
RF-FH+Lex	67,37	58,50	42,50	49,30	70,60	82,10	75,90
LR-FH	63,28	61,90	3,70	6,90	63,30	**98,70**	77,10
LR-FH+Lex	70,57	**67,00**	41,30	51,10	71,60	88,00	**78,90**
MNB-FH	62,54	47,20	4,80	8,70	63,10	96,80	76,40
MNB-FH+Lex	70,41	64,40	45,90	53,60	72,60	84,90	78,30
ENS(LR+RF+MNB)-FH	64,59	39,80	32,00	35,50	63,80	71,30	67,40
ENS(LR+RF+MNB)-FH+Lex	**70,62**	57,70	53,70	**55,60**	**73,60**	76,70	75,10
Conjunto de dados Strict OMD							
Método	Acc.(%)	Classe Positiva			Classe Negativa		
		Precision(%)	Recall(%)	F1(%)	Precision(%)	Recall(%)	F1(%)
SVM-FH	51,31	39,80	55,30	46,30	64,20	48,90	55,50
SVM-FH + Lex	62,99	51,30	47,00	49,00	69,20	72,80	71,00
RF-FH	61,36	48,50	31,70	38,30	65,60	79,40	71,90
RF-FH+Lex	72,60	69,00	50,10	58,10	73,90	86,30	79,60
LR-FH	65,29	65,30	17,90	28,10	65,30	94,20	77,10
LR-FH+Lex	73,03	68,10	54,20	60,40	75,20	84,50	79,60
MNB-FH	60,70	36,20	4,90	8,60	62,00	**94,70**	75,00
MNB-FH+Lex	71,39	66,20	50,10	57,00	73,50	84,40	78,60
ENS(LR+RF+MNB)-FH	65,17	59,60	23,30	33,50	65,90	90,30	76,20
ENS(LR+RF+MNB)-FH+Lex	**74,56**	**70,20**	**57,10**	**63,00**	**76,50**	85,20	**80,60**
Sanders – Twitter Sentiment Corpus							
Método	Acc.(%)	Classe Positiva			Classe Negativa		
		Precision(%)	Recall(%)	F1(%)	Precision(%)	Recall(%)	F1(%)
SVM-FH	49,75	46,30	49,80	48,00	53,20	49,70	51,40
SVM-FH+Lex	75,00	74,10	71,20	72,60	75,70	78,30	77,00
RF-FH	55,64	52,10	59,80	55,70	59,80	52,00	55,60
RF-FH+Lex	71,63	68,00	73,90	70,80	75,40	69,70	72,40
LR-FH	56,94	54,70	43,50	48,50	58,20	68,70	63,00
LR-FH+Lex	75,98	74,40	**73,90**	74,10	**77,40**	77,80	77,60
MNB-FH	54,25	51,00	45,30	48,00	56,50	62,10	59,20
MNB-FH+Lex	75,08	73,30	73,20	73,20	76,60	76,80	76,70
ENS(LR+RF+MNB)-FH	57,84	53,70	49,30	51,40	58,80	63,00	60,80
ENS(LR+RF+MNB)-FH+Lex	**76,63**	**75,40**	73,20	**74,30**	77,20	**79,20**	**78,20**
Melhor amostragem com o conjunto Stanford - Amostragem com 3000 tweets							
Método	Acc.(%)	Classe Positiva			Classe Negativa		
		Precision(%)	Recall(%)	F1(%)	Precision(%)	Recall(%)	F1(%)
SVM-FH	47,63	47,50	31,90	38,20	47,70	63,80	54,60
SVM-FH+Lex	54,32	54,00	66,50	59,60	54,80	41,80	47,40
RF-FH	47,63	52,50	58,80	55,40	51,60	45,20	48,20
RF-FH+Lex	70,47	70,40	72,00	71,20	70,50	68,90	69,70
LR-FH	55,71	56,20	57,10	56,70	55,20	54,20	54,70
LR-FH+Lex	78,55	76,60	83,00	79,70	**80,90**	74,00	77,30
MNB-FH	54,32	54,50	59,30	56,80	54,00	49,20	51,50
MNB-FH+Lex	78,27	**77,40**	**80,80**	**79,00**	79,30	**75,70**	**77,50**
ENS(LR+RF+MNB)-FH	57,38	55,30	54,40	54,80	53,90	54,80	54,30
ENS(LR+RF+MNB)-FH+Lex	**79,11**	76,90	78,60	77,70	77,50	**75,70**	76,60
Conjunto de dados HCR							
Método	Acc.(%)	Classe Positiva			Classe Negativa		
		Precision(%)	Recall(%)	F1(%)	Precision(%)	Recall(%)	F1(%)
SVM-FH	67,22	28,40	27,30	27,80	78,30	79,30	78,80
SVM-FH+Lex	69,92	20,50	10,40	13,80	76,50	87,90	81,80
RF-FH	63,16	25,10	29,90	27,30	77,60	73,20	75,30
RF-FH+Lex	72,48	38,70	**45,50**	**41,80**	**82,60**	78,30	80,40
LR-FH	67,52	26,90	23,40	25,00	77,80	80,80	79,30
LR-FH+Lex	77,6	50	17,50	26,00	79,20	94,70	**86,30**
MNB-FH	73,83	34,80	10,40	16,00	77,70	94,10	85,10
MNB-FH+Lex	75,34	46,50	26,00	33,30	80,30	91,00	85,30
ENB(LR+RF+MNB)-FH	76,09	27,50	7,10	11,30	77,10	94,30	84,90
ENB(LR+RF+MNB)-FH+Lex	**78,35**	**52,90**	29,90	38,20	81,30	92,00	**86,30**

Tabela 2.7: Resultados usando *feature hashing* (FH) – melhores resultados em negrito.

Comparando pares de classificadores com e sem léxicos observa-se que o uso de atributos derivados de léxicos proporcionou melhores resultados em todos os experimentos. Como mencionado anteriormente, o léxico utilizado foi construído para comentários de produtos vendidos *on-line* (Liu, 2012). Tal léxico consiste de palavras informais, informação típica encontrada em *tweets*. No conjunto de dados *Stanford* a melhora foi relevante, uma vez que os tweets nesse conjunto foram coletados a partir de *emoticons*. Neste tipo de conjunto de dados, não há um domínio específico, enquanto que nos outros conjuntos de dados são abordados tópicos mais específicos, tais como tecnologia (Sanders), política (*Strict* OMD and OMD) e saúde (Conjunto de dados HCR).

Considerando os resultados com *feature hashing* (Tabela 2.7), os *ensembles* reportados nesta tese apresentaram maiores valores de acurácia do que os classificadores de base em todos os conjuntos de dados. Levando em consideração a medida *F-Measure*, foram obtidos os melhores resultados em 80% dos casos com BoW. Os resultados com o uso de *feature hashing* são piores do que os resultados alcançados com o uso de BoW na maioria das bases de dados, exceto para o conjunto de dados HCR, no qual os resultados foram melhores com *feature hashing* em comparação aos resultados apresentados na literatura — incluindo os próprios resultados com *ensembles* baseados em BoW reportados nesta tese. Entretanto, como esperado, *feature hashing* possibilita uma significante redução de dimensionalidade, como mostrado na Tabela 2.8, e há um importante *trade-off* entre a acurácia da classificação e a economia em recursos computacionais. É importante enfatizar que nenhum dos resultados reportados na literatura fazem uso de técnicas de redução de dimensionalidade e, portanto, estritamente não podem ser comparados aos resultados considerando o uso de *feature hashing*. Como já mencionado anteriormente, *feature hashing* não havia ainda sido avaliada em conjuntos de dados públicos para análise de sentimentos de *tweets*.

Conjunto de dados	**# atributos BoW**	**# atributos FH**
Conjunto de dados OMD	1.352	13
Conjunto de dados *Strict* OMD	759	10
Sanders - Twitter Sentiment Corpus	1.203	23
Stanford Dataset	2.137	21
Conjunto de dados HCR	1.432	10

Tabela 2.8: Número de atributos usando *bag-of-Words* e usando *feature hashing*.

Alguns experimentos adicionais para o conjunto de dados *Stanford* foram também conduzidos. Devido a limitações computacionais, não foram feitos experimentos com o conjunto de treinamento completo. Para driblar tal limitação foram consideradas amostragens aleatórias balanceadas com tamanhos variando de 500 a 3.000 *tweets*. A Figura 2.9 apresenta uma visão geral dos algoritmos de aprendizagem de máquina considerados no experimento. Note que o *ensemble* obtido a partir da BoW e léxicos alcançou os melhores resultados. Outro ponto a ser considerado é que, em contraste com outras abordagens, os valores de acurácia alcançados fo-

ram obtidos a partir de amostragens pequenas dos dados, enfatizando que tais valores poderiam ser ainda melhores com um conjunto de treinamento maior. O melhor índice de acurácia reportado na literatura para o conjunto de treinamento completo (formado por 1.600.000 *tweets*) é $87{,}20\%$, enquanto que a acurácia a partir do *ensemble* treinado com apenas $0{,}03\%$ dos dados (reportados nesta tese) alcançou $81{,}06\%$ no conjunto de validação proposto em Go *et al.* (2009).

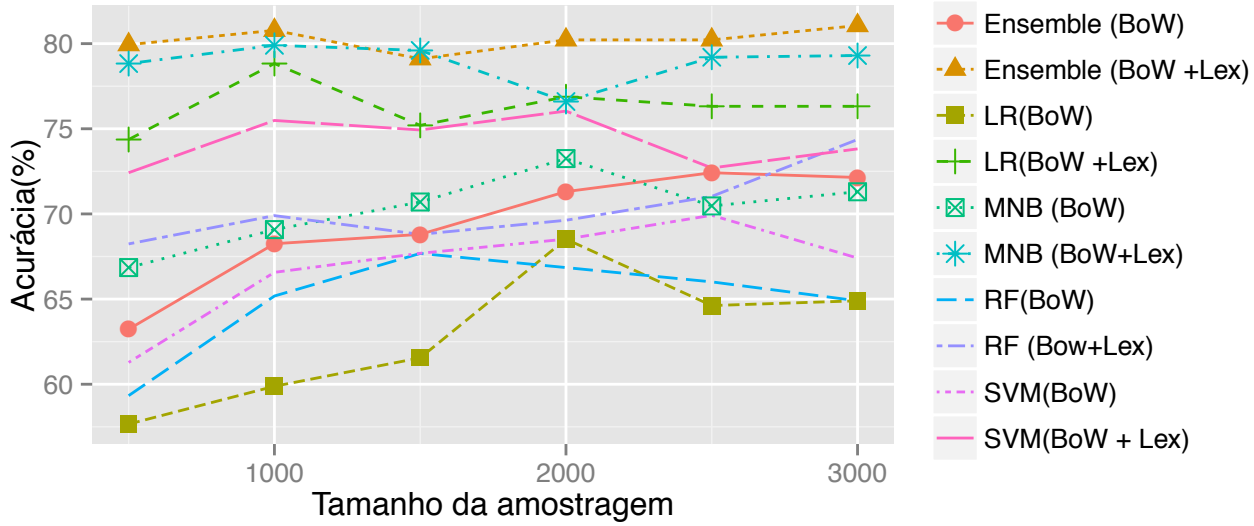


Figura 2.9: Acurácias a partir de diferentes amostragens do conjunto de treinamento — Stanford.

2.4 Considerações Finais

Antes de qualquer consideração, é importante ressaltar que os trabalhos encontrados na literatura (Lin & Kolcz, 2012; Clark & Wicentwoski, 2013; Rodrıguez-Penagos *et al.*, 2013; Hassan *et al.*, 2013; Kanakaraj & Guddeti, 2015; Hagen *et al.*, 2015), com exceção de Lin & Kolcz (2012) não haviam sido publicados quando se deu o início da validação e implementação da hipótese de pesquisa 1 reportada neste capítulo. Outro ponto a ser enfatizado é que o trabalho reportado por Lin & Kolcz (2012) não fez uso de uma base de dados pública, o que inviabilizaria um estudo comparativo. A constatação da realização de diversos trabalhos correlatos, concomitantemente ao trabalho desta tese, evidencia a grande relevância do tema.

Os resultados mencionados neste capítulo foram publicados em da Silva *et al.* (2014c,b) e cumprem o primeiro objetivo desta tese, validando a primeira hipótese de pesquisa: mostrar evidências de que agregadores de classificadores têm o poder preditivo comparável ou superior ao poder preditivo de estratégias de classificação tradicionais (*single classifier*) para análise de sentimento em *tweets*.

Capítulo

3

Agregadores de Classificadores e Agrupadores

A principal motivação para se combinar classificadores e agrupadores provém da idéia de combinar informações de dados rotulados e não rotulados, tendo como benefício as propriedades que tornaram os agregadores de classificadores e de agrupadores ferramentas conceituais amplamente usadas na prática. Adicionalmente, outras propriedades desejáveis podem emergir da combinação dessas duas ferramentas. Por exemplo, modelos não supervisionados podem fornecer uma série de restrições complementares para auxiliar na classificação de novos dados em um conjunto-alvo (Acharya *et al.*, 2014, 2011; Basu *et al.*, 2008). A partir deste ponto de vista, a premissa subjacente é que objetos similares no conjunto-alvo são mais propensos a compartilharem o mesmo rótulo de classe. Esta premissa é intuitiva e suportada por diferentes estudos (Gan *et al.*, 2013; Gao *et al.*, 2013; Rahman & Verma, 2013; Duval-Poo *et al.*, 2012; Soares *et al.*, 2012; Verma & Rahman, 2012; Acharya *et al.*, 2014, 2011).

Diversos trabalhos na literatura têm reportado os benefícios de se usar informações de dados rotulados e não rotulados em aplicações de aprendizado de máquina. Fundamentalmente, resultados de agrupamentos podem ajudar a aumentar a capacidade de generalização de classificadores (Bousquet *et al.*, 2004), enquanto vários trabalhos que abordam *agrupamento supervisionado* (Pedrycz & Vukovich, 2004; Papadimitriou *et al.*, 2001; Timm, 2001) e *agrupamento semissupervisionado* (Kulis *et al.*, 2005; Zhu, 2008; Basu *et al.*, 2004) fornecem evidências teóricas e práticas sobre as vantagens de se usar a informação das classes para agrupar dados. Do ponto de vista de classificação, a idéia de se combinar classificadores e agrupadores tem como benefícios (Cai *et al.*, 2010, 2009; Gao *et al.*, 2013, 2009): **(i)** aumentar a qualidade e a robustez do classificador; e **(ii)** facilitar a interpretação do modelo na medida em que se torna possível a identificação das relações subjacentes entre grupos e classes. Em aplicações de fluxos contínuos de dados (*data streams*), algoritmos de agrupamento de dados são úteis para identificar mudan-

ças de *conceito* e novas classes não previstas no conjunto de treinamento[1]. Sob este ponto de vista, a informação de grupos pode ser útil para se projetar métodos de aprendizado que sejam "conscientes" das possíveis diferenças entre as distribuições dos dados de treinamento e dos dados do conjunto-alvo (Acharya *et al.*, 2014). Outras aplicações envolvem a identificação de *outliers*, evitando a sua associação com alguma das classes (Kollios *et al.*, 2003; Miller & Browning, 2003) e a preocupação em tornar o modelo mais interpretável (Yong *et al.*, 2009; Bai *et al.*, 2007; Zhang *et al.*, 2005).

A maioria dos trabalhos que se propõem a combinar agrupadores e classificadores, realizam esta tarefa de maneira *sequencial* ou *simultânea*:

- Nas abordagens sequenciais, os dados são primeiramente agrupados e posteriormente classificados a partir da estrutura de grupos obtida. Por exemplo, Setnes & Babuska (1999) inicialmente usaram o algoritmo *Fuzzy C-Means* (FCM) (Bezdek, 1981) para induzir grupos e, na sequência, uma matriz de relação *fuzzy* — a qual indica as relações entre grupos e classes — é construída para auxiliar na classificação de novos objetos. Cai *et al.* (2007) aperfeiçoaram este algoritmo, tornando-o mais estável e capaz de atuar em conjuntos de dados com distribuições não necessariamente hiper-esféricas. Li & Ye (2006) introduziram um algoritmo que se baseia em agrupamento supervisionado para construir um classificador. Em De Castro *et al.* (2004) um algoritmo genético para agrupamento é combinado com o *K-Médias* (KM) (MacQueen, 1967) para automaticamente estimar a localização e o número de funções de base radial em uma rede neural RBF. Embora esses trabalhos indiquem que as tarefas de classificação e agrupamento possam ser complementares, tais abordagens sequenciais nem sempre favorecem a otimalidade dos dois tipos de aprendizagem (Cai *et al.*, 2009), *i.e.*, na maioria das vezes, a tarefa de agrupamento apenas auxilia na classificação, mas não se beneficia desta. Cai *et al.* (2009) também argumentam que, por meio do conhecimento das relações entre grupos e classes, pode-se obter classificadores mais interpretáveis e robustos, desta forma motivando o uso de abordagens simultâneas para se combinar classificadores e agrupadores.

- Buscando combinar um classificador com um agrupador de maneira simultânea, Cai *et al.* (2009) introduziram um algoritmo que se baseia na otimização de uma única função objetivo, projetada para avaliar tanto o classificador quanto o agrupador, de maneira que se possa alcançar simultaneamente resultados robustos e eficazes para os dois modelos. Posteriormente, em Cai *et al.* (2010), esse algoritmo foi estendido para um cenário de otimização multi-objetivo, no qual se permite a otimização do problema por meio de vários critérios (tanto para agrupamento quanto para classificação). Para ilustrar a abordagem proposta, os autores utilizaram duas funções objetivo baseadas nos protótipos dos grupos obtidos pelo algoritmo de agrupamento FCM (taxa de erro de classificação e compactação/separabilidade entre os grupos). Inspirado por estes trabalhos, Qian *et al.* (2012)

[1] A mudança de conceito também é conhecida como *concept drift*, e o surgimento de novas classes envolve a evolução de conceito também chamada de *concept evolution* (Masud *et al.*, 2013, 2010a,b).

desenvolveram uma abordagem mais rápida e flexível. O algoritmo tem garantias de convergência e foi generalizado para aprendizado semissupervisionado.

3.1 Combinação de Classificadores e Agrupadores em análise de sentimentos

A informação oriunda de tópicos encontrados em textos tem sido aplicada em diferentes problemas no cenário de análise de sentimentos. Os trabalhos apresentados em Mei *et al.* (2007); Jo & Oh (2011); He *et al.* (2012) propõem o uso de informação de tópicos para extrair os aspectos relacionados ao domínio em questão e posteriormente classificar o sentimento quanto à sua polaridade. As fontes textuais são provenientes de *blogs*, comentários de filmes e livros. Para exemplificar a abordagem segundo Mei *et al.* (2007), considere que o domínio seja *Dell Laptop*. Para o domínio considerado, após a modelagem de tópicos, são encontrados os tópicos: *(i) relacionado ao aspecto preço* e *(ii) relacionado ao aspecto bateria*. Na sequência, os textos relacionados aos tópicos (aspectos) são classificados quanto ao seu respectivo sentimento, positivo, negativo ou neutro.

Dentre as principais motivações para a combinação de classificadores e agrupadores no contexto de análise de sentimentos está o fato de que a maior parte dos trabalhos que consideram apenas modelos de classificação utilizam somente informação local proveniente dos *tweets* (e.g., léxicos, unigramas, bigramas, *part-of-speech* etc; inferidos na etapa conhecida como engenharia de atributos). Tais abordagens não consideram a informação global de mais alto nível promovida pelos tópicos, ou seja, a informação gerada pelo agrupamento de *tweets*. Em particular, a mesma palavra pode ter polaridades de sentimento controversas em domínios diferentes. Por exemplo, embora o adjetivo "*complexo*" na frase "*O livro é complexo e emocionante!*" pode ter uma orientação positiva em uma resenha do livro, ele também poderia ter uma orientação negativa na frase "*É difícil usar um celular tão complexo.*" em um texto proveniente de comentários sobre aparelhos eletrônicos. Portanto, é mais adequado analisar tópicos e sentimentos simultaneamente.

No âmbito da análise de sentimentos em *tweets*, o cenário em que os dados são primeiramente agrupados e posteriormente classificados a partir da estrutura de grupos obtida se repete, e os dois trabalhos existentes na literatura (Si *et al.*, 2013; Xiang & Zhou, 2014) presumem o agrupamento dos dados como uma tarefa de "pré-processamento", uma vez que só após a realização desta é feita classificação de sentimentos.

Si *et al.* (2013) emprega uma extensão não paramétrica de um modelo de tópicos descrito em Teh *et al.* (2006) para estimar o número de tópicos em um um conjunto de *tweets* (*streaming snapshot*) extraído diariamente. Na sequência, é construído um classificador de sentimentos para os tópicos inferidos diariamente pelo modelo de tópicos.

Em Xiang & Zhou (2014), uma abordagem semissupervisionada baseada em tópicos é usada para análise de sentimentos em *tweets*. Os autores propõem construir um modelo de tópicos nos dados rotulados. A informação de tópicos é obtida a partir da implementação do principal

modelo de tópicos da literatura, conhecido como *Latent Dirichlet Allocation* (LDA) e proposto em Blei *et al.* (2003). A partir dos tópicos obtidos, um modelo de sentimentos específico é construído em cada grupo encontrado. A Figura 3.1 sumariza esta abordagem. Primeiramente, um classificador é inferido a partir dos *tweets* rotulados. Na sequencia, tal modelo é usado para estimar as probabilidades de classe dos *tweets* não rotulados (este passo ocorre somente uma vez). Procede-se então com o passo 4, no qual um subconjunto de *tweets* com probabilidade de classe maior que um *threshold* é selecionado. Estes tweets são então incluídos no conjunto de *tweets* rotulados. No passo 5, os *tweets* rotulados são usados para construir um modelo de tópicos, a partir do qual serão obtidas distribuições de tópicos que serão armazenadas para cada *tweet*. Por fim, um agrupamento baseado na distribuição de tópicos é inferido e um modelo de sentimentos particular é treinado para cada grupo.

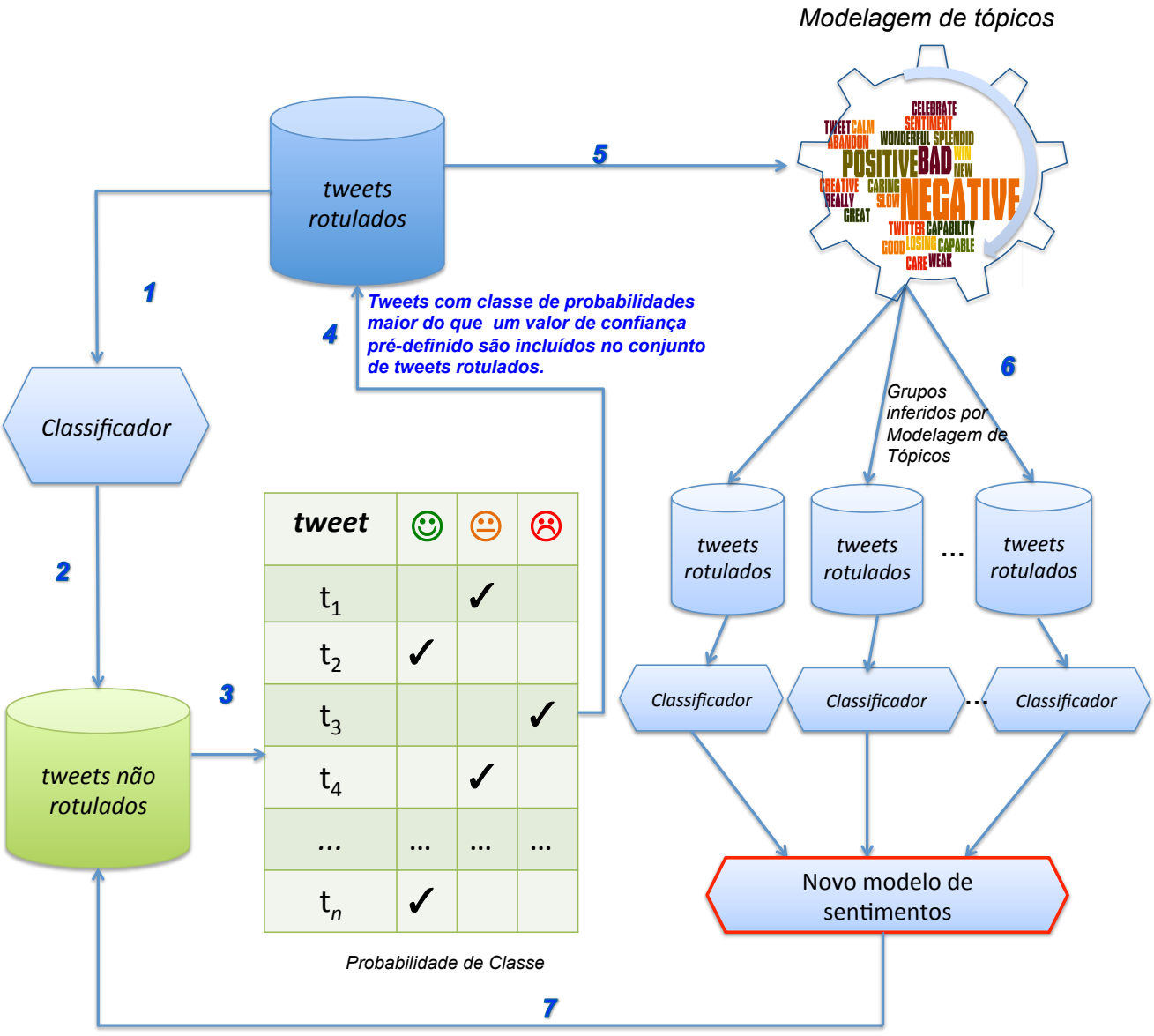


Figura 3.1: Abordagem de análise de sentimentos que usa informação de tópicos como pré-processamento para a etapa de classificação (Xiang & Zhou, 2014).

Os trabalhos de Acharya *et al.* (2014, 2011) e Gao *et al.* (2009) propõem combinar diferentes classificadores e diferentes agrupadores para lidar com o problema de classificação de dados. Em particular o algoritmo, já existente, C³E (*Consensus between Classification and Clustering Ensembles*) (Acharya *et al.*, 2014, 2011) aplicado em diversos domínios e com resultados promissores nos mesmos é objeto de estudo no cenário de análise de sentimentos em *tweets* nesta tese. Tal algoritmo classifica novos dados a partir da combinação de dois modelos: um resultante de um modelo de classificação obtido com um agregador de vários classificadores (Capítulo 2) e outro resultante da agregação de agrupadores. Combinar um classificador com um agrupador de maneira simultânea é o foco deste Capítulo. Tal algoritmo é descrito mais detalhadamente a seguir.

3.2 O Algoritmo C³E

A abordagem do algoritmo C³E (*Consensus between Classification and Clustering Ensembles*) (Acharya *et al.*, 2014, 2011) para agregar classificadores e agrupadores gerando uma classificação consolidada dos dados é ilustrada na Figura 3.2. Em resumo, este algoritmo recebe como entradas uma distribuição de probabilidades de classes, $\{\boldsymbol{\pi}_i\}_{i=1}^n$, e uma matriz de similaridades, $\mathbf{S}$, referentes aos objetos de um conjunto-alvo, $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n$, que se deseja classificar. Após processar estas entradas, o algoritmo fornece como resultado um conjunto de vetores $\{\mathbf{y}_i\}_{i=1}^n$ correspondentes às estimativas (refinadas) das probabilidades de classes *a posteriori* para cada objeto de $\mathcal{X}$, ou seja, $\mathbf{y}_i = p(C \mid \mathbf{x}_i)$.

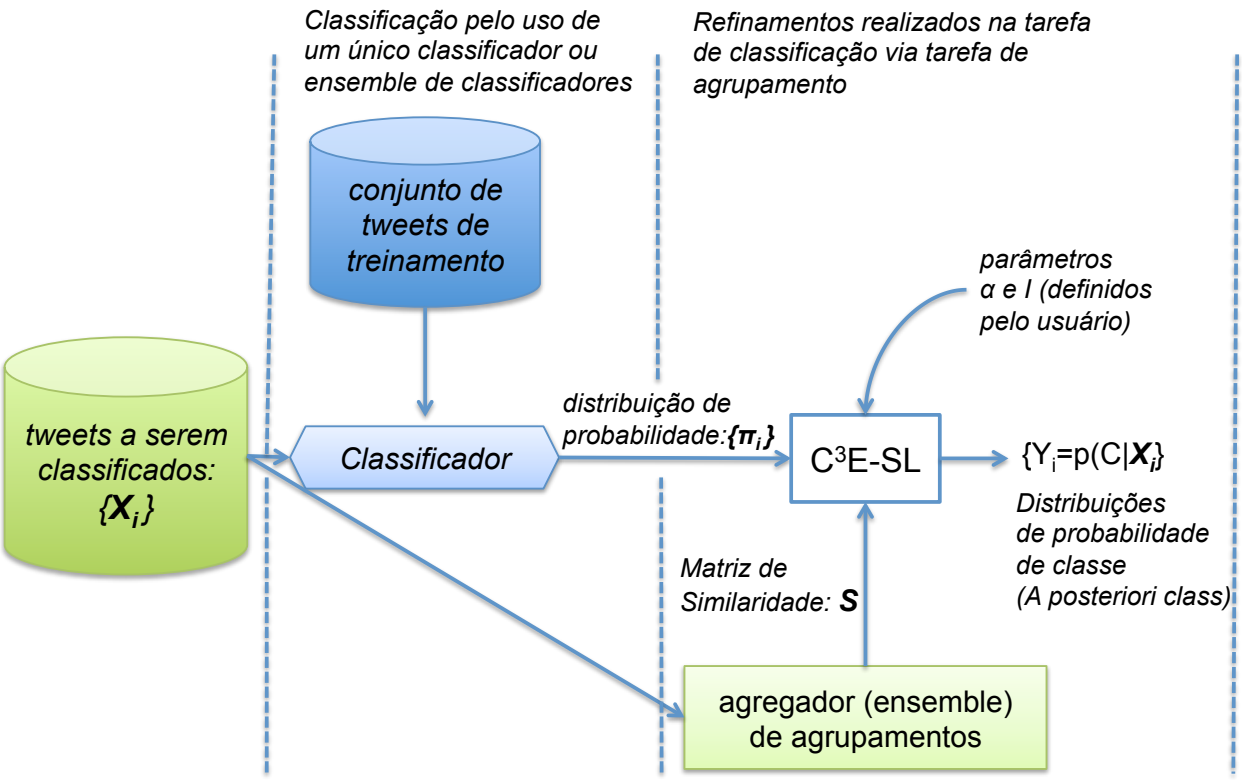


Figura 3.2: Visão geral do algoritmo C³E (Acharya *et al.*, 2014, 2011).

Mais formalmente, considere que r_1 ($r_1 \geq 1$) classificadores, indexados por q_1, e r_2 ($r_2 \geq 1$) agrupadores, indexados por q_2, são empregados para obter uma classificação consolidada dos objetos do conjunto-alvo. Os seguintes passos (I – III) sumarizam a abordagem proposta em (Acharya *et al.*, 2014, 2011) — os passos I e II podem ser vistos como passos preliminares para se obter as entradas (*inputs*) para o algoritmo C^{3}E, enquanto o passo III se constitui no algoritmo de otimização propriamente dito.

Passo I - Obter entradas do C^{3}E a partir da agregação de classificadores: a saída (*output*) de um classificador, q_1, para o objeto $\mathbf{x}_i \in \mathcal{X}$ é um vetor c-dimensional que captura a distribuição de probabilidades de classes $\boldsymbol{\pi}_i^{(q_1)}$. Mais especificamente, este vetor denota as probabilidades atribuídas a cada uma das classes de um dado problema. Tais probabilidades podem refletir tanto atribuições rígidas de classes (caso específico no qual os vetores são formados por valores binários) ou suaves (caso geral). A partir do conjunto de tais vetores, $\{\boldsymbol{\pi}_i^{(q_1)}\}_{q_1=1}^{r_1}$, um vetor médio pode ser computado para cada objeto $\mathbf{x}_i$ do conjunto-alvo:

$$\boldsymbol{\pi}_i = \frac{1}{r_1} \sum_{q_1=1}^{r_1} \boldsymbol{\pi}_i^{(q_1)}. \tag{3.1}$$

Passo II - Obter entradas do C^{3}E a partir da agregação de agrupadores: após a execução dos agrupadores, r_2 partições dos dados do conjunto-alvo estão disponíveis para se construir uma matriz de similaridades (co-associação) $\mathbf{S}$.

Dada uma dessas partições, (rígida) com n objetos, uma matriz binária e simétrica, $\mathbf{S}_{n\times n}^{(q_2)}$, — com o elemento $s_{ij}^{(q_2)}$ sendo igual a 1 se os objetos i e j estão no mesmo grupo e igual a 0 caso contrário — pode ser obtida[2]. Em outras palavras, dado que a q_2-ésima partição possui $k^{(q_2)}$ grupos, o resultado desse agrupamento pode ser denotado por um vetor de rótulos $\boldsymbol{\lambda}^{(q_2)} \in \mathbb{Z}_+^n$, e os elementos da respectiva matriz binária, $\mathbf{S}^{(q_2)}$, podem ser computados da seguinte forma:

$$s_{ij}^{(q_2)} = \begin{cases} 1 & (i,j) \in C_g(\boldsymbol{\lambda}^{(q_2)}) \text{ para algum } g \in \{1,2,...,k^{(q_2)}\} \\ 0 & \text{caso contrário.} \end{cases}, \tag{3.2}$$

onde $C_g(\boldsymbol{\lambda}^{(q_2)})$ é o g-ésimo grupo na q_2-ésima partição. A partir da média de todas as r_2 matrizes — cada uma delas produzida a partir de um componente do agregador — uma matriz de similaridades pode ser computada:

$$\mathbf{S} = \frac{1}{r_2} \sum_{q_2=1}^{r_2} \mathbf{S}^{(q_2)}. \tag{3.3}$$

[2]Uma matriz de similaridades pode também ser definida para partições suaves (*soft*) ou probabilísticas — *e.g.*, Punera & Ghosh (2008).

Passo III - Obter resultados de classificação consolidados a partir do C³E: após definir as entradas para o C³E ($\{\boldsymbol{\pi}_i\}_{i=1}^n$ e $\mathbf{S}$), o problema de se combinar classificadores e agrupadores pode ser formulado como um problema de otimização cujo objetivo é minimizar J na Equação (3.4) com relação ao conjunto de vetores de probabilidades $\{\mathbf{y}_i\}_{i=1}^n$, no qual $\mathbf{y}_i$ é a estimativa refinada da distribuição de probabilidades de classes *a posteriori* para um determinado objeto de $\mathcal{X}$:

$$J = \frac{1}{2}\sum_{i\in\mathcal{X}}\|\mathbf{y}_i - \boldsymbol{\pi}_i\|^2 + \alpha\frac{1}{2}\sum_{(i,j)\in\mathcal{X}} s_{ij}\|\mathbf{y}_i - \mathbf{y}_j\|^2 \tag{3.4}$$

Mantendo $\{\mathbf{y}_j\}_{j=1}^n \setminus \{\mathbf{y}_i\}$ fixos, é possível minimizar J na Equação (3.4) para todo $\mathbf{y}_i$, fazendo:

$$\frac{\partial J}{\partial \mathbf{y}_i} = \mathbf{0}. \tag{3.5}$$

Considerando a propriedade de simetria da matriz de similaridades $\mathbf{S}$ e observando que $\frac{\partial\|\mathbf{x}\|2}{\partial\mathbf{x}} = 2\mathbf{x}$, obtém-se:

$$\mathbf{y}_i = \frac{\boldsymbol{\pi}_i + \alpha'\sum_{j\neq i} s_{ij}\mathbf{y}_j}{1 + \alpha'\sum_{j\neq i} s_{ij}}, \tag{3.6}$$

onde $\alpha' = 2\alpha$ é usado por conveniência matemática. A equação (3.6) pode ser computada iterativamente, para todo $i \in \{1, 2, ..., n\}$, até um número máximo de iterações (I), ao passo que também são obtidas as distribuição de probabilidades de classes *a posteriori* para as instâncias em $\mathcal{X}$.

3.2.1 Estimando os parâmetros do C³E

Os parâmetros definidos pelo usuário em C³E, α e I, podem ser estimados por meio do algoritmo *Dynamic Differential Evolution* (D²E) proposto em Coletta *et al.* (2015b), o qual estende o algoritmo DE (Storn & Price, 1997; Price, 1996) realizando a amostragem de seus parâmetros, F e Cr, quando, em duas gerações consecutivas, não ocorrerem mudanças no *fitness* médio da população corrente. Para isso, o algoritmo D²E utiliza a seguinte regra:

$$<F, Cr>_{G+1} = \begin{cases} <d_3, d_4>, & \text{if } (\bar{f}_G = \bar{f}_{G-1}) \\ <F, Cr>_G, & \text{caso contrário} \end{cases} \tag{3.7}$$

onde $d_3 \in [0.1, 1]$, $d_4 \in [0, 1]$ são valores aleatoriamente amostrados a partir de uma distribuição uniforme, $\bar{f}_G$ e $\bar{f}_{G-1}$ representam o *fitness* médio da população nas gerações G and $G-1$, respectivamente. O algoritmo se inicia com parâmetros de controle fornecidos pelo usuário (e.g., população inicial — *number of parents* — $Np = 20$, *weighting factor* $F = 0{,}5$ e *crossover constant* $Cr = 0{,}9$). O D²E, então, permite que F e Cr assumam novos valores caso não

ocorram melhorias em duas iterações subsequentes. Esta ideia evita mínimos locais (com valores suficientemente altos de F e/ou Cr), bem como pode-se enfatizar busca local (com valores suficientemente baixos de F e/ou Cr). Esta dinâmica minimiza o impacto dos valores iniciais (definidos pelo usuário) tanto para F, como para Cr. Em outras palavras, esta adaptação permite uma boa relação entre exploração global e local, tornando o algoritmo mais robusto aos valores iniciais de seus parâmetros de controle. D^{2}E é baseado na estratégia "DE/best/2/bin" (Price, 1996) e começa com $Np = 20$, $F = 0{,}25$, e $Cr = 0{,}25$.

Para otimizar os parâmetros do C^{3}E-SL, o D^{2}E busca a minimização do erro de classificação apresentado pelo algoritmo no conjunto de validação pela combinação de diferentes valores de $\alpha = \{0, 0.001, 0.002, ..., 0.15\}$ e $I = \{1, 2, ..., 10\}$. Na sequencia, C^{3}E-SL com α^* e I^* pode ser usado para classificar novos *tweets* no conjunto de teste (*target/test set*).

3.3 Experimentos

Experimentos foram conduzidos a fim de avaliar as melhorias alcançadas quando um classificador SVM é combinado com um *ensemble de agrupadores* por meio do algoritmo C^{3}E-SL (descrito na Seção 3.2). Comparações foram feitas considerando o classificador SVM e suas saídas refinadas pelo algoritmo C^{3}E-SL. Adicionalmente, para efeitos de comparação, os melhores resultados de classificação relatados na literatura são também apresentados.

Para os experimentos realizados foram utilizados os conjuntos de *tweets* Sanders, Stanford, Debate Obama-McCain e Reforma da Saúde (*Health Care Reform*). Estes são conjuntos representativos, obtidos a partir de diferentes assuntos e disponíveis na literatura correlata (Hassan Saif & Alani, 2013). Tais conjuntos de dados foram descritos em detalhes na seção 2.3.3. O pré-processamento também foi feito de acordo com os passos já mencionados no Capítulo 2 e o leitor interessado pode recorrer ao mesmo para uma maior explicação.

Delimitando o cenário experimental, nesta fase de estudo foram escolhidas as melhores configurações de atributos – no caso a combinação de *bag-of-words* e léxicos – bem como o algoritmo de classificação SVM a partir dos experimentos realizados no capítulo anterior.

3.3.1 SVM e a configuração do *ensemble* de agrupadores

Por simplicidade e com o objetivo de oferecer uma prova de conceito, os resultados reportados neste capítulo foram alcançados usando um *ensemble* formado por um único componente — SVM não linear com *kernel* RBF. Os parâmetros para o *kernel* RBF, C e γ, foram estimados por *grid-search* no conjunto de treinamento de acordo com Hsu *et al.* (2010). Para o componente de agrupamento, foi construído uma matriz de similaridade a partir de cinco partições induzidas pelo uso do algoritmo de agrupamento *K-Medoids* (Kaufman & Rousseeuw, 1987) com similaridade cosseno. Baseado em Kuncheva *et al.* (2006) e Kuncheva & Hadjitodorov (2004), cada partição possui um valor específico para o número de grupos $k = \{2{,}3{,}5{,}6{,}8\}$. Estes valores foram aleatoriamente selecionados e fixados em todos os conjuntos de dados.

3.3.2 Otimização dos parâmetros do algoritmo C^{3}E-SL

A otimização dos parâmetros do algoritmo C^{3}E-SL, α and I, tem por objetivo maximizar a acurácia da classificação. Procura-se então por um par de valores ótimos (α^* and I^*). Para estimar estes valores, foi empregado o algoritmo D^{2}E (abordado na Seção 3.2.1).

O procedimento adotado envolve três conjuntos de dados distintos: treinamento, validação e teste (Witten & Frank, 2005). O conjunto de teste (com instâncias a serem classificadas) não é utilizado no processo de construção do *ensemble* de classificadores — *i.e.*, ele é um conjunto de *tweets* independente não usado para otimizar nenhum parâmetro do algoritmo. Desta forma, como usual, e em conformidade com aplicações reais (em classificação de *tweets*), somente os conjuntos de treinamento e validação são usados para otimizar os parâmetros de C^{3}E-SL. Neste Capítulo, o conjunto de validação é formado por metade do conjunto de treinamento disponível. Com isso, primeiramente um modelo de classificação a partir do SVM é construído com 50% do conjunto de treinamento. Então, D^{2}E estima α^* e I^*, minimizando o erro de classificação no conjunto de validação (formado com os outros 50% do conjunto de treinamento disponível), no qual o *ensemble* de agrupadores foi induzido. Na sequencia, os valores estimados dos parâmetros, α^* e I^*, são finalmente empregados para avaliar a acurácia no conjunto de teste, no qual novamente e como uma requisição do C^{3}E-SL, um *ensemble* de agrupadores similar ao que foi construído para o conjunto de validação deve ser induzido.

Para o conjunto de dados *Health Care Reform* (Saif *et al.*, 2012b) e *Stanford* (Go *et al.*, 2009), os resultados foram obtidos a partir de 10 execuções do mesmo conjunto de treinamento e teste. Em particular, para o conjunto *Stanford* (Go *et al.*, 2009), diferentes conjuntos de treinamento (com diferentes tamanhos) a partir do conjunto de treinamento original foram amostrados. Para os conjuntos de dados *Obama-McCain Debate (OMD)* e *Sanders*, os resultados também foram obtidos a partir de 10 execuções, mas considerando 2×5 *cross-validation* (validação cruzada) (Witten & Frank, 2005).

3.3.3 Resultados

A Tabela 3.1 mostra os resultados alcançados com o algoritmo SVM e seus refinamentos pelo uso do algoritmo C^{3}E-SL em cada conjunto de dados. Os melhores resultados na literatura também são reportados. O *F-Score* (F1) também foi mostrado devido ao desbalanceamento dos dados. Pode-se observar que a combinação de um classificador com agrupadores, como feito pelo algoritmo C^{3}E-SL, apresenta resultados competitivos em comparação com o uso do classificador SVM, o qual é considerado um dos classificadores mais utilizados no estado da arte de análise de sentimentos em *tweets* (TSA), conforme mencionado no Capítulo 2.

Para o conjunto de dados *HCR*, o C^{3}E-SL apresentou relevante melhora em relação aos resultados apresentados pelo algoritmo SVM. Além disso, apresentou melhores resultados se comparados aos reportados na literatura (Speriosu *et al.*, 2011; Saif *et al.*, 2012b).

Para os conjuntos de dados *OMD* e *Sanders*, é possível observar melhoras suaves, próximas dos resultados apresentados na literatura (Hu *et al.*, 2013b; Saif *et al.*, 2012b; Ziegelmayer

Conjunto de dados	Acurácia (%)		F1 - Classe Positiva (%)		F1 - Classe Negativa (%)	
	SVM	C^{3}E-SL	SVM	C^{3}E-SL	SVM	C^{3}E-SL
Health Care Reform (HCR)	74,29	**79,62**	**43,56**	32,13	83,35	**88,00**
Literatura (Speriosu *et al.*, 2011; Saif *et al.*, 2012b)	71,20		50,30		86,00	
Obama-McCain Debate (OMD)	75,15	**75,18**	80,14	**80,58**	**67,24**	64,24
Literatura (Hu *et al.*, 2013b; Saif *et al.*, 2012b)	76,30		70,30		85,40	
Sanders	82,11	**82,15**	80,43	**80,80**	**83,48**	83,27
Literatura (Ziegelmayer & Schrader, 2012)	84,40		-		-	
Stanford - treinado com 500 tweets	74,65	**77,69**	74,65	**77,33**	74,65	**78,02**
Stanford - treinado com 1.000 tweets	77,72	**79,69**	77,40	**80,21**	78,02	**79,13**
Stanford - treinado com 2.000 tweets	**78,83**	77,80	79,79	**80,92**	**77,78**	73,46
Stanford - treinado com 3.000 tweets	79,39	**81,84**	79,56	**82,36**	79,21	**81,27**
Literatura (Bakliwal *et al.*, 2012; Saif *et al.*, 2012b)	87,20*		82,50*		85,30*	

* Usando o conjunto todo como conjunto de treinamento — i.e., 1,6 milhões tweets.

Tabela 3.1: Acurácia na Classificação (%) e *F-Scores* (%) obtidos usando SVM e C^{3}E-SL em cada base de dados. Para o conjunto de dados *Stanford*, foram amostrados conjuntos de treinamentos de diferentes tamanhos (melhores resultados em negrito). Para efeitos de comparação, os melhores resultados de classificação relatados na literatura são também apresentados.

& Schrader, 2012). Considerando a classe positiva e a medida F1 para a base de dados *OMD*, C^{3}E-SL alcançou melhor resultado (80,58%) do que o melhor resultado encontrado na literatura (70,30%).

Para o conjunto de dados *Stanford*, foram realizados experimentos em quatro conjuntos diferentes de treinamento (amostrados com 500, 1.000, 2.000 e 3.000 conjuntos de tweets balanceados a partir do conjunto de treinamento original). Os refinamentos providos pelo C^{3}E-SL evidenciaram melhores resultados do que com o uso do classificador SVM em todos estes cenários (exceto para o conjunto de treinamento com 2.000 tweets). C^{3}E-SL também mostrou resultados competitivos se comparados aos trabalhos apresentados na literatura. Entretanto, é importante enfatizar que os resultados reportados em Bakliwal *et al.* (2012); Saif *et al.* (2012b) foram obtidos a partir do conjunto de treinamento completo (com 1,6 milhões de tweets). Sendo assim, os resultados são muito encorajadores, sugerindo que o algoritmo C^{3}E-SL é promissor para análise de sentimentos.

Os trabalhos relacionados utilizam de mais informações oriundas da estrutura da rede social (Speriosu *et al.*, 2011; Saif *et al.*, 2012b), recursos linguísticos incorporados com o auxílio de especialistas e correlações entre conceitos semânticos e sentimentos (Saif *et al.*, 2012b), bem como a necessidade de bases gigantescas de treinamento (Bakliwal *et al.*, 2012; Saif *et al.*, 2012b) para a apuração de índices razoáveis de acurácia. Com base nestes apontamentos é importante mencionar a importância dos resultados obtidos com o C^{3}E-SL, uma vez que para

tanto apenas atributos do tipo *unigramas*, bem como léxicos de propósito geral foram utilizados. Tendo em vista uma comparação justa entre apenas os algoritmos SVM e C^{3}E-SL, o algoritmo objeto de validação, C^{3}E-SL ganhou em $85{,}00\%$ dos testes realizados, considerando as métricas acurácia e F1 para a classe positiva. Considerando o F1 para a classe negativa o ganho do algoritmo C^{3}E-SL também é relevante, uma vez que em $57{,}15\%$ dos casos o mesmo supera o algoritmo SVM.

3.4 Considerações Finais

A agregação de classificadores e agrupadores tem como motivação as diversas vantagens inerentes a cada um desses modelos (Acharya *et al.*, 2014, 2011; Cai *et al.*, 2009; Gao *et al.*, 2009), e, em geral, os dados são primeiramente agrupados e posteriormente classificados a partir da estrutura de grupos obtida. Em análise de sentimentos este cenário se repete, uma vez que o agrupamento é realizado somente como uma estratégia de pré-processamento (Xiang & Zhou, 2014). Realizar essas duas tarefas (a agregação de classificadores e agrupadores) simultaneamente traz como vantagem a possibilidade dos modelos não supervisionados aumentarem a precisão e a capacidade de generalização dos classificadores. Baseado nisso, o algoritmo C^{3}E, proposto por Acharya *et al.* (2014, 2011) foi aplicado à análise de sentimentos em *tweets* nesta tese com o objetivo de comprovar os benefícios que já vêm sendo evidenciados em outras aplicações, ou seja, que a agregação de classificadores e agrupadores, feita de forma simultânea, de fato contribui para a obtenção de uma classificação consolidada dos dados a partir da combinação de dois modelos: um modelo resultante da agregação de classificadores (Oza & Tumer, 2008; Kuncheva, 2004) e outro resultante da agregação de agrupadores (Acharya *et al.*, 2011; Ghaemi *et al.*, 2009).

Os resultados mencionados neste Capítulo evidenciam a veracidade e validade da segunda hipótese de pesquisa, i.e., é verdade que o conhecimento inferido no conjunto-alvo ao fazer uso de um agrupador de dados permite refinar o modelo de classificação de sentimentos inferido no conjunto de treinamento, fornecendo resultados competitivos ou superiores aos resultados inferidos a partir de classificadores derivados apenas de fontes de informações supervisionadas. Tais resultados foram publicados em da Silva *et al.* (2014a).

Capítulo

4

Aprendizado semissupervisionado agregando classificadores e agrupadores

No aprendizado supervisionado, considera-se que o conjunto de treinamento contém apenas objetos rotulados. Deste modo, os objetos podem ser vistos como tuplas $< \mathbf{x}_i^T, C_\ell^i >$ que relacionam a ℓ-ésima classe com o i-ésimo objeto do conjunto de treinamento $\mathcal{X}^T$. A tarefa de classificação, então, tem como objetivo realizar a predição da classe de novos objetos (não rotulados). Em outras palavras, busca-se inferir uma função f de tal forma que $\{<\mathbf{x}_i^T, C_\ell^i>, i = 1, ..., N; \ell = 1, ..., c\} \mapsto f$, a qual espera-se ser apropriada para rotular objetos não observados. O algoritmo C^{3}E-SL (apresentado na Sessão 3.2) tem sido usado para induzir (refinar) modelos supervisionados. Entretanto, na prática, pode-se encontrar obstáculos para este tipo de aprendizagem. Em muitas aplicações reais (e.g. análise de sentimentos em *tweets*), embora seja relativamente fácil adquirir dados não rotulados, o esforço de rotulação é tipicamente caro. Este fato motivou o surgimento do que se convencionou a chamar de aprendizado *semissupervisionado* (Goldberg, 2010; Zhu & Goldberg, 2009; Chapelle *et al.*, 2006).

No aprendizado semissupervisionado, presume-se que o conjunto de treinamento possui, além de objetos rotulados, também objetos que não estão rotulados (sendo estes últimos comumente em maior quantidade). Este tipo de aprendizado pode ser definido como *indutivo* ou *transdutivo*:

Caso Indutivo: busca-se predizer a classe de novos objetos, como é feito normalmente em aprendizado supervisionado, porém presumindo-se que se dispõe também dos objetos não rotulados $\{\mathbf{x}_i^T, i = N+1, ..., U\}$ no conjunto de treinamento. O problema consiste, então, em inferir uma função f de tal forma que $\{<\mathbf{x}_i^T, C_\ell^i>, i = 1, ..., N; \ell = 1, ..., c\} \cup \{\mathbf{x}_i^T, i = N + 1, ..., U\} \mapsto f$.

Caso Transdutivo: objetiva-se apenas predizer a classe dos objetos não rotulados no próprio

conjunto de treinamento de tal forma que $\{<\mathbf{x}_i^T, C_\ell^i>, i = 1, ..., N; \ell = 1, ..., c\} \cup \{\mathbf{x}_i^T, i = N+1, ..., U\} \mapsto \{C_\ell^i, i = N+1, ..., U\}$.

A partir da formalização apresentada, este capítulo introduz uma extensão do algoritmo C^{3}E-SL para o contexto semissupervisionado indutivo, com foco em análise de sentimentos em *tweets*. Dados não rotulados são fontes de conhecimento pouco exploradas neste cenário, e dada a sua imensa disponibilidade, possuem grande potencial de agregar valor às tradicionais formas de classificação supervisionada.

A abordagem semissupervisionada (SSL) desenvolvida estende o estudo apresentado no Capítulo 3, no qual a combinação de classificadores e agrupadores foi usada para classificar *tweets* de maneira supervisionada. Esta extensão utiliza o algoritmo C^{3}E-SL para combinar máquinas de vetores de suporte (*Support Vector Machines* — SVMs (Cortes & Vapnik, 1995; Boser *et al.*, 1992)) com a informação de similaridades entre *tweets* de um conjunto alvo que se deseja classificar. Neste contexto, o C^{3}E-SL se torna parte de um processo iterativo de *self-training* (Zhu & Goldberg, 2009), no qual predições mais confiáveis aprimoram o modelo de classificação final. Visto que esta ideia atua com uma quantidade relativamente pequena de dados rotulados, o classificador gerado é especialmente útil para a classificação de *tweets*. Em particular, o experimento ilustrativo da Seção 4.1.3 mostra como a informação não supervisionada do C^{3}E-SL auxilia também na aprendizagem semissupervisionada. Experimentos comparativos mais elaborados são reportados e discutidos a partir da Seção 4.2.

4.1 Abordagens para Aprendizado Semissupervisionado

Aprendizado semissupervisionado (SSL) situa-se em um meio termo entre aprendizado supervisionado e não supervisionado, uma vez que se aproveita tanto de dados rotulados quanto não rotulados para induzir modelos preditivos (Goldberg, 2010). Neste contexto, Nigam *et al.* (2000) projetaram um modelo generativo que utiliza o algoritmo *Expectation-Maximization* (EM) (Dempster *et al.*, 1977) para gerar misturas de Gaussianas a partir de dados não rotulados. Estas distribuições compensam a falta de conhecimento *a priori* e ajudam a melhorar a capacidade de generalização do modelo. Semelhantemente, modelos discriminativos têm sido também estendidos para o contexto semissupervisionado. Por exemplo, e como uma extensão das máquinas de vetores de suporte (Cortes & Vapnik, 1995; Boser *et al.*, 1992), *Semi-Supervised* SVMs (S3VMs) encontram uma fronteira de máxima margem usando dados rotulados e não rotulados (Xu *et al.*, 2008; Sindhwani & Keerthi, 2006; Joachims, 1999b,a).

A teoria dos grafos também tem ajudado a modelar aprendizado semissupervisionado. Vários estudos da literatura têm representado objetos rotulados e não rotulados como sendo vértices de um grafo (Zhu, 2005; Belkin *et al.*, 2004; Zhou *et al.*, 2004; Zhu *et al.*, 2003a; Jaakkola & Szummer, 2002; Blum & Chawla, 2001). Arestas (com pesos) são, então, usadas para definir as similaridades entre pares de vértices. A partir desta representação, algoritmos basicamente propagam a informação de vértices rotulados para os demais vértices (assumindo-se suavidade na propagação) até que algum critério de convergência seja alcançado. Espera-se que, neste pro-

cesso, dois objetos *similares* possam (até certo ponto) compartilhar o mesmo rótulo de classe. Infelizmente, a maioria dos algoritmos baseados em grafos possuem ordem de complexidade computacional cúbica com o número de objetos, tornando a aplicação limitada a pequenas bases de dados (Zhu, 2005).

Modelos semissupervisionados também são reproduzidos como um processo iterativo que pode ter como base componentes supervisionados. Blum & Mitchell (1998), por exemplo, utilizam dois componentes supervisionados que "ensinam" cada um o outro de forma iterativa. Estes componentes operam em diferentes espaços de atributos, fazendo uma analogia às diferentes *visões* que se podem ter dos dados. Os objetos com maior confiança nas predições de cada um dos componentes são incorporados no conjunto de dados rotulados (o qual, então, é usado para retreinar os componentes). Este processo, denominado de *co-training*, tem sido estudado e melhorado por inúmeros autores (Wang & Zhou, 2010; Zhou & Li, 2005). Com ideia similar, algoritmos de *self-training* (Scudder, 1965) — usam predições de um único componente supervisionado para melhorar o modelo. A aprendizagem se inicia com um classificador treinado a partir de (poucos) dados rotulados. A cada iteração, após a predição de novos dados, promove-se para o conjunto de dados rotulados os objetos para os quais há grande confiança sobre as estimativas dos rótulos de classe. A partir disso, o classificador é reconstruído e usado para classificar novos dados numa próxima iteração. Esta prática funciona bem se os objetos promovidos estão atrelados aos rótulos de classes corretas (Zhu & Goldberg, 2009). Além disso, *self-training* se destaca por ser uma abordagem bem flexível, na qual praticamente qualquer algoritmo (supervisionado) pode ser usado como componente.

4.1.1 Aprendizado semissupervisionado aplicado à análise de sentimentos

Tipicamente, em análise de sentimentos em *tweets*, os algoritmos para aprendizado semissupervisionado são baseados em: ***grafos*** (Tan *et al.*, 2011; Pozzi *et al.*, 2013; Johnson *et al.*, 2011), ***self-training*** (Becker *et al.*, 2013; Hong *et al.*, 2014; Baugh, 2013; Drury *et al.*, 2011), ***co-training*** (Wan, 2009; Li *et al.*, 2011; Yu, 2014; Liu *et al.*, 2013b) e ***tópicos*** (Xiang & Zhou, 2014).

SSL baseado em grafos

De modo geral, os métodos de classificação de sentimentos semissupervisionados baseados em grafos utilizam de informações da rede social (tais como conexões entre seus usuários) para propagar conhecimento ao longo do grafo que representa a rede social.

Tan *et al.* (2011) propõem como escopo de trabalho a análise de sentimentos de cada usuário, uma tarefa em que a opinião global do usuário é determinada com base em uma coleção de *tweets* postados por ele em um período de tempo e também nas conexões da rede de usuários à qual o mesmo pertence. É pressuposto pelos autores a existência de um tópico de interesse e um grafo, no qual uma proporção relativamente pequena de usuários já tenham sido rotulados com suas respetivas opiniões. A tarefa é predizer os rótulos de classe para os usuários não rotulados do grafo em questão. A principal motivação apresentada pelos autores é que os usuários

que estão de alguma forma "conectados"podem ser mais propensos a ter opiniões semelhantes (princípio da homofilia) (Lazarsfeld & Merton, 1954; McPherson *et al.*, 2001; Thelwall, 2010). Sendo assim, informações de relacionamento podem complementar o que é extraído a partir das fontes textuais providas pelos usuários. Observe que o problema tratado nesta tese é o de obter a opinião para cada *tweet* e não o de inferir a polaridade do usuário em relação à um tópico como o tratado pelos autores mencionados.

Pozzi *et al.* (2013) argumentam que apesar de as relações sociais desempenharem um papel importante em análise de sentimentos em *microblogs*, considerar as conexões de amizade é uma estratégia fraca para modelagem da homofilia, uma vez que dois amigos podem não compartilhar a mesma opinião sobre um determinado tópico. Por esta razão, os autores propõem modelar a classificação da polaridade do usuário através da integração de conteúdos textuais e das relações de aprovação (por exemplo, o botão curtir ou *like* no Facebook e a opção *retweet* no Twitter), com o objetivo de melhor representar o princípio da homofilia. Assim como no trabalho de Tan *et al.* (2011), o objetivo do trabalho proposto por Pozzi *et al.* (2013) é o de inferir a polaridade do usuário em relação a um tópico e não o de obter a opinião para cada *tweet*.

Jiliang *et al.* (2015) propõem uma abordagem semissupervisionada cuja rede é construída com base em um léxico, em *tweets* rotulados e não rotulados, além de teorias sociais, como o contágio emocional (Hatfield *et al.*, 1993) e a consistência dos sentimentos (Abelson, 1983). A motivação, segundo os autores, é que um *tweet* tem a probabilidade de ser positivo se ele contém muitas palavras com sentimento positivo, e se ele está conectado a outros *tweets* (conectados por *retweets*, por exemplo) que também possuem a classe de sentimento positivo. Analogamente, uma palavra provavelmente possui um sentimento positivo se ela está associada com um certo número de *tweets* com sentimento positivo também, além de se correlacionar com outras palavras que possuem já associado o rótulo de sentimento positivo. Este raciocínio também pode ser aplicado a outras classes de sentimento, como negativo e neutro. Os autores denominam este fenômeno de reforço mútuo entre os sentimentos dos *tweets* e das palavras que o compõem, e o processo de propagação é baseado neste conceito (ver Figura 4.1).

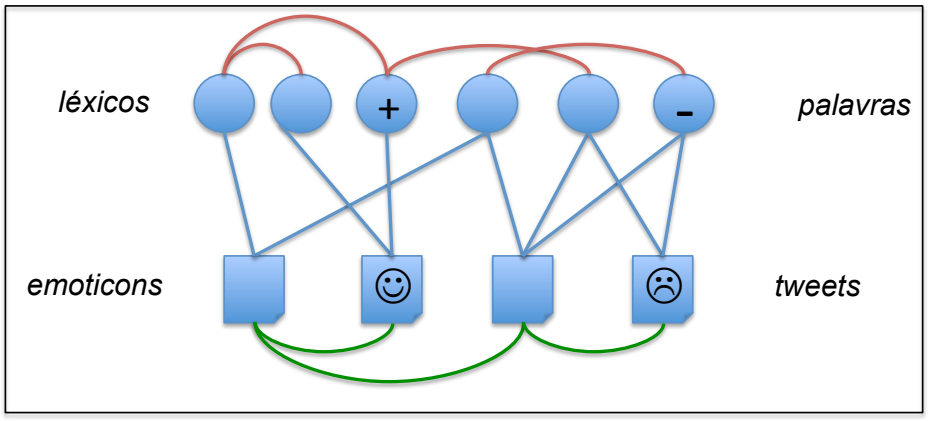


Figura 4.1: Adaptada de Jiliang *et al.* (2015)

O trabalho de Matsuno *et al.* (2015) realiza a análise de sentimentos com granularidade de aspectos (esta tarefa foi apresentada na seção 1.1) e utiliza o aprendizado semissupervisionado

tanto para a classificação de aspectos quanto para a classificação de sentimentos do aspecto em questão. Os autores propõem o *framework* denominado ASPHN (*Aspect-Based Sentiment Propagation on Heterogeneous Networks*), em que são integrados vários tipos de informações como vértices em uma única rede: atributos linguísticos, atributos candidatos a aspectos, atributos gramaticais, dentre outros. O aprendizado é baseado na propagação da informação de exemplos rotulados por meio das relações topológicas entre os vértices. Os autores aplicaram o método em *tweets* cujos tópicos são relacionados a restaurantes e laptops.

SSL baseado em *self-training*

O *self-training* é uma das abordagens semissupervisionadas mais aplicadas à análise de sentimentos em *tweets* (Zimmermann *et al.*, 2014; Becker *et al.*, 2013; Baugh, 2013; Yu, 2013; He & Zhou, 2011; Drury *et al.*, 2011; Qiu *et al.*, 2009).

Becker *et al.* (2013) propõem o *framework* denominado *AVAYA*, o qual foi aplicado à análise de sentimentos em *tweets* com granularidade de aspectos e de documento. Os autores avaliaram sua abordagem em uma competição conhecida como SemEval (Nakov *et al.*, 2013), a qual provê um conjunto de dados de treinamento e alguns conjuntos de testes, que são distribuídos à todos os participantes do evento. De posse dos mesmos conjuntos de dados, é possível comparar e avaliar as abordagens propostas. Considerando as duas bases de testes apresentadas pelos organizadores, Becker *et al.* (2013) alcançaram a 5ª posição nos dois conjuntos de testes, para a granularidade de documento. Para a granularidade de aspectos, os autores alcançaram o 2º e 3º lugares respectivamente. A estratégia usada por eles foi considerar todo o conjunto de treinamento fornecido pela organização do evento como dados rotulados, e coletar um conjunto extra de 475.000 *tweets* a serem usados como dados não rotulados. Os autores utilizaram de atributos do tipo *bag-of-words*, léxicos, atributos extraídos do *microblog*, atributos *part-of-speech tag* e sintáticos.

Baugh (2013) também participou da competição SemEval (Nakov *et al.*, 2013) propondo um método hierárquivo em que inicialmente os *tweets* são classificados segundo a sua subjetividade, ou seja, são classificados quanto à presença (*tweets* subjetivos) ou ausência de sentimentos (*tweets* objetivos ou neutros). Na sequência os *tweets* classificados anteriormente como subjetivos são novamente classificados, mas agora quanto à sua polaridade (*tweets* positivos ou negativos). O autor utiliza um procedimento iterativo com base em *self-training* para o classificador de polaridade. Assim como na abordagem de Becker *et al.* (2013), Baugh (2013) utilizou o conjunto de treinamento fornecido pela competição como dados rotulados e um conjunto de 910.000 *tweets*, coletado pelo próprio autor, como dados não rotulados. Para a granularidade de documentos, os autores alcançaram o 8º e 13º nas duas bases de dados de teste fornecidas pelos organizadores, sendo que ao todo 44 equipes participaram da tarefa.

He & Zhou (2011) e Qiu *et al.* (2009) propõem métodos de análise de sentimento semelhantes: ambos utilizam de um léxico para rotulação de um conjunto de treinamento, ou seja, um conjunto inicial de mensagens é classificado de acordo com um dicionário de sentimentos. Este conjunto inicial é então utilizado em um processo iterativo baseado em *self-training* para

classificar os casos em que o dicionário utilizado não foi suficiente.

Zimmermann *et al.* (2014) abordam o problema de classificação de sentimentos face o desafio de aprender a partir de um conjunto limitado de dados em um ambiente de fluxo contínuo. Os autores utilizam um método baseado em *self-training* adaptado para o cenário de fluxo contínuo de dados com o objetivo de classificar *tweets* quanto ao seu sentimento. Os mesmos mostraram que os resultados alcançados possuem desempenho comparável ou ainda melhor do que considerando a base de dados totalmente rotulada em um aprendizado supervisionado.

Drury *et al.* (2011) também propõem um classificador semissupervisionado baseado em um procedimento de *self-training*. Contudo, as amostras rotuladas são usadas para a construção de um classificador baseado em regras, as quais são inferidas a partir de conhecimento linguístico. Os autores avaliaram o método proposto em comentários (*reviews*) sobre refeições aéreas, professores universitários e música.

Transferência de conhecimento entre domínios em análise de sentimentos é o objeto de pesquisa de Yu (2013). Este estudo investigou a viabilidade do *self-training* para lidar com o problema de transferência de conhecimento entre domínios, aplicando dados rotulados de um domínio em contextos diferentes e desconhecidos. O autor avalia sua abordagem em dados oriundos de comentários de filmes, reportagens e *tweets*, e enfatiza o quão promissor é o *self-training* nesses casos.

SSL baseado em *co-training*

Wan (2009) foi o precursor nos estudos da aplicação do algoritmo *co-training* em análise de sentimentos. O autor direciona esforços para tratar o problema da classificação de sentimentos *cross-lingual*, o qual aproveita um *corpus* que está em inglês como base de treinamento, para classificar textos em chinês. A tradução automática é usada para eliminar a diferença de linguagem entre o conjunto de treinamento e o conjunto de teste, bem como os atributos que serão obtidos a partir do inglês e a partir do chinês. Os atributos extraídos do inglês e do chinês são considerados como duas visões independentes do problema de classificação. As duas visões (inglês e chinês) cooperaram entre si para a classificação de instâncias não rotuladas — que estão em chinês (ver Figura 4.2).

Aplicar uma estratégia baseada em *co-training* para análise de sentimentos em *tweets* foi a proposta de Liu *et al.* (2013b). Neste caso, os autores consideraram uma visão construída a partir de características textuais com o uso de um dicionário de sentimentos e uma outra visão que inclui *emoticons*, características temporais e pontuação. Liu *et al.* (2013a) e Liu *et al.* (2015) são continuações do trabalho de Liu *et al.* (2013b) em que os autores aperfeiçoam a técnica utilizando um maior conjunto de atributos textuais, bem como realizando mais testes e enfatizando a possibilidade de o algoritmo *co-training* ser uma técnica que provê uma maior adaptabilidade a novos tópicos. Segundo os autores, tal método obteve desempenho superior a algoritmos supervisionados conhecidos e a *ensembles* de classificadores.

Recentemente, Carter & Inkpen (2015) apresentam uma adaptação do algoritmo *co-training* para identificar, simultaneamente, tanto aspectos de um produto (características) expressas em

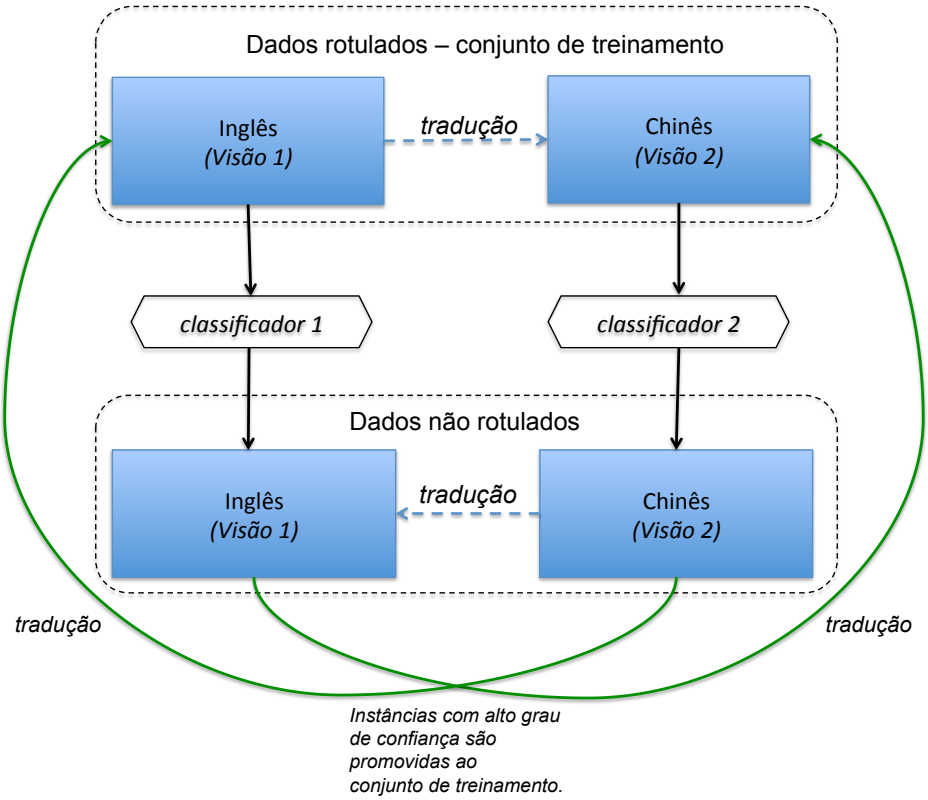


Figura 4.2: *Co-training* segundo Wan (2009)

textos fornecidos por *tweets* quanto os sentimentos expressos sobre tais aspectos. Os autores utilizam de duas visões, uma baseada em lexemas[1] e outra baseada em características sintáticas.

Yang *et al.* (2015) propõem uma abordagem baseada em *deep learning* (Bengio *et al.*, 2015; Ranzato *et al.*, 2015; Zhang & LeCun, 2015; Tang *et al.*, 2015a,b; Tang, 2015) para inferência de uma visão, bem como de um dicionário de sentimentos (léxico) para inferência de uma segunda visão. As duas visões colaboram entre si para a obtenção de um classificador consolidado para análise de sentimentos em *tweets* e comentários sobre filmes escritos em inglês, bem como outras fontes textuais escritas em chinês.

SSL baseado em tópicos

Como já abordado no Capítulo 3, Xiang & Zhou (2014) propõem uma abordagem semi-supervisionada baseada em tópicos que é usada para análise de sentimentos em *tweets*. Os autores propõem construir um modelo de tópicos nos dados rotulados. A informação de tópicos é obtida a partir da implementação do principal modelo de tópicos da literatura, conhecido como *Latent Dirichlet Allocation* (LDA) e proposto em Blei *et al.* (2003). A partir dos tópicos obtidos, um modelo de sentimentos específico é construído em cada grupo encontrado. Embora Xiang & Zhou (2014) indiquem que as tarefas de classificação e agrupamento possam ser complementares, tais abordagens foram implementadas de forma sequencial, o que nem sempre favorece a otimalidade dos dois tipos de aprendizagem (Cai *et al.*, 2009), *i.e.*, o agrupamento

[1]Unidade mínima distintiva do sistema semântico de uma língua que reúne todas as flexões de uma mesma palavra. Em termos simplificados, é a parte de uma palavra que constitui uma unidade mínima dotada de significado lexical.

neste caso "trabalhou" apenas como uma ferramenta de pré-processamento para a tarefa de classificação.

A seguir, uma nova abordagem semissupervisionada é descrita, a qual explora estruturas intrínsecas dos dados para refinar modelos de classificação com o algoritmo C^{3}E-SL em um procedimento de *self-training*.

4.1.2 C^{3}E-SL Semissupervisionado

O algoritmo C^{3}E-SL, conforme abordado na Seção 3.2, recebe distribuições de probabilidades de classes que serão refinadas em um processo de otimização. Tipicamente, estas distribuições são resultantes de uma agregação de classificadores (Coletta *et al.*, 2015a,b; Acharya *et al.*, 2014, 2011), mas o uso de um único classificador operando de maneira independente também é possível. Mais detalhadamente, a versão semissupervisionada do C^{3}E-SL apresentada nesse trabalho refina resultados de um classificador SVM. Neste ponto, cumpre observar que o SVM possui grande capacidade de generalização, além de alcançar melhores resultados do que algoritmos comumente usados na área de análise de sentimentos — como o *Naive Bayes* e a Regressão Logística (Nakov *et al.*, 2013). Seguindo Mohammad *et al.* (2013), adotou-se uma SVM com um *kernel* linear e parâmetro $C = 0{,}005$.

Além das distribuições de probabilidades de classes para objetos de um conjunto alvo, o C^{3}E-SL requer também uma matriz de similaridades construída a partir deste conjunto. Em Coletta *et al.* (2015b), Acharya *et al.* (2014, 2011) e da Silva *et al.* (2014a) esta matriz provém de agregadores de agrupadores. Já na versão semissupervisionada do algoritmo apresentada nesta tese, a matriz é obtida diretamente dos dados brutos seguindo os três passos a seguir:

Passo I - Geração de uma matriz de distâncias (D): formada pelo cômputo das distâncias euclidianas entre pares de objetos presentes no conjunto alvo;

Passo II - Normalização da matriz D: para conter somente valores entre 0 e 1. A Equação (4.1) é usada para calcular a distância normalizada, $\overline{d_{ij}}$, para os objetos i e j, sendo d_{ij} a distância Euclidiana calculada no Passo I. Os termos d_{min} e d_{max} referem-se, respectivamente, às distâncias mínimas e máximas observadas.

$$\overline{d_{ij}} = \frac{d_{ij} - d_{min}}{d_{max} - d_{min}} \tag{4.1}$$

Passo III - Obtenção da matriz de similaridades (S): a matriz D normalizada pode ser convertida em uma matriz de similaridades, S, calculando-se para cada um de seus elementos:

$$s_{ij} = 1 - \overline{d_{ij}} \ . \tag{4.2}$$

Tendo definidas as entradas do C^{3}E-SL, a abordagem semissupervisionada do algoritmo é sumarizada nas Figuras 4.3–4.5. Em particular, a Figura 4.3 ilustra a calibração do C^{3}E-SL a partir de um procedimento de validação cruzada que estima os parâmetros α e I do algoritmo.

Ou seja, dividiu-se os objetos rotulados em um conjunto de treinamento e outro de validação. Um classificador SVM é induzido a partir do conjunto de treinamento e, então, o algoritmo D^{2}E (apresentado na Seção 3.2.1) é usado para estimar os valores "ótimos" dos parâmetros no conjunto de validação (no qual uma matriz de similaridades foi produzida usando-se os Passos I–III). Os valores de α^* e I^* ("ótimos") são, então, fixados e utilizados para a classificação de dados não rotulados nos conjuntos de teste (alvo). Esta etapa é apresentada na Figura 4.4. Como resultado, obtém-se um conjunto de vetores, $\{\mathbf{y}_j\}_{j=1}^{n}$, que são estimativas refinadas das distribuições de probabilidades de classes dos objetos do conjunto alvo. A partir destas estimativas, alguns objetos são promovidos para o conjunto de dados rotulados como parte da prática de *self-training* (Zhu & Goldberg, 2009). Portanto, em resumo, o C^{3}E-SL é empregado em uma estrutura algorítmica para desempenhar *self-training*, segundo o qual o modelo de classificação é realimentado com novos dados a partir de um processo iterativo que aproveita a informação não supervisionada para aumentar a confiança das predições. A Figura 4.5 apresenta o esquema completo do C^{3}E-SL semissupervisionado.

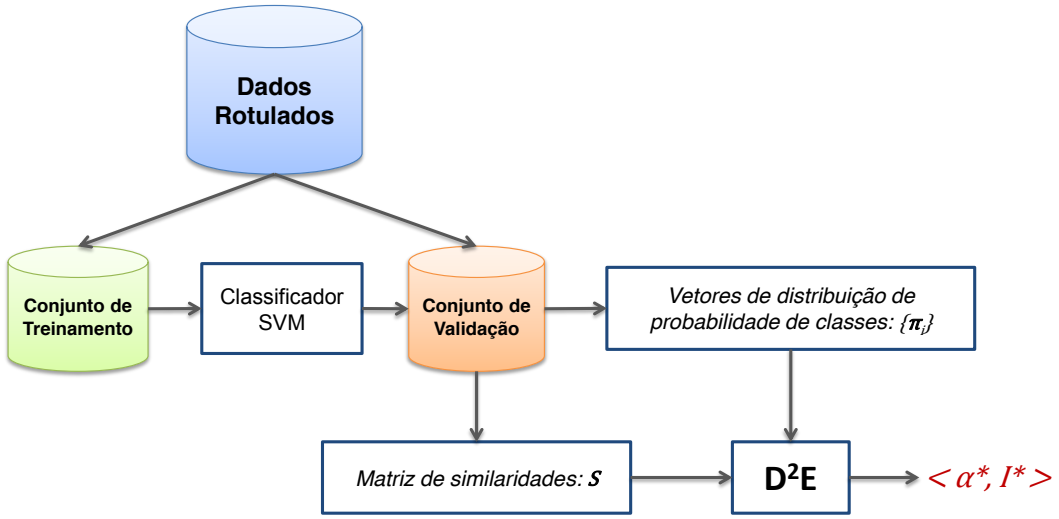


Figura 4.3: Calibração do C^{3}E-SL — parâmetros α e I são estimados em um conjunto de validação por meio do algoritmo D^{2}E (Coletta *et al.*, 2015b).

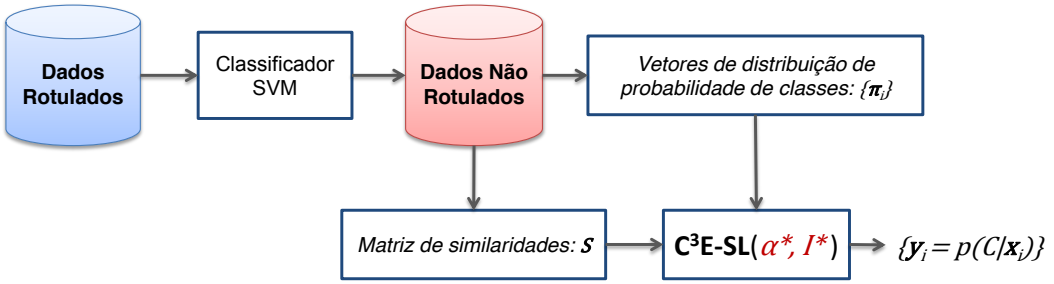


Figura 4.4: Os valores "ótimos" para α e I são fixados e usados no C^{3}E-SL para refinar resultados do SVM em um conjunto alvo (com dados não rotulados que se deseja predizer).

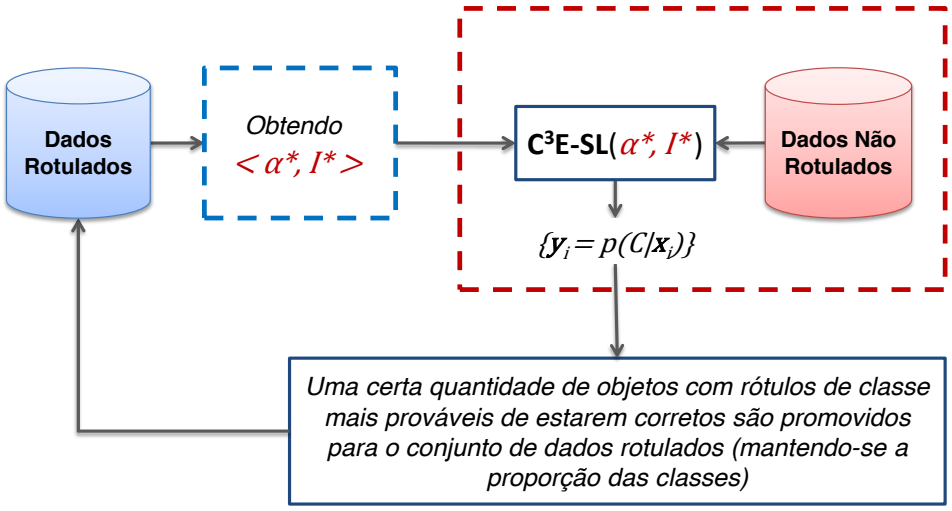


Figura 4.5: Esquema do C^{3}E-SL semissupervisionado.

4.1.3 Exemplo Didático

Esta seção ilustra o funcionamento do C^{3}E-SL semissupervisionado. A Tabela 4.1 descreve quatro *tweets* provenientes da base de dados *Twitter2014* (Rosenthal *et al.*, 2014). Na Figura 4.6, estes *tweets* são representados pelo número de palavras de sentimento positivo e negativo — estes atributos, como será visto a seguir, são atributos usados para se construir a matriz de similaridades (**S**). Mais especificamente, para t_1 há duas palavras de sentimento positivo: "*wins*" e "*earned*", e não há palavras de sentimento negativo. Para t_2 há duas palavras de sentimento positivo: "*better*" e "*better*", mas há duas palavras de sentimento negativo também: "*sad*" e "*can't*". Para t_3 há apenas uma palavra de sentimento positivo, "*ready*", e não há palavras de sentimento negativo. Para t_4, finalmente, há uma palavra de sentimento positivo ("*decent*") e uma palavra de sentimento negativo ("*not*"). Estes termos estão destacados na Tabela 4.1. Conforme ilustrado em **(a)** da Figura 4.6, o *tweet* t_1 pertence à classe *positiva*, enquanto que t_2, t_3 e t_4 pertencem à classe *neutra*.

ID	Texto
t_1	@$LUFC\_SOCCER$ Gators soccer ***wins*** 12th SEC title: The three points ***earned*** put Florida on top with 33 ...
t_2	I hope my ***better*** half is feeling ***better*** today! I will be so ***sad*** if we ***can't*** climb Mount Snowdon tomorrow! Was so
t_3	Getting ***ready*** for Knollwood P.S. bazaar, tonight and tomorrow!
t_4	@SeanJohnGerard @gmurph25 @tompcotter It's ***decent*** but ***not*** +3 standard yet, teeing it up at MJ on Monday. Ho

Tabela 4.1: Exemplo de *tweets* da base de dados *Twitter2014* (Rosenthal *et al.*, 2014).

Modelos de classificação foram induzidos a partir de um conjunto de dados rotulados que representou 5% da base de dados *SemEval2013* (Nakov *et al.*, 2013). Em **(b)** e **(c)** da Figura 4.6 nota-se que o classificador SVM (operando individualmente) não foi capaz de predizer corretamente os rótulos de classes dos quatro *tweets* em destaque. Em **(c)**, em particular, avaliou-se o SVM como componente supervisionado em um procedimento de *self-training*. Por outro lado, ao se utilizar a abordagem semissupervisionada do C^{3}E-SL (esquematizada na Figura 4.5) todos os *tweets* foram corretamente classificados — observe **(a)** e **(d)** na Figura 4.6. Neste caso,

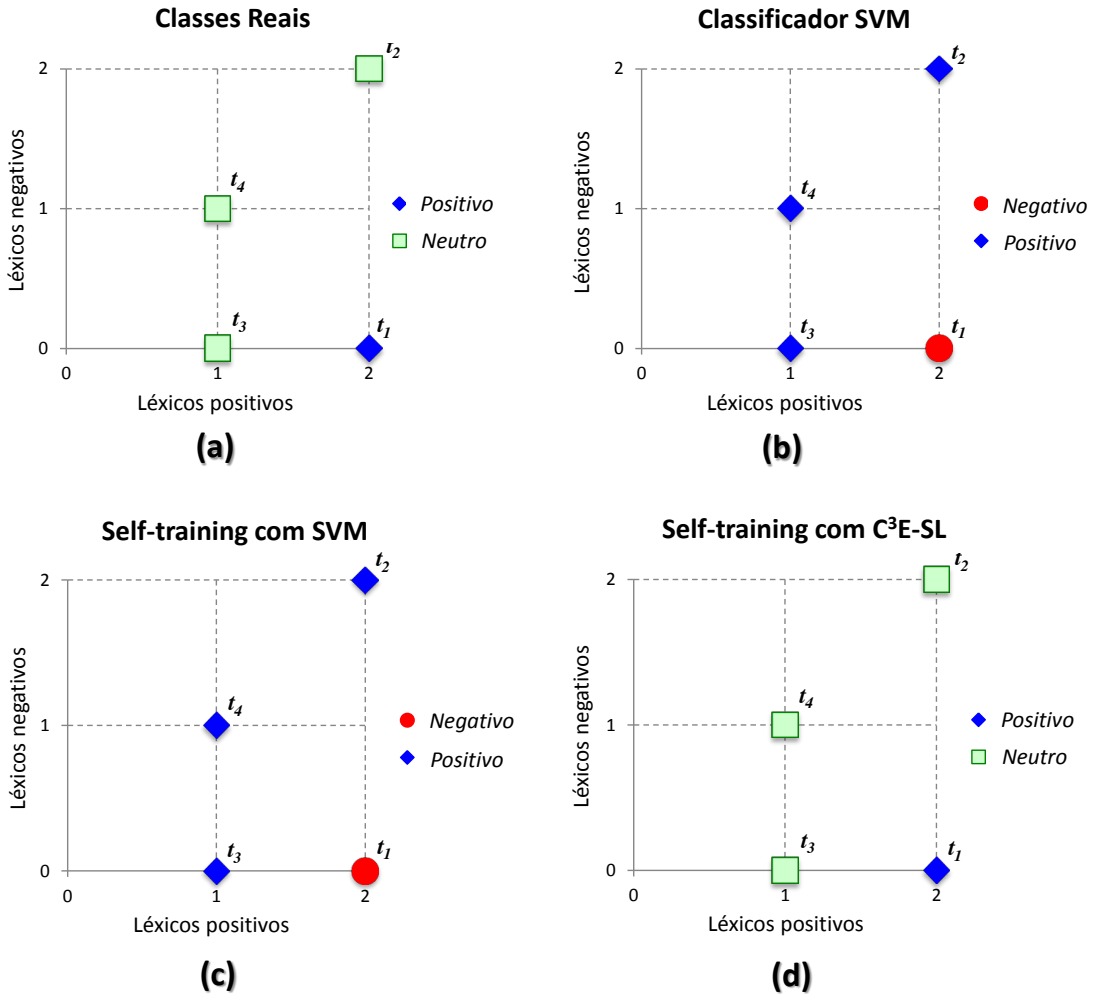


Figura 4.6: Espaço bidimensional formado pela contagem de palavras de sentimento positivo e de palavras de sentimento negativo presentes nos *tweets* da Tabela 4.1. Em **(a)** são apresentadas as classes reais destes *tweets*. Em **(b)** e **(c)** são apresentadas as classificações com o uso do SVM (independente e como componente na abordagem de *self-training*). Em **(d)**, por fim, é apresentada a predição feita pela abordagem do C^{3}E-SL semissupervisionado.

a informação não supervisionada codificada pela matriz de similaridades ajudou na predição correta destes *tweets*, pois o C^{3}E-SL conseguiu refinar (e corrigir) classificações prévias feitas pelo SVM ao explorar estruturas intrínsecas dos dados em um espaço de atributos formado por palavras de sentimento positivo e de sentimento negativo. A Tabela 4.2 apresenta parte da matriz de similaridades usada pelo C^{3}E-SL. Nesta matriz, o *tweet* t_1 tem similaridade máxima com relação aos *tweets* t_{20} e t_{1853} — pois possuem a mesma contagem de léxicos positivos e negativos. Os *tweets* t_{20} e t_{1853} pertencem à classe *positiva* e, assim, somaram evidências para que, durante o processo de otimização do C^{3}E-SL, t_1 pudesse ser classificado como sendo da classe *positiva* também. Portanto, este é um exemplo que mostra que o refinamento produzido pelo C^{3}E-SL ajuda a melhorar classificações e, em especial, aumenta as chances de realimentar o conjunto de dados rotulados com objetos (*tweets*) corretos a partir do *self-training*.

	t_1	t_2	t_3	t_4	$\cdots$	t_{20}	$\cdots$	t_{50}	$\cdots$	t_{125}	$\cdots$	t_{1853}
t_1	1,0	0,5	0,5	0,0	$\cdots$	**1,0**	$\cdots$	0,3	$\cdots$	0,0	$\cdots$	**1,0**
t_2	0,5	1,0	0,2	0,3	$\cdots$	0,5	$\cdots$	1,0	$\cdots$	0,2	$\cdots$	0,0
t_3	0,5	0,2	1,0	0,5	$\cdots$	0,5	$\cdots$	0,5	$\cdots$	0,8	$\cdots$	0,4
t_4	0,0	0,3	0,5	1,0	$\cdots$	0,2	$\cdots$	0,6	$\cdots$	0,7	$\cdots$	0,2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
t_{20}	**1,0**	0,5	0,5	0,2	$\cdots$	1,0	$\cdots$	0,2	$\cdots$	0,3	$\cdots$	**1,0**
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
t_{50}	0,3	1,0	0,5	0,3	$\cdots$	0,2	$\cdots$	1,0	$\cdots$	0,3	$\cdots$	0,2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
t_{125}	0,0	0,2	0,8	0,7	$\cdots$	0,3	$\cdots$	0,3	$\cdots$	1,0	$\cdots$	0,2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
t_{1853}	**1,0**	0,0	0,3	0,1	$\cdots$	**1,0**	$\cdots$	0,2	$\cdots$	0,2	$\cdots$	1,0

Tabela 4.2: Valores de similaridades entre *tweets* contidos na matriz de similaridades usada pelo C^{3}E-SL. Em destaque estão os valores máximos calculados para os *tweets* t_1, t_{20} e t_{1853}.

4.2 Avaliação Empírica na Classificação de Tweets

Na análise de sentimentos em *tweets*, a obtenção de dados rotulados depende de um processo custoso de anotação manual que muitas vezes necessita de conhecimento de domínio (e de especialistas). Isso limita bastante a quantidade de *tweets* rotulados oferecidos para classificadores. Os experimentos realizados se basearam nestes cenários para explorar as capacidades do C^{3}E-SL semissupervisionado (Seção 4.1.2). Análises empíricas foram conduzidas de maneira comparativa a partir de bases de dados representativas da área.

A Tabela 4.3 apresenta as características das bases de dados usadas nos experimentos. Estas bases de dados são recomendadas pelo *International Workshop on Semantic Evaluation* (SemEval)[2], cujos organizadores sugerem particularmente o uso da base de dados *SemEval2013* como fonte de dados rotulados para o treinamento dos modelos (Nakov *et al.*, 2013; Rosenthal *et al.*, 2014). Os modelos induzidos foram, então, avaliados em cinco conjuntos de teste (alvo), os quais, na realidade, são bases de dados amplamente conhecidas da área: *LiveJournal*, *SMS2013*, *Twitter2013*, *Twitter2014* e *TwitterSarcasm2014*. As bases de dados *LiveJournal* e *SMS2013* foram usadas com o objetivo de investigar como classificadores treinados a partir de mensagens de *Twitter* se comportam em outras fontes de dados (i.e., de *weblogs* e mensagens de celular). Já as bases de dados *Twitter2013*, *Twitter2014* e *TwitterSarcasm2014* foram obtidas pela identificação de entidades nomeadas extraídas de milhões de *tweets* coletados em um período entre Janeiro de 2012 e Janeiro de 2013 através da API pública do *Twitter*. A partir do reconhecimento de temas populares relacionados às entidades nomeadas (os quais são frequentemente mencionados em associação com uma data específica), os *tweets* foram manualmente rotulados. A base de dados *Twitter2013* possui três temas em comum com os temas presentes no conjunto usado para treinamento (i.e., *SemEval2013*). As bases de dados *Twitter2014* e *TwitterSarcasm2014* foram formadas mais recentemente e não contêm necessariamente temas presentes em *SemEval2013* — a última base de dados, em particular, foi coletada através da hashtag $\#sarcasm$

[2] http://en.wikipedia.org/wiki/SemEval

com o intuito de determinar como o sarcasmo e a ironia afetam a polaridade dos *tweets*. Todas as bases de dados foram rotuladas por meio de anotadores cadastrados no *Amazon Mechanical Turk*[3].

Base de Dados	Classe Positiva	Classe Negativa	Classe Neutra	Total
Dados rotulados para a aprendizagem dos modelos				
SemEval2013	4.215 (37%)	1.807 (15%)	5.325 (48%)	11.338
Novos dados para classificação (conjuntos de teste)				
LiveJournal	427 (37%)	304 (27%)	411 (36%)	1.142
SMS2013	492 (23%)	394 (19%)	1.207 (58%)	2.093
Twitter2013	1.572 (41%)	601 (16%)	1.640 (43%)	3.813
Twitter2014	982 (53%)	202 (11%)	669 (36%)	1.853
TwitterSarcasm2014	33 (38%)	40 (47%)	13 (15%)	86

Tabela 4.3: Distribuição das classes *positiva*, *negativa* e *neutra* nas seis bases de dados utilizadas. Os classificadores foram treinados com *tweets* rotulados da base de dados *SemEval2013* Nakov *et al.* (2013). Os resultados reportados foram obtidos a partir de classificações em cinco conjuntos de teste — *LiveJournal*, *SMS2013*, *Twitter2013*, *Twitter2014* e *TwitterSarcasm2014*.

A forma de representação de *tweets* não é consistentemente semelhante nos diferentes trabalhos encontrados na literatura, como já discutido no Capítulo 2. Por exemplo, muitas abordagens têm usado apenas atributos gerados a partir de N-gramas (Baugh, 2013), enquanto que outras incluem também informações de *emoticons* (Liu *et al.*, 2013b,a). Alguns trabalhos mais recentes também consideram um espaço de atributos mais complexo derivado de *part-of-speech tags*, léxicos e *hashtags* (Becker *et al.*, 2013; Xiang & Zhou, 2014; Zhao *et al.*, 2014). O conjunto de atributos adotado nesta pesquisa foi inspirado no trabalho de Mohammad *et al.* (2013), que conquistou o primeiro lugar na competição promovida pelo evento SemEval 2013 (Nakov *et al.*, 2013). Esta representação também alcançou os melhores resultados durante o SemEval 2014 — nas bases de dados *LiveJournal*, *SMS2013* e *TwitterSarcasm2014* (Rosenthal *et al.*, 2014). O conjunto de atributos adotado é formado por:

(i) *N-gramas:* um N-grama é uma sequência contínua de N itens (palavras) presentes em uma fonte textual ou falada. Foram considerados unigramas, bigramas e trigramas, os quais são subsequências formadas, respectivamente, por uma, duas e três palavras. Este modelo de linguagem é baseado em *bag-of-words*, no qual um atributo representa a frequência com que um N-grama aparece na mensagem considerada (Pang *et al.*, 2002). Assim como no Capítulo 2, os *tweets* são representados por meio de uma tabela na qual as colunas representam os n-gramas existentes nas mensagens, e os valores associados às colunas são referentes à presença destes no *tweet*.

(ii) *Léxicos:* geram atributos que representam a contagem de certas palavras consideradas positivas ou negativas de acordo com o método descrito em (Mohammad *et al.*, 2013);

(iii) *Negação:* indica o número de contextos negados. Segundo Pang *et al.* (2002), um contexto negado é um segmento de mensagem que se inicia com uma palavra de negação (e.g., *"no"*, *"shouldn't"*) e termina com uma pontuação (vírgula, ponto, dois pontos, ponto

[3] https://www.mturk.com

e vírgula, ponto de exclamação ou ponto de interrogação). Um contexto negado afeta os atributos formados por N-gramas e léxicos. Assim, adiciona-se o sufixo *"NEG"* em cada palavra (e.g., *"good"* torna-se *"good_NEG"*). Uma lista de palavras de negação foi adotada a partir do tutorial produzido por Christopher Potts[4];

(iv) *Part-of-speech:* usando-se o aplicativo Ark-twitter NLP (Owoputi *et al.*, 2013b) extraiu-se *tags* de *part-of-speech* (e.g., *emoticons*, *hashtags*, adjetivos, advérbios, dentre outros). O número de ocorrências de cada *tag* é representado na forma de atributo;

(v) *Estilo de Escrita:* atributos foram derivados a partir da presença de três ou mais caracteres repetidos em palavras, da sequência de três ou mais pontuações e do número de palavras com todas as letras em maiúsculo;

(vi) *Atributos do Microblogging:* considerou-se atributos que contam o número total de *hashtags* de sentimentos no texto, de ítens léxicos e *emoticons* (Thelwall *et al.*, 2010; Mohammad *et al.*, 2013).

Para a extensão semissupervisionada do C^{3}E-SL, o componente supervisionado (SVM) foi treinado no espaço de atributos formado pelos itens ***(i)–(vi)***, enquanto que os atributos responsáveis pela criação da matriz de similaridades correspondem aos léxicos e contextos negados — ou seja, os itens ***(ii)*** e, ***(iii)***. Tal configuração visa minimizar os efeitos negativos da *maldição da dimensionalidade* (ou *curse of dimensionality*)[5] e tornar o componente não supervisionado mais simples e interpretável.

A avaliação experimental levou em conta o cenário de aprendizado semissupervisionado indutivo. Uma proporção p de *tweets* (rotulados) foi aleatoriamente amostrada a partir da base de dados *SemEval2013* (Nakov *et al.*, 2013) (veja a Tabela 4.3) — manteve-se o balanço entre as classes *positiva*, *negativa* e *neutra*. O conjunto de *tweets* dado por p foi usado como instâncias de treinamento para construção de um modelo inicial. Os *tweets* contidos na importância sobressalente $(1 - p)$ foram usados na fase de aprendizagem das abordagens. Seguindo a filosofia do *self-training*, a cada iteração, 100 *tweets* foram incorporados no conjunto p de *tweets* rotulados (mantendo-se a proporção das classes). Este procedimento foi repetido até todos os $(1 - p)$ *tweets* da base de dados *SemEval2013* serem rotulados e promovidos ao conjunto p. Os modelos resultantes foram usados para classificar *tweets* nas bases de dados de teste (*LiveJournal*, *SMS2013*, *Twitter2013*, *Twitter2014* e *TwitterSarcasm2014*). Os resultados para cada uma destas bases de dados foram anotados e são reportados a seguir. Adotou-se como medida de desempenho o *F-score* para as classes *positivas* e *negativas*, bem como o *F-score* geral dado por $\overline{F} = (F_{Pos} + F_{Neg})/2$ (Becker *et al.*, 2013; Baugh, 2013; Zhao *et al.*, 2014; Xiang & Zhou, 2014).

[4] http://sentiment.christopherpotts.net/lingstruc.html

[5] Termo que se refere a vários fenômenos que surgem na análise de dados em espaços com muitas dimensões (atributos) (Hastie *et al.*, 2009).

4.2.1 Análise Comparativa

O C^{3}E-SL semissupervisionado foi comparado com uma abordagem similar, a qual usa o classificador SVM como componente do *self-training* e com o *co-training* (Zhou & Li, 2005), que pode ser vista como uma versão multi-descrição do algoritmo *self-training*. O conjunto inicial de *tweets* rotulados foi composto por 5% da base de dados *SemEval2013* (realizou-se amostragem aleatória estratificada). As médias e desvios padrões referentes aos resultados nos conjuntos de teste estão sumarizados nas Tabelas 4.4–4.8 (para obter estas estatísticas de interesse, repetiu-se o experimento 10 vezes para cada base de dados). Como referência, são reportados também os resultados para o SVM atuando de maneira independente (sem *self-training*). Os resultados deste SVM independente, denominado de SVM supervisionado, correspondem às médias de resultados de classificação computados a partir dos conjuntos de treinamento iniciais — com 5% da base de dados *SemEval2013*. Assim, enquanto o SVM semissupervisionado (baseado em *self-training*) possui interdependência das predições em diferentes iterações, cada rodada do classificador SVM supervisionado independe de suas outras execuções (por selecionar aleatoriamente conjuntos de treinamento em cada uma delas). Uma abordagem não supervisionada considerando léxicos também foi considerada como baseline para a análise, sendo que tal abordagem foi baseada na contagem das palavras de sentimento nos *tweets* de acondo com léxicos existentes. Para a abordagem *self-training*, foram usados todos os atributos descritos anteriormente, enquanto que para o *co-training*, foram consideradas duas descrições (visões): uma fornecida pelos Ngramas ***(i)*** e outra contendo o restante dos atributos ***(ii)-(vi)***. Esta configuração favorece a observação das vantagens de uso do *self-training* e do *co-training* no contexto das máquinas de vetores de suporte.

As Tabelas 4.4–4.8 mostram que todas as abordagens semissupervisionadas oferecem resultados tipicamente melhores que o SVM operando de maneira supervisionada. A versão semissupervisionada do C^{3}E-SL, em particular e comparativamente à abordagem *self-training*, alcançou, na maioria das vezes, os melhores resultados. Em comparação à abordagem *co-training*, em 60% dos casos a versão semissupervisionada do C^{3}E-SL obteve os melhores resultados. Em contrapartida, algumas vezes o desempenho das abordagens iterativas (semissupervisionadas) foi inferior ao alcançado pelo SVM supervisionado (e.g., execuções $\#1$ e $\#6$ na Tabela 4.8). Pressupõe-se que, nestes casos, os dados rotulados representaram razoavelmente bem as distribuições das classes, enquanto que dados não rotulados podem naturalmente favorecer superajuste (*overfitting*). Neste contexto particular, dados não rotulados nem sempre são úteis para se alavancar a classificação — entretanto, como visto, na média geral vantagens significativas foram observadas ao se usar a informação não supervisionada por meio do algoritmo C^{3}E-SL.

Execuções	SVM Supervisionado (Independente)			*Self-training*			*Co-training*			C^{3}E-SL Semissupervisionado		
	F_{Pos}	F_{Neg}	$\overline{F}$	F_{Pos}	F_{Neg}	$\overline{F}$	F_{Pos}	F_{Neg}	$\overline{F}$	F_{Pos}	F_{Neg}	$\overline{F}$
#1	61,10	52,31	56,71	68,40	53,20	60,80	69,24	43,92	56,58	**70,07**	**54,81**	**62,44**
#2	67,19	50,31	58,75	67,71	50,75	59,23	**69,44**	53,98	61,71	68,59	**54,90**	**61,75**
#3	50,30	46,05	48,18	62,97	50,66	56,82	**69,46**	**54,33**	**61,90**	63,47	46,88	55,18
#4	56,66	40,47	48,56	67,42	46,54	56,98	**69,24**	43,55	56,40	68,05	**47,59**	**57,82**
#5	66,85	60,91	63,88	67,51	57,89	62,70	**69,53**	53,98	61,75	68,38	**62,36**	**65,37**
#6	62,19	41,47	51,83	66,12	43,27	54,69	**68,72**	**51,31**	**60,01**	67,24	46,02	56,63
#7	67,56	50,42	58,99	66,46	54,41	60,43	**69,62**	53,98	61,80	68,23	**55,60**	**61,92**
#8	68,38	35,53	51,96	66,03	47,64	56,83	**68,53**	44,08	56,31	67,72	**50,53**	**59,13**
#9	62,64	38,41	50,53	63,76	37,96	50,86	**68,84**	42,99	**55,91**	66,88	**44,94**	**55,91**
#10	63,56	37,41	50,48	63,76	**51,21**	**57,48**	**68,78**	44,80	56,79	66,32	44,34	55,33
$\overline{F}$	62,64	45,33	53,99	66,01	49,35	57,68	**69,14**	48,69	58,92	67,50	**50,80**	**59,15**
Std. deviation	05,65	08,08	05,26	01,90	5,75	**03,37**	**0,39**	**5,17**	**2,75**	01,74	05,90	03,54

Tabela 4.4: F-scores (%) obtidos pelas abordagens semissupervisionadas e o SVM supervisionado na base de dados LiveJournal. Melhores resultados estão em negrito.

Execuções	SVM Supervisionado (Independente)			*Self-training*			*Co-training*			C^{3}E-SL Semissupervisionado		
	F_{Pos}	F_{Neg}	$\overline{F}$	F_{Pos}	F_{Neg}	$\overline{F}$	F_{Pos}	F_{Neg}	$\overline{F}$	F_{Pos}	F_{Neg}	$\overline{F}$
#1	49,91	49,31	49,61	53,98	47,58	50,78	**60,18**	47,01	53,60	56,52	**51,72**	**54,12**
#2	54,92	47,09	51,01	**60,00**	40,68	50,34	56,18	**49,28**	52,73	59,94	49,16	**54,55**
#3	49,64	41,89	45,77	52,05	39,66	45,85	55,94	**49,34**	**52,64**	**56,26**	**43,43**	49,84
#4	51,98	39,56	45,77	53,93	42,87	48,40	**60,56**	**47,34**	**53,95**	57,96	46,11	52,04
#5	53,38	45,58	49,48	55,08	41,80	48,44	56,38	**49,58**	**52,98**	**56,78**	48,36	52,58
#6	48,67	39,59	44,13	49,31	38,56	43,94	**59,48**	**49,67**	**54,57**	54,05	42,38	48,21
#7	55,27	47,19	51,23	53,58	43,07	48,32	56,15	49,40	52,77	**58,66**	**51,16**	**54,90**
#8	54,53	37,04	45,78	53,47	41,33	47,40	**60,00**	46,75	**53,37**	58,26	**47,18**	52,72
#9	48,67	38,27	43,47	46,89	34,94	40,92	**60,26**	**46,81**	**53,50**	52,99	41,69	47,34
#10	47,92	39,12	43,52	49,01	44,66	46,83	**60,28**	**46,73**	**53,50**	49,61	46,64	48,12
$\overline{F}$	51,49	42,46	46,98	52,73	41,52	47,12	**58,54**	**48,19**	**53,37**	56,10	46,78	51,44
Std. deviation	02,86	04,42	03,06	03,69	03,45	02,96	**2,06**	**1,34**	**0,60**	03,09	03,47	02,84

Tabela 4.5: F-scores (%) obtidos pelas abordagens semissupervisionadas e o SVM supervisionado na base de dados SMS2013. Melhores resultados estão em negrito.

Execuções	SVM Supervisionado (Independente)			*Self-training*			*Co-training*			C^{3}E-SL Semissupervisionado		
	F_{Pos}	F_{Neg}	$\overline{F}$	F_{Pos}	F_{Neg}	$\overline{F}$	F_{Pos}	F_{Neg}	$\overline{F}$	F_{Pos}	F_{Neg}	$\overline{F}$
#1	60,50	45,66	53,08	64,98	45,83	55,41	**66,64**	38,28	52,46	63,99	**48,44**	**56,21**
#2	61,18	42,41	51,79	65,13	**45,66**	**55,39**	64,87	38,34	51,60	**65,30**	41,29	53,30
#3	57,55	**42,62**	50,08	62,21	41,91	52,06	64,38	38,44	51,41	**67,95**	41,30	**54,63**
#4	60,42	44,89	52,65	64,80	**47,43**	**56,11**	**66,79**	38,58	52,68	65,66	44,32	54,99
#5	61,43	**44,97**	53,20	64,74	42,73	**53,73**	64,89	38,49	51,69	**65,32**	39,02	52,17
#6	62,30	43,91	53,11	64,53	43,81	54,17	66,22	37,64	51,93	**67,49**	**44,91**	**56,20**
#7	60,80	43,13	51,97	63,73	44,70	54,21	64,71	38,50	51,61	**68,62**	**45,64**	**57,13**
#8	62,59	35,39	48,99	64,74	**45,61**	**55,17**	**66,72**	38,37	52,55	64,32	45,42	54,86
#9	60,00	41,91	50,96	62,16	41,27	**51,72**	66,66	37,81	52,24	**68,62**	**42,27**	**55,44**
#10	63,85	42,42	53,14	64,29	**46,00**	55,14	66,72	38,46	52,59	**68,98**	43,58	**56,28**
$\overline{F}$	61,06	42,73	51,90	64,13	**44,49**	54,31	65,86	38,29	52,08	**66,62**	43,63	**55,12**
Std. deviation	01,70	02,88	01,47	01,10	02,00	01,46	**1,01**	**0,31**	**0,48**	01,90	02,71	01,49

Tabela 4.6: F-scores (%) obtidos pelas abordagens semissupervisionadas e o SVM supervisionado na base de dados Twitter2013. Melhores resultados estão em negrito.

Execuções	SVM Supervisionado (Independente)			*Self-training*			*Co-training*			C^{3}E-SL Semissupervisionado		
	F_{Pos}	F_{Neg}	$\overline{F}$	F_{Pos}	F_{Neg}	$\overline{F}$	F_{Pos}	F_{Neg}	$\overline{F}$	F_{Pos}	F_{Neg}	$\overline{F}$
#1	57,33	43,72	50,52	64,63	**47,03**	**55,83**	**66,55**	42,68	54,62	63,69	45,00	54,35
#2	60,84	42,13	51,49	**64,92**	44,33	54,62	64,01	33,49	48,75	64,76	**44,80**	**54,79**
#3	57,52	**44,92**	51,22	63,21	41,87	52,54	**63,88**	33,65	48,76	63,77	44,02	**53,89**
#4	58,73	44,32	51,53	62,43	**46,31**	54,37	**66,55**	42,33	**54,44**	64,40	42,66	53,54
#5	61,58	**45,96**	53,77	**64,40**	45,23	**54,81**	64,15	33,17	48,66	64,12	42,32	53,22
#6	60,64	40,87	50,76	64,12	43,26	**53,69**	**66,63**	35,16	50,89	65,30	**45,01**	55,15
#7	60,42	43,14	51,78	63,30	43,41	53,35	63,85	33,73	48,79	**64,29**	**43,70**	**53,99**
#8	61,59	40,12	50,85	64,25	45,15	54,70	**66,63**	42,07	54,35	65,39	**47,96**	**56,68**
#9	58,88	40,58	49,73	61,66	40,72	51,19	**66,48**	42,68	**54,84**	62,80	**43,74**	53,27
#10	62,73	45,58	54,16	63,03	45,23	54,13	**66,52**	42,81	54,66	64,23	**47,15**	**55,69**
$\overline{F}$	60,02	43,14	51,58	63,60	44,25	53,92	**65,52**	38,18	51,85	64,28	**44,64**	**54,46**
Std. deviation	01,82	02,12	01,39	01,05	01,96	01,31	1,33	4,60	2,89	**0,77**	**01,79**	**01,12**

Tabela 4.7: F-scores (%) obtidos pelas abordagens semissupervisionadas e o SVM supervisionado na base de dados Twitter2014. Melhores resultados estão em negrito.

Execuções	SVM Supervisionado (Independente)			*Self-training*			*Co-training*			C^{3}E-SL Semissupervisionado		
	F_{Pos}	F_{Neg}	$\overline{F}$	F_{Pos}	F_{Neg}	$\overline{F}$	F_{Pos}	F_{Neg}	$\overline{F}$	F_{Pos}	F_{Neg}	$\overline{F}$
#1	50,00	**25,53**	**37,77**	50,63	13,64	32,13	**55,00**	13,95	34,47	50,75	21,74	36,24
#2	49,32	13,64	31,48	52,05	24,49	38,27	**55,26**	**45,28**	**50,27**	54,55	25,53	40,03
#3	46,15	17,39	31,77	50,63	13,64	32,13	**56,75**	**44,44**	**50,60**	50,00	21,28	35,64
#4	44,44	09,30	26,87	48,72	09,30	29,01	**55,69**	13,94	**34,81**	51,85	**17,78**	**34,81**
#5	55,70	33,96	44,83	52,63	24,49	38,56	56,00	**44,44**	**50,22**	**56,47**	34,62	45,54
#6	49,28	25,53	37,40	48,10	17,39	32,75	**55,69**	**36,73**	**46,21**	51,28	09,30	30,29
#7	**58,54**	13,64	36,09	55,42	25,53	40,48	55,26	**45,28**	**50,27**	56,41	21,74	39,07
#8	54,76	09,30	32,03	52,50	**25,00**	**38,75**	55,69	13,95	34,82	**62,37**	09,52	35,94
#9	49,35	17,78	33,56	51,85	**21,28**	**36,56**	**56,41**	13,95	35,18	53,16	17,78	35,47
#10	43,84	09,30	26,57	52,38	**17,78**	**35,08**	**56,41**	13,95	**35,18**	49,35	**17,78**	33,56
$\overline{F}$	50,14	17,54	33,84	51,49	19,25	35,37	**55,81**	**28,59**	**42,20**	53,62	19,70	36,66
Std. deviation	04,86	08,38	05,46	02,10	05,77	03,74	**0,57**	15,62	7,80	03,96	**0,73**	**0,41**

Tabela 4.8: F-scores (%) obtidos pelas abordagens semissupervisionadas e o SVM supervisionado na base de dados Twitter Sarcasm 2014. Melhores resultados estão em negrito.

4.2.2 Abordagem Não Supervisionada Baseada em Léxicos

Uma abordagem não supervisionada baseada em léxicos também foi considerada na análise. Foram utilizados somente os léxicos propostos por Mohammad *et al.* (2013) e Hu & Liu (2004). O sentimento foi computado baseado na contagem de palavras de opinião positiva e negativa, i.e., se o número de palavras positivas é maior que o número de palavras negativas, o *tweet* é considerado positivo, caso contrário o *tweet* é considerado negativo. Se não há palavras de opinião ou o número de palavras positivas e negativas é o mesmo, o *tweet* é considerado neutro. Os experimentos reportados na seção anterior, considerando o cenário que adotou 5% dos dados rotulados da base *SemEval2013* foi mantido. As mesmas abordagens comparadas na seção anterior foram usadas — ou seja, levou-se em conta os resultados do C^{3}E-SL semissupervisionado, *self-training* com SVM, *co-training* com SVM e também um classificador SVM supervisionado operando de maneira independente (sem *self-training*). Os resultados obtidos com todos os algoritmos estão reportados nas Tabelas 4.4–4.8, exceto os resultados com a abordagem não supervisionada baseada em léxicos os quais são reportados na Tabela 4.9.

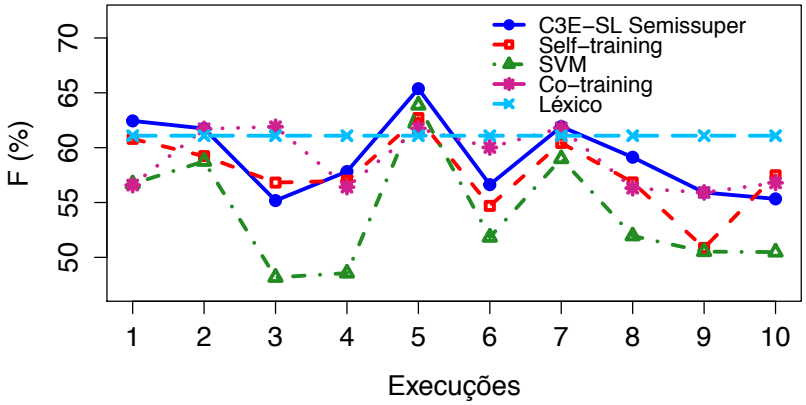


Figura 4.7: $\overline{F}$ (%) obtidos pelas abordagens semissupervisionadas, pelo SVM supervisionado e pela abordagem não supervisionada baseada em léxicos — LiveJournal.

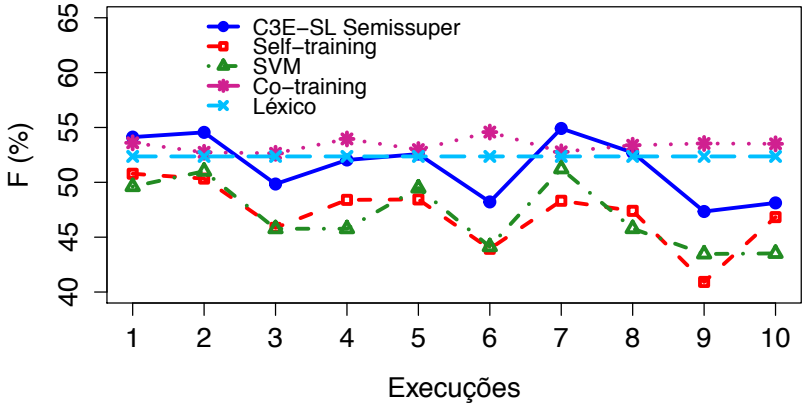


Figura 4.8: $\overline{F}$ (%) obtidos pelas abordagens semissupervisionadas, pelo SVM supervisionado e pela abordagem não supervisionada baseada em léxicos — SMS.

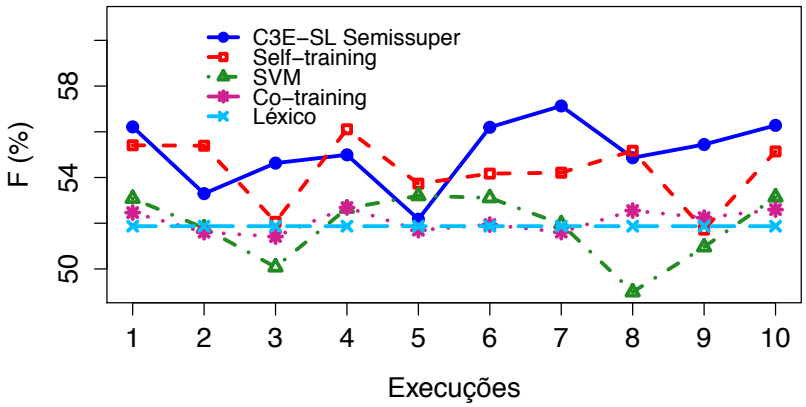


Figura 4.9: $\overline{F}$ (%) obtidos pelas abordagens semissupervisionadas, pelo SVM supervisionado e pela abordagem não supervisionada baseada em léxicos — Twitter2013.

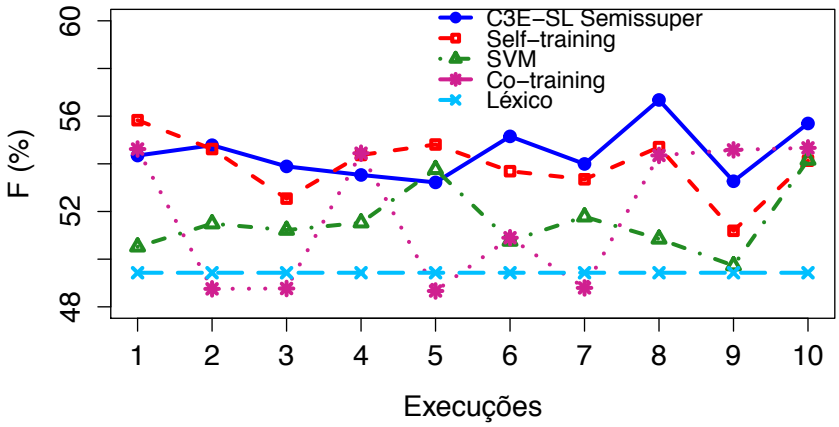


Figura 4.10: $\overline{F}$ (%) obtidos pelas abordagens semissupervisionadas, pelo SVM supervisionado e pela abordagem não supervisionada baseada em léxicos —Twitter2014.

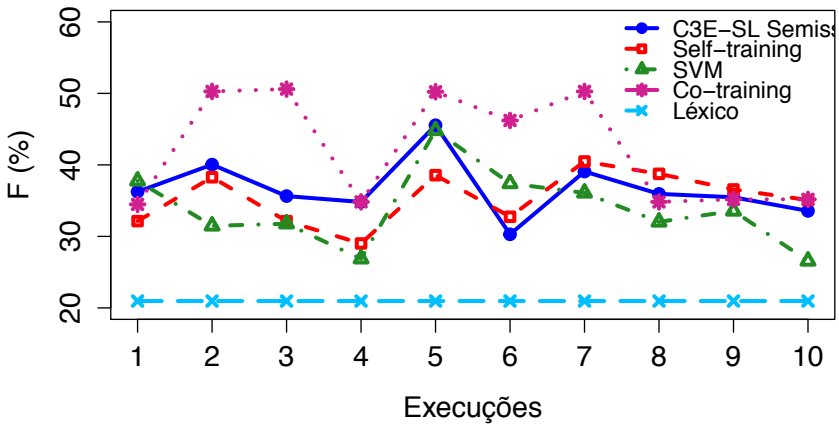
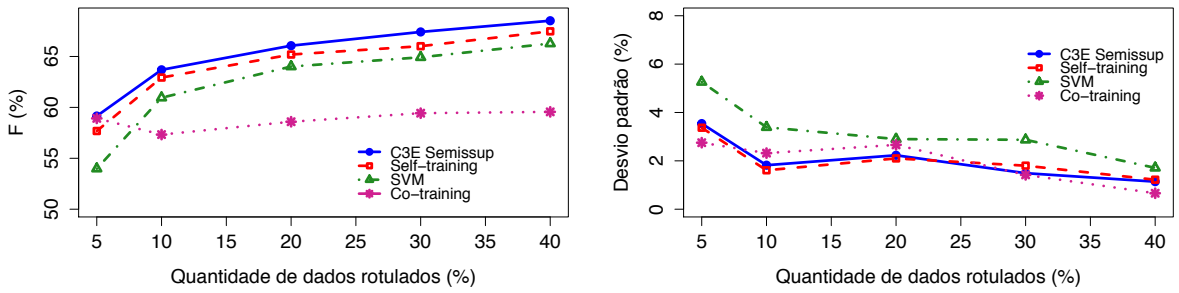


Figura 4.11: $\overline{F}$ (%) obtidos pelas abordagens semissupervisionadas, pelo SVM supervisionado e pela abordagem não supervisionada baseada em léxicos — Twitter Sarcasm 2014.

4.2.3 Impacto da Quantidade de Dados Rotulados Disponíveis

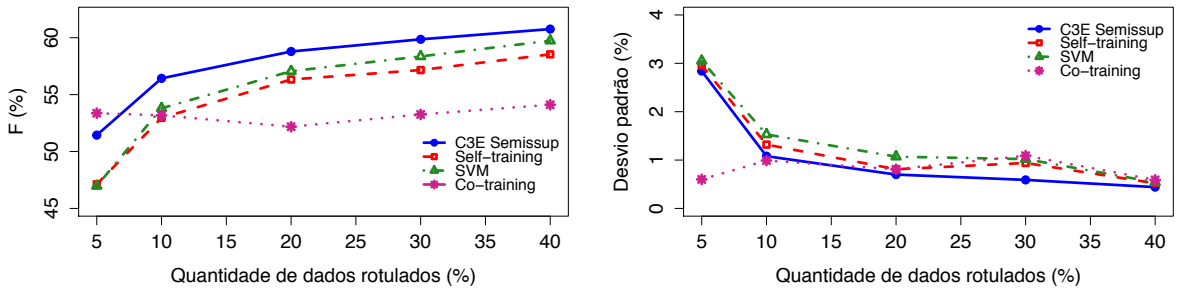
Os experimentos das seções 4.2.1 e 4.2.2 foram conduzidos em um cenário que adotou 5% dos dados rotulados da base *SemEval2013*. Esta seção investiga os impactos da variação deste percentual, i.e., do tamanho do conjunto de dados rotulados que se dispõe inicialmente para gerar um modelo de classificação. Conforme já destacado, uma das principais limitações práticas da tarefa de classificação é a escassez de dados rotulados. Portanto, espera-se que as abordagens semissupervisionadas se comportem bem em cenários de significativa escassez de dados rotulados. Considerando o compromisso entre quantidade de dados rotulados e desempenho de classificação, as discussões apresentadas podem ajudar na escolha da alternativa mais vantajosa para o problema que se tem em mãos. As mesmas abordagens comparadas na seção anterior foram usadas — ou seja, levou-se em conta os resultados do C^{3}E-SL semissupervisionado, *self-training* com SVM, *co-training* com SVM e um classificador SVM supervisionado operando de maneira independente (sem *self-training*).

As Figuras 4.12–4.16 apresentam os resultados gerais ($\overline{F}$ e desvio-padrão) para cada abordagem, considerando 10 execuções independentes a partir de uma certa quantidade (inicial) de dados rotulados. As quantidades usadas foram de 5%, 10%, 20%, 30% e 40% de dados rotulados amostrados da base *SemEval2013*. Como esperado, os resultados melhoraram com o aumento da quantidade de dados rotulados disponíveis. Comparativamente, os resultados do C^{3}E-SL semissupervisionado foram sensivelmente melhores em todas as bases de dados consideradas. Na base de dados *Twitter2013*, as abordagens semissupervisionadas alcançaram resultados modestos em relação ao classificador SVM. Entretanto, convém destacar que, em todas as bases de dados, para quantidades relativamente pequenas de dados rotulados, as abordagens semissupervisionadas foram as mais vantajosas. A Tabela 4.9 exibe as estatísticas obtidas, além de apresentar o desempenho de abordagens supervisionadas que utilizaram o conjunto de dados completo.



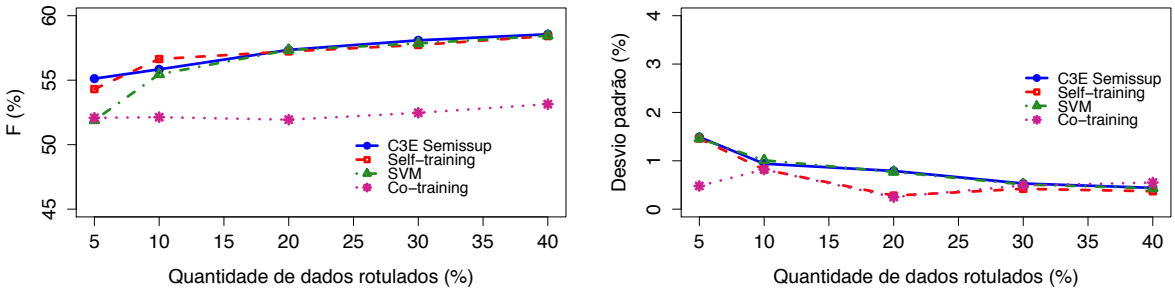
(a) $\overline{F}$ (%) (b) Desvios-padrão

Figura 4.12: Resultados para diferentes percentuais de dados rotulados — LiveJournal.



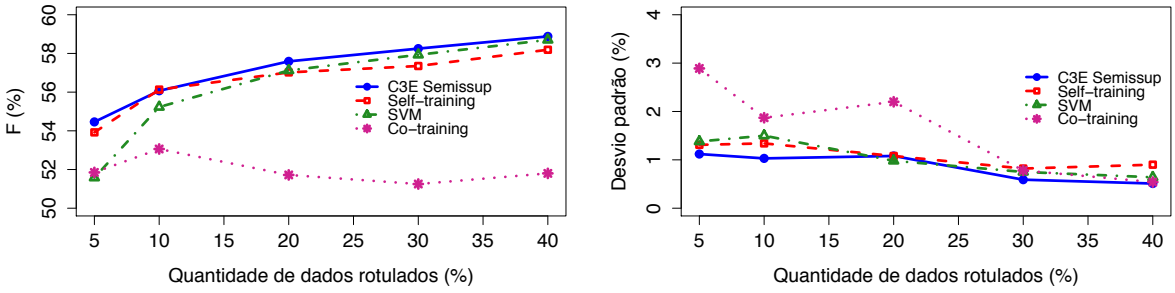
(a) $\overline{F}$ (%) (b) Desvios-padrão

Figura 4.13: Resultados para diferentes percentuais de dados rotulados — SMS2013.



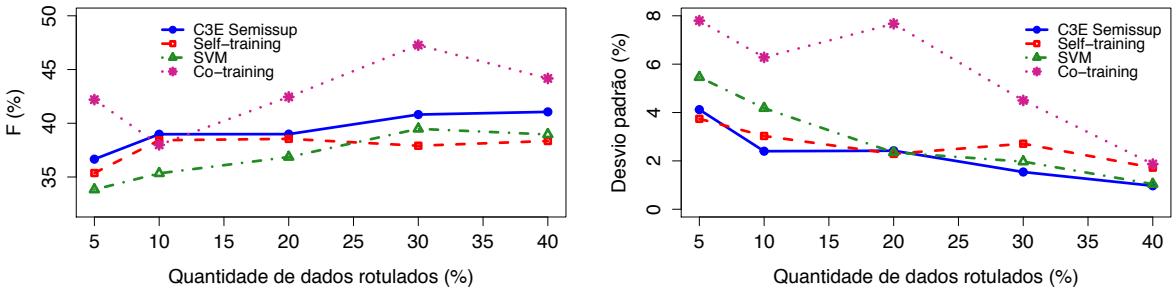
(a) $\overline{F}$ (%) (b) Desvios-padrão

Figura 4.14: Resultados para diferentes percentuais de dados rotulados — Twitter2013.



(a) $\overline{F}$ (%) (b) Desvios-padrão

Figura 4.15: Resultados para diferentes percentuais de dados rotulados — Twitter2014.



(a) $\overline{F}$ (%) (b) Desvios-padrão

Figura 4.16: Resultados para diferentes percentuais de dados rotulados — Twitter Sarcasm 2014.

LiveJournal																		
Approach	**5%**			**10%**			**20%**			**30%**			**40%**			**100%**		
	F-pos	F-neg	$\overline{F}$	F-pos	F-neg	$\overline{F}$	F-pos	F-neg	$\overline{F}$	F-pos	F-neg	$\overline{F}$	F-pos	F-neg	$\overline{F}$	F-pos	F-neg	$\overline{F}$
C^{3}E-SL	67,50	50,80	59,15	**69,10**	**58,27**	**63,69**	**69,87**	**62,25**	**66,06**	**70,21**	**64,61**	**67,41**	**70,81**	**66,21**	**68,51**	—	—	—
Semissup	±1,74	±5,90	±3,54	±0,86	±2,86	±1,82	±0,67	±4,30	±2,23	±0,66	±2,59	±1,49	±0,65	±2,06	±1,14			
Self-	66,01	49,35	57,68	68,09	57,77	62,93	69,13	61,24	65,19	69,54	62,49	66,01	70,22	64,71	67,47	—	—	—
training	±1,89	±5,75	±3,37	±0,79	±2,77	±1,61	±0,69	±4,08	±2,10	±0,66	±3,13	±1,80	±0,84	±2,18	±1,22			
Co-	**69,14**	48,69	58,92	68,51	46,14	57,33	67,79	49,38	58,59	67,08	51,78	59,43	66,48	52,64	59,56	—	—	—
Training	±0,39	±5,17	±2,75	±0,61	±4,17	±2,32	±0,91	±4,49	±2,66	±0,75	±2,59	±1,42	±0,73	±0,69	±0,66			
SVM	62,64	45,33	53,98	66,54	55,35	60,95	68,33	59,69	64,01	69,01	60,82	64,92	69,21	63,34	66,27	68,84	61,75	67,34
	±5,64	±8,08	±5,26	±3,21	±5,58	±3,39	±1,57	±5,64	±2,90	±0,87	±5,38	±2,87	±1,18	±3,05	±1,71			
Léxico	66,58	**55,59**	**61,09**	66,58	55,59	61,09	66,58	55,59	61,09	66,58	55,59	61,09	66,58	55,59	61,09	66,58	55,59	61,09
Literatura	Supervised approach proposed by Zhu *et al.* (2014)																	74,84

SMS2013																		
Approach	**5%**			**10%**			**20%**			**30%**			**40%**			**100%**		
	F-pos	F-neg	$\overline{F}$	F-pos	F-neg	$\overline{F}$	F-pos	F-neg	$\overline{F}$	F-pos	F-neg	$\overline{F}$	F-pos	F-neg	$\overline{F}$	F-pos	F-neg	$\overline{F}$
C^{3}E-SL	56,10	46,78	51,44	**59,97**	**52,89**	**56,43**	**62,06**	**55,51**	**58,79**	**63,34**	**56,38**	**59,86**	**63,87**	**57,67**	**60,76**	—	—	—
Semissup	±3,09	±3,47	±2,84	±1,10	±1,23	±1,08	±1,10	±1,00	±0,70	±0,81	±0,83	±0,59	±0,49	±0,91	±0,44			
Self-	52,73	41,52	47,12	56,72	49,20	52,96	60,50	52,16	56,33	60,60	53,73	57,16	61,67	55,41	58,54	—	—	—
training	±3,69	±3,44	±2,96	±2,22	±1,01	±1,32	±0,81	±0,99	±0,81	±0,87	±1,22	±0,94	±0,63	±0,78	±0,52			
Co-	**58,54**	48,19	**53,37**	59,32	47,05	53,18	57,38	46,99	52,18	57,69	48,84	53,26	58,38	49,84	54,11	—	—	—
Training	±2,06	±1,34	±0,60	±0,99	±2,13	±0,99	±1,03	±2,28	±0,81	±1,10	±1,89	±1,09	±0,56	±0,92	±0,59			
SVM	51,48	42,46	46,97	56,74	50,85	53,80	60,35	53,77	57,06	61,56	55,15	58,36	62,78	56,72	59,74	59,31	51,69	55,50
	±2,86	±4,42	±3,05	±2,12	±1,29	±1,53	±1,56	±1,06	±1,07	±1,15	±1,28	±1,02	±0,62	±1,04	±0,54			
Léxico	54,43	**50,29**	52,36	54,43	50,29	52,36	54,43	50,29	52,36	54,43	50,29	52,36	54,43	50,29	52,36	54,43	50,29	52,36
Literatura	Supervised approach proposed by Mohammad *et al.* (2013)																	68,46

Twitter2013																		
Approach	**5%**			**10%**			**20%**			**30%**			**40%**			**100%**		
	F-pos	F-neg	$\overline{F}$	F-pos	F-neg	$\overline{F}$	F-pos	F-neg	$\overline{F}$	F-pos	F-neg	$\overline{F}$	F-pos	F-neg	$\overline{F}$	F-pos	F-neg	$\overline{F}$
C^{3}E-SL	**66,62**	43,62	**55,12**	66,10	45,58	55,84	66,63	48,06	**57,35**	**67,38**	48,80	**58,09**	**67,86**	49,26	**58,56**	—	—	—
Semissup	±1,90	±2,71	±1,49	±0,65	±1,60	±0,94	±0,56	±1,64	±0,79	±0,41	±0,88	±0,53	±0,32	±0,76	±0,44			
Self-	64,13	**44,49**	54,31	66,38	**46,90**	**56,64**	**66,78**	47,67	57,23	67,13	48,33	57,73	67,68	49,15	58,42	—	—	—
training	±1,09	±1,99	±1,46	±0,39	±1,54	±0,82	±0,41	±0,50	±0,28	±0,47	±1,06	±0,42	±0,26	±0,76	±0,37			
Co-	65,86	38,29	52,08	**66,45**	37,80	52,13	66,14	37,74	51,94	66,56	38,37	52,47	66,85	39,43	53,14	—	—	—
Training	±1,01	±0,31	±0,48	±0,49	±1,63	±0,82	±0,71	±0,66	±0,25	±0,39	±0,89	±0,49	±0,25	±1,06	±0,55			
SVM	61,06	42,73	51,89	64,80	46,13	55,46	66,39	**48,29**	57,34	66,83	**48,88**	57,85	67,50	**49,40**	58,45	70,70	61,41	66,05
	±1,70	±2,87	±1,46	±0,79	±1,69	±1,01	±0,50	±1,37	±0,77	±0,65	±0,83	±0,51	±0,29	±0,83	±0,43			
Léxico	61,75	42,00	51,87	61,75	42,00	51,87	61,75	42,00	51,87	61,75	42,00	51,87	61,75	42,00	51,87	61,75	42,00	51,87
Literatura	Supervised approach proposed by Mohammad *et al.* (2013)																	69,02

Twitter2014																		
Approach	**5%**			**10%**			**20%**			**30%**			**40%**			**100%**		
	F-pos	F-neg	$\overline{F}$	F-pos	F-neg	$\overline{F}$	F-pos	F-neg	$\overline{F}$	F-pos	F-neg	$\overline{F}$	F-pos	F-neg	$\overline{F}$	F-pos	F-neg	$\overline{F}$
C^{3}E-SL	64,28	**44,64**	**54,46**	65,88	**46,26**	56,07	**66,73**	**48,45**	**57,59**	**67,80**	**48,69**	**58,25**	**68,18**	**49,57**	**58,88**	—	—	—
Semissup	±0,77	±1,80	±1,12	±0,91	±1,57	±1,03	±0,79	±2,12	±1,08	±0,42	±1,06	±0,59	±0,56	±0,71	±0,51			
Self-	63,59	44,25	53,92	66,01	46,26	**56,13**	66,71	47,33	57,02	67,31	47,40	57,35	68,06	48,33	58,19	—	—	—
training	±1,04	±1,95	±1,31	±0,79	±2,05	±1,34	±0,73	±1,78	±1,08	±0,52	±1,37	±0,82	±0,35	±1,54	±0,90			
Co-	**65,52**	38,18	51,85	**66,16**	39,97	53,06	65,63	37,82	51,72	66,23	36,27	51,25	66,72	36,88	51,80	—	—	—
Training	±1,33	±4,60	±2,89	±0,69	±3,23	±1,87	±0,91	±3,62	±2,20	±0,45	±1,49	±0,78	±0,49	±0,66	±0,54			
SVM	60,02	43,13	51,58	65,17	45,29	55,23	66,44	47,61	57,12	67,57	48,30	57,93	68,14	49,26	58,70	71,87	55,51	63,69
	±1,82	±2,12	±1,38	±1,25	±2,28	±1,50	±0,79	±1,97	±0,98	±0,75	±1,39	±0,75	±0,44	±1,08	±0,64			
Léxico	61,75	37,11	49,43	61,75	37,11	49,43	61,75	37,11	49,43	61,75	37,11	49,43	61,75	37,11	49,43	61,75	37,11	49,43
Literatura	Supervised approach proposed by Miura *et al.* (2014)																	70,96

Twitter Sarcasm 2014																		
Approach	**5%**			**10%**			**20%**			**30%**			**40%**			**100%**		
	F-pos	F-neg	$\overline{F}$	F-pos	F-neg	$\overline{F}$	F-pos	F-neg	$\overline{F}$	F-pos	F-neg	$\overline{F}$	F-pos	F-neg	$\overline{F}$	F-pos	F-neg	$\overline{F}$
C^{3}E-SL	53,62	19,71	36,66	54,85	23,13	**38,98**	54,51	23,46	38,99	55,75	25,86	40,81	**58,19**	23,93	41,06	—	—	—
Semissup	±3,96	±7,38	±4,12	±2,89	±4,65	±2,40	±1,71	±4,97	±2,42	±2,43	±2,73	±1,54	±1,73	±2,53	±0,97			
Self-	51,49	19,25	35,37	52,85	**23,99**	38,42	53,49	23,62	38,55	54,21	21,62	37,91	54,64	22,06	38,35	—	—	—
training	±2,09	±5,76	±3,74	±1,55	±5,24	±3,03	±1,79	±4,33	±2,29	±1,41	±4,68	±2,71	±1,02	±3,34	±1,72			
Co-	**55,81**	**28,59**	**42,20**	**56,73**	19,24	37,99	**57,37**	**27,54**	**42,45**	**57,18**	**37,35**	**47,27**	53,67	**34,66**	**44,17**	—	—	—
Training	±0,57	±15,62	±7,80	±0,49	±12,46	±6,28	±1,02	±15,34	±7,67	±1,24	±8,61	±4,50	±1,65	±2,86	±1,87			
SVM	50,13	17,53	33,83	50,40	20,28	35,34	52,2	21,12	36,86	54,66	24,31	39,48	53,95	23,96	38,96	57,15	25,00	41,08
	±4,86	±8,37	±5,46	±3,26	±6,46	±4,18	±2,63	±4,23	±2,35	±2,23	±3,05	±1,97	±1,45	±2,02	±1,04			
Léxico	34,86	7,58	20,96	34,86	7,58	20,96	34,86	7,58	20,96	34,86	7,58	20,96	34,86	7,58	20,96	34,86	7,58	20,96
Literatura	Supervised approach proposed by Zhu *et al.* (2014)																	58,16

Tabela 4.9: F-scores para diferentes percentuais de dados rotulados para abordagens semissupervisionadas, supervisionadas e baseada em léxicos – incluindo os melhores resultados da literatura ao se utilizar o conjunto completo de dados rotulados. Os melhores resultados estão em negrito.

A Tabela 4.10, que foca apenas nos resultados alcançados pelas abordagens semissupervisionadas, reporta a proporção de F-scores mais altos e menores desvios padrões encontrados na Tabela 4.9. Pode-se constatar, portanto, que o *self-training* baseado no C^{3}E-SL apresentou melhores resultados do que se utilizar apenas o SVM como componente do *self-training*.

Base de Dados	**C^{3}E-SL Semissuper.**		***Self-training***		***Co-training***		**SVM**		**Léxico***
	Maiores F-scores	*Menores Desvios-Padrões*	*Maiores F-scores*	*Menores Desvios-Padrões*	*Maiores F-scores*	*Menores Desvios-Padrões*	*Maiores F-scores*	*Menores Desvios-Padrões*	*Maiores F-scores*
LiveJournal	**80.00**	40.00	0.00	26.67	6.66	33.33	0.00	0.00	13.33
SMS2013	**80.00**	**40.00**	0.00	26.66	13.33	33.33	0.00	0.00	6.66
Twitter2013	**46.66**	6.66	26.66	40.00	6.66	**46.00**	20.00	6.66	0.00
Twitter2014	**80.00**	**60.00**	6.66	20.00	13.33	13.33	0.00	6.66	0.00
Twitter Sarcasm 2014	**80.00**	**33.33**	6.66	**33.33**	13.33	26.66	0.00	6.66	0.00

*Os desvios-padrão para a abordagem baseada em léxicos é igual a zero.

Tabela 4.10: Proporção de maiores F-scores e menores desvios padrões ao se comparar apenas os resultados da Tabela 4.9 para o C^{3}E-SL semissupervisionado e sua contraparte, a qual emprega *self-training* com SVM, *co-training*, SVM supervisionado e a abordagem não-supervisionada baseada em léxicos. Os melhores resultados estão em negrito.

4.3 Considerações Finais

Em da Silva *et al.* (2014a) combinou-se classificadores e agrupadores para produzir melhores modelos supervisionados para a classificação de *tweets*. Contudo, abordagens semissupervisionadas têm se mostrado alternativas mais promissoras neste contexto (Becker *et al.*, 2013; He & Zhou, 2011; Crammer *et al.*, 2009; Qiu *et al.*, 2009). A partir desta perspectiva, uma abordagem semissupervisionada do algoritmo C^{3}E-SL foi avaliada na classificação de sentimentos em *tweets*. Esta abordagem combina máquinas de vetores de suporte — *Support Vector Machines* (SVMs) (Cortes & Vapnik, 1995; Boser *et al.*, 1992) — com a informação de similaridades entre objetos para se gerar classificações mais refinadas. Baseando-se em um procedimento de *self-training* (Zhu & Goldberg, 2009), predições mais confiáveis aprimoram o modelo de forma iterativa. Como é requerida uma quantidade pequena de dados rotulados, o C^{3}E-SL semissupervisionado é particularmente útil para a classificação de *tweets*.

Os experimentos envolveram seis bases de dados amplamente conhecidas da área de análise de sentimentos em *tweets* (Mohammad *et al.*, 2013; Rosenthal *et al.*, 2014; Nakov *et al.*, 2013). O C^{3}E-SL semissupervisionado foi comparado com o algoritmo *self-training*, o qual usa o classificador SVM como componente supervisionado, bem como foi comparado também ao algoritmo *co-training* composto por duas visões distintas. Adicionalmente, máquinas de vetores de suporte operando de forma independente (sem *self-training*) foram também testadas e finalmente, uma abordagem não supervisionada baseada no uso de léxicos foi investigada. O C^{3}E-SL semissupervisionado obteve os melhores resultados na maioria dos cenários investigados, apresentando-se como uma alternativa promissora para a aplicação analisada. Estes resultados e discussões foram reportadas no artigo *"Using Unsupervised Information to Improve Semi-supervised Tweet Sentiment Classification"* da Silva *et al.* (2016b).

Em resumo, os artigos produzidos nesta pesquisa e que estudam aplicações do C^{3}E-SL na classificação de *tweets* são:

- SILVA, N. F. F.; COLETTA, L. F. S.; HRUSCHKA, E. R.; HRUSCHKA JR., E. R., *Combining Classification and Clustering for Tweet Sentiment Analysis*. In: Brazilian Conference on Intelligent Systems (BRACIS), 2014, São Carlos. Proceedings on Brazilian Conference on Intelligent Systems, p. 210-215, 2014 — da Silva *et al.* (2014a).

- SILVA, N. F. F.; COLETTA, L. F. S.; HRUSCHKA, E. R.; HRUSCHKA JR., E. R., *Using Unsupervised Information to Improve Semi-supervised Tweet Sentiment Classification*, *Information Sciences*, 2016 — da Silva *et al.* (2016b).

- SILVA, N. F. F.; COLETTA, L. F. S.; HRUSCHKA, E. R., *A Survey and Comparative Study of Tweet Sentiment Analysis via Semi-Supervised Learning*, *ACM Computing Surveys*, 2016 — da Silva *et al.* (2016a).

Como tópico emergente, o uso de aprendizado semissupervisionado para a análise de sentimento em *tweets* motiva estudos em diferentes linhas. Para a abordagem proposta, a qual emprega o algoritmo C^{3}E-SL, a escolha dos melhores atributos, tanto para o componente supervisionado quanto para o não supervisionado, requer maiores investigações. Por exemplo, na presença de sarcasmo e ironia, existem atributos que são reconhecidamente mais promissores (Carvalho *et al.*, 2009; Gonzalez-Ibanez *et al.*, 2011; Vanin *et al.*, 2013; de Freitas *et al.*, 2014). Neste sentido, um mecanismo mais dinâmico seria selecionar durante o aprendizado os atributos que são mais adequados, como um procedimento intrínseco ao processo iterativo desempenhado pelo C^{3}E-SL semissupervisionado. Além disso, métodos semissupervisionados ainda dependem significativamente de uma amostragem *ideal* dos dados. Em outras palavras, em cenários com alto custo de rotulamento, deseja-se selecionar alguns (poucos) objetos para rotular que sejam, de fato, significativos para o propósito de treinamento do modelo, a exemplo das contribuições de Hajmohammadi *et al.* (2015) e Smailović *et al.* (2014). Nestes trabalhos, os autores utilizam-se de aprendizado ativo — *active learning* (Settles, 2012) — para maximizar a representação dos dados em pequenas amostras de treinamento.

Capítulo

5

Conclusão

Nesta tese, diferentes abordagens para classificar sentimentos em textos curtos, especialmente para *tweets*, foram estudadas. O ponto de partida foi a aplicação de algoritmos de aprendizagem supervisionada e agregadores de classificadores (*ensembles* de classificadores). Posteriormente, o algoritmo C^{3}E (*Consensus between Classification and Clustering Ensembles*) desenvolvido por Acharya *et al.* (2014, 2011) e que combina classificadores e agrupadores foi aplicado. Este algoritmo recebe como entradas estimativas de distribuições de probabilidades de classes para objetos de um conjunto alvo, bem como uma matriz de similaridades entre esses objetos. O C^{3}E produz como saída uma classificação consolidada dos objetos do conjunto alvo por meio de refinamentos nas distribuições de probabilidades de classes. A idéia intuitiva é que a informação de grupos (codificada pela matriz de similaridades) possa prover restrições complementares que auxiliem na classificação de sentimentos. A partir dessa noção, foi desenvolvida uma versão semissupervisionada do C^{3}E, que visa superar algumas das dificuldades práticas da classificação de sentimentos em *tweets* — especialmente quando se dispõe de poucos *tweets* rotulados e/ou uma amostragem deficiente dos *tweets* de treinamento. Em particular, o algoritmo concebido fornece um classificador mais "consciente" das possíveis diferenças entre as distribuições dos *tweets* usados no treinamento e dos *tweets* do conjunto alvo.

5.1 Sumário das Principais Contribuições

Neste trabalho, foi apresentada a proposta de uso de agregadores de modelos de classificação supervisionada para análise de sentimentos em textos curtos e oriundos do Twitter. Não foram encontradas na literatura pesquisas anteriores a esta propondo o uso de *ensembles* de classificadores em conjunto com léxicos para análise de sentimentos em textos curtos e informais.

Uma abordagem para combinar classificadores e agrupadores com o objetivo de classificar

novos *tweets* foi apresentada. Tal abordagem se materializa por meio do algoritmo existente C^{3}E (Coletta *et al.*, 2015a; Acharya *et al.*, 2014; da Silva *et al.*, 2014a; Coletta *et al.*, 2013; Acharya *et al.*, 2011), o qual não havia sido aplicado ao cenário de pesquisa desta tese. Os resultados experimentais reportados mostram que o C^{3}E é promissor, comprovando a ideia intuitiva de que as restrições fornecidas por meio de partições de dados (agrupamento) são úteis para classificar novos *tweets* quanto ao seu sentimento.

Uma versão do algoritmo C^{3}E foi estendida para o cenário semissupervisionado e aplicada em análise de sentimentos em *tweets*. Os resultados experimentais reportados demonstram que o método é promissor. Trata-se de unir os benefícios de combinar classificadores e agrupadores em um cenário, no qual se dispõe de poucos *tweets* rotulados. Como já mencionado, tal abordagem reduz custos e esforços em uma etapa prévia de anotação do conjunto de treinamento obrigatoriamente necessária em problemas de classificação supervisionada.

Em resumo, as abordagens propostas cumpriram com êxito os objetivos desta tese sendo capazes de fornecer classificadores com propriedades bastante apreciadas e com grande potencial prático. Desta forma, as principais contribuições desta tese são listadas a seguir:

- A estratégia de combinar vários classificadores de sentimentos se mostrou bastante eficaz, levando a uma acurácia preditiva estável e, muitas vezes, superior àquela obtida por um classificador isoladamente.

- Com a combinação de classificadores e agrupadores, comprovou-se que as restrições fornecidas por meio de partições de dados (agrupamento) são úteis para classificar novos *tweets* quanto ao seu sentimento. Os resultados experimentais reportados mostraram, por meio do algoritmo C^{3}E (Coletta *et al.*, 2015a; Acharya *et al.*, 2014; Coletta *et al.*, 2015b, 2013; Acharya *et al.*, 2011), que tal premissa contribui para análise de sentimentos em textos curtos, em especial para *tweets*.

- Em relação à versão do algoritmo C^{3}E semissupervisionado, demonstrou-se com os experimentos a ampla abrangência de aplicação, pois as propriedades dos algoritmos são especialmente úteis em cenários com poucos *tweets* rotulados.

- Ainda em relação à versão do algoritmo C^{3}E semissupervisionado, este possui uma menor dependência do usuário/especialista de domínio, uma vez que seu uso é livre de parâmetros considerados críticos.

- Destaca-se a eficiência das abordagens, especialmente pela simplicidade e baixo custo computacional do algoritmo C^{3}E-SL(Coletta *et al.*, 2015a; Acharya *et al.*, 2014; da Silva *et al.*, 2014a; Coletta *et al.*, 2013; Acharya *et al.*, 2011) e ainda a versatilidade da ferramenta, pois o algoritmo C^{3}E-SL aceita o uso de (praticamente) qualquer componente nos agregadores, independente de atribuições rígidas ou suaves de classes e/ou grupos.

- Outro quesito importante está relacionado ao suporte à computação distribuída dos métodos propostos, pois os dados de treinamento não necessitam estar disponíveis após a

construção do agregador de classificadores ou após construção da matriz de similaridades.

- É interessante mencionar a possibilidade de paralelização da aplicação, pois esta importante característica está presente em todos componentes envolvidos – classificadores e agrupadores.

- Outra importante contribuição deste trabalho caracteriza-se pelo estudo comparativo realizado com as abordagens supervisionadas e semissupervisionadas do estado da arte, bem como as estratégias propostas nesta tese.

- A engenharia e seleção de atributos, uma importante etapa que antecede a definição e escolha do método de classificação de sentimentos, também foi estudada neste trabalho, sendo que várias configurações e tipos de atributos foram avaliadas: a *bag-of-words* tradicional, léxicos, *feature hashing*, o uso de *part-of-speech*, *hashtags*, menções, *emoticons*, e variações de *n-gramas*. Neste cenário, apesar de não existir um consenso sobre quais atributos (Pang *et al.*, 2002; Dave *et al.*, 2003; Kim & Hovy, 2006; Kouloumpis *et al.*, 2011; Lin & Kolcz, 2012; Ghiassi *et al.*, 2013; Zhu *et al.*, 2014) favorecem o desempenho do futuro classificador empregado, concluiu-se apartir dos experimentos realizados, que:

  - *Feature hashing* é uma técnica de redução de esparsidade (Lin & Kolcz, 2012; Caragea *et al.*, 2011; Weinberger *et al.*, 2009; Shi *et al.*, 2009; Ganchev & Dredze, 2008; Forman & Kirshenbaum, 2008), e portanto cumpre o seu papel ao fornecer ao classificador uma representação compacta, o que minimiza a complexidade das operações realizadas pelo mesmo. No entanto, esta redução pode afetar diretamente o desempenho do classificador, podendo reduzir o seu poder preditivo. Com uma decisão acertada de projeto pode-se minimizar este impacto determinando um limiar de redução de esparsidade aceitável em detrimento do desempenho e acurácia desejados.

  - O uso de léxicos e *emoticons* tanto em conjunto com *feature hashing* quanto combinados à *bag-of-words* contribuem para a capacidade preditiva do classificador. Entretato o uso isolado de léxicos e *emoticons* não provê informação suficiente para desempenho superior ou comparável às combinações mencionadas anteriormente.

  - Em um segundo momento o uso de *part-of-speech*, atributos exclusivos de redes sociais como *hashtags* e menções, foram também adicionados e também contribuíram para uma maior acurácia preditiva dos métodos de classificação.

  - Embora a combinação de unigramas, bigramas e trigramas contribua para o aumento da esparsidade, esta configuração contribui significativamente para uma maior acurácia preditiva dos métodos de classificação de sentimentos estudados.

5.2 Publicações Geradas e Artigos Submetidos

As contribuições obtidas com o desenvolvimento desta tese de doutorado foram (ou estão em fase de serem) divulgadas por meio de publicações em conferências e periódicos internacionais de destaque. A seguir, é apresentada uma lista completa dos artigos publicados e submetidos:

- SILVA, N. F. F.; COLETTA, L. F. S.; HRUSCHKA, E. R.; HRUSCHKA JR., E. R., *Combining Classification and Clustering for Tweet Sentiment Analysis*. In: Brazilian Conference on Intelligent Systems (BRACIS), 2014, São Carlos. Proceedings on Brazilian Conference on Intelligent Systems, p. 210-215, 2014 — (da Silva *et al.*, 2014a).

- SILVA, N. F. F.; HRUSCHKA, E. R.; HRUSCHKA JR., E. R., *Biocom Usp: Tweet Sentiment Analysis with Adaptive Boosting Ensemble*. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), 123-128, Dublin, Ireland, 2014 — (da Silva *et al.*, 2014b)

- SILVA, N. F. F.; HRUSCHKA, E. R.; HRUSCHKA JR., E. R., *Tweet sentiment analysis with classifier ensembles*. In: Decision Support Systems, Volume 66, Outubro 2014, Pages 170-179 ***(Qualis A2)*** — (da Silva *et al.*, 2014c).

- SILVA, N. F. F.; COLETTA, L. F. S.; HRUSCHKA, E. R.; HRUSCHKA JR., E. R., *Using Unsupervised Information to Improve Semi-supervised Tweet Sentiment Classification*, 2016 ***(Qualis A1)*** — (da Silva *et al.*, 2016b).

- SILVA, N. F. F.; COLETTA, L. F. S.; HRUSCHKA, E. R., *A Survey and Comparative Study of Tweet Sentiment Analysis via Semi-Supervised Learning*, 2015 ***(Qualis A1)*** — (da Silva *et al.*, 2016a).

5.3 Limitações e Perspectivas Futuras

No decorrer deste trabalho foram feitas opções que determinaram um rumo a seguir. Outras decisões de projeto poderiam ter sido tomadas e certamente outros resultados seriam encontrados e novas perspectivas se abririam. Desta forma, esta tese não representa um projeto finalizado, mas apenas uma parte do mesmo. Por conta disso, serão elencadas algumas sugestões para trabalhos a desenvolver que visam não só complementar o trabalho realizado como abrir novos percursos de investigação.

Nesta tese comprovou-se efetivamente que combinar classificadores contribui para a análise de sentimentos em *tweets*. A estratégia de combinação de classificadores adotada neste trabalho foi simplificada, conforme apresentado na seção 2.3. Entretanto, existem formas mais sofisticadas de combinação. Um exemplo disso é a extratégia de combinação por meio de meta-aprendizagem (Rokach, 2010), em que o objetivo é minimizar a taxa de erro dos classificadores transformando as predições dos mesmos em instâncias de treinamento, que por sua vez serão utilizadas para a geração de um novo classificador. Outro quesito a ser melhor investigado está

relacionado à importância e contribuição de cada classificador base para o *ensemble*, sendo possível por meio da definição de um *ranking* baseado no desempenho de classificadores definir pesos para a participação e contribuição dos mesmos para o agregador.

Considerando a combinação de classificadores e agrupadores realizada pelo C^{3}E-SL no cenário de análise de sentimentos, diversos aspectos podem ser melhor estudados em trabalhos futuros. O impacto da quantidade e dos tipos de classificadores usados no agregador, bem como maneiras eficientes de se obter matrizes de similaridades para o propósito de consenso são tópicos que merecem maior investigação. Em uma perspectiva mais geral, a importância relativa de cada um dos componentes dos agregadores poderia ser capturada na forma de pesos que, então, poderiam ser ajustados de acordo com a aplicação. Especificamente, pode-se explorar formas de se produzir entradas para o C^{3}E-SL que contenham um certo ganho de informação e não sejam redundantes a exemplo das estratégias adotadas para se obter componentes diversificados em agregadores (Naldi *et al.*, 2013; Windeatt, 2005; Kuncheva, 2004; Kuncheva & Whitaker, 2003a). Nesse contexto, um mecanismo mais inteligente para se gerar matrizes de similaridades poderia ser estudado ao se levar em conta a indução de agrupamentos a partir de diferentes "visões" dos dados com ênfase para a seleção de conjuntos de atributos mais promissores (os quais, por exemplo, poderiam ser aqueles que melhor discriminam as classes). A partir disso, é importante conduzir pesquisas sobre medidas de desempenho úteis para a avaliação das partições de dados e a matriz de similaridades especialmente em *tweets*, por se tratar de textos curtos e provenientes de redes sociais.

Vários estudos na literatura buscaram explorar abordagens de aprendizagem de máquina para resolver tarefas de análise de sentimento de diferentes perspectivas nos últimos 15 anos, como demonstrado nesta tese. Uma vez que o desempenho do algoritmo de aprendizagem é fortemente dependente das escolhas de representações de dados, muitos estudos se dedicaram à engenharia e seleção de atributos (Mohammad *et al.*, 2013; Zhu *et al.*, 2014). Recentemente, abordagens de aprendizagem em profundidade (*deep learning*) (Bengio *et al.*, 2015; Ranzato *et al.*, 2015; Zhang & LeCun, 2015; Tang *et al.*, 2015a,b; Tang, 2015) emergiram como modelos computacionais poderosos com a promessa de descobrir representações semânticas intrínsecas em textos, automaticamente a partir dos dados, sem o recurso da engenharia. Essas abordagens têm melhorado o estado da arte em muitas tarefas de análise de sentimento incluindo a classificação de sentimento de sentenças/documentos e de aprendizagem de léxicos sentimento. Pretende-se com isso incorporar em métodos de análise de sentimentos semissupervisionados técnicas de *deep-learning*.

Ao se utilizar a versão semissupervisionada do C^{3}E-SL, estudos para considerar a informação não supervisionada, em conjunto com as probabilidades de classes, para se obter um critério mais refinado de seleção de objetos no autotreinamento (*self-training*) se mostram bem interessantes. Neste contexto, a maior dificuldade reside na obtenção de informação não supervisionada complementar que possa, de fato, alavancar o modelo semissupervisionado. Portanto, investigações que explorem a geração eficiente de matrizes de similaridades que sejam proveitosas para o autotreinamento são significativamente úteis.

Nesta tese não foi explorado o conceito de análise de sentimentos com granularidade de as-

pectos. Tal conceito vêm sendo recentemente explorado no cenário semissupervisionado (Matsuno *et al.*, 2015; Carter & Inkpen, 2015), com resultados significativos e promissores. Avaliar a cooperação do agrupamento e classificação como método de extração de aspectos e análise de sentimentos ocorrendo de forma simultânea é também um desafio interessante de pesquisa.

Nesta tese, tanto nas análises realizadas com o aprendizado supervisionado quanto nas análises realizadas com o aprendizado semissupervisionado, admitiu-se a existência de um conjunto de treinamento que é dado e fixo. Na prática, é muito conveniente utilizar o *active learning* (Hajmohammadi *et al.*, 2015; Smailović *et al.*, 2014; Settles, 2012; Zhu *et al.*, 2003b; Tong & Koller, 2001; Cohn *et al.*, 1995) ou aprendizado ativo em conjunto com o aprendizado semissupervisionado. Essa abordagem pode permitir que o algoritmo de aprendizado escolha um conjunto de *tweets* a serem rotulados por um especialista de domínio, ao invés de utilizar de uma seleção aleatória dos mesmos.

Outro viés de pesquisa sugere o quão interessante seria incorporar "a força" do sentimento ao trabalho, ou atribuições mais suaves de classes — área de pesquisa conhecida como *opinion strength* (Turney & Littman, 2003). Nesta tese, foram considerados os sentimentos positivo, negativo ou neutro relacionados à um *tweet*, enquanto que em linguagem natural, as opiniões não são tão rígidas. Se um consumidor está usando um sistema comparativo de avaliações de sentimento, é particularmente vantajoso que a força das opiniões seja conhecida e agregada (e.g., um celular com muitas opiniões fracamente negativas deveria ser preferível a um similar com muitas opiniões fortemente negativas). Existem poucos trabalhos com foco em classificação baseada na força do sentimento (*strength-based sentiment classification*), por exemplo, Wilson *et al.* (2004) e Turney & Littman (2003).

Na vida real os *tweets* chegam como fluxos de dados ou *data streams* (Domingos & Hulten, 2000; Dahal *et al.*, 2015). Desta forma, é importante que a abordagem usada para classificação de sentimentos seja dinâmica e capaz de manipular com eficácia e eficiência fluxos contínuos de *tweets*. Esta caraterística não foi explorada nesta tese, uma vez que para os experimentos foram considerados conjuntos de dados estáticos. Fluxo de dados em análise de sentimentos vêm sendo muito estudado (Lourenco Jr. *et al.*, 2014; Kim *et al.*, 2013; Mejova & Srinivasan, 2012; Bifet *et al.*, 2011; Bifet & Frank, 2010), porém sem foco na estratégia de usar agrupadores para refinar e melhorar o poder de generalização de classificadores.

Referências Bibliográficas

(2002). Forecasting the {NYSE} composite index with technical analysis, pattern recognizer, neural network, and genetic algorithm: a case study in romantic decision support. *Decision Support Systems*, v.32, n.4, p.361 – 377.

(2004). Detection of land-cover transitions by combining multidate classifiers. *Pattern Recognition Letters*, v.25, n.13, p.1491 – 1500. Pattern Recognition for Remote Sensing (PRRS 2002).

(2004). Model selection for medical diagnosis decision support systems. *Decision Support Systems*, v.36, n.3, p.247 – 259.

Abelson, R. P. (1983). Whatever became a consistency theory. *Pesonality and Social Psychology Bulletin*, v.9, p.37–54.

Acharya, A.; Hruschka, E.; Ghosh, J.; Acharyya, S. (2011). C3E: A framework for combining ensembles of classifiers and clusterers. *Multiple Classifier Systems*, v. 6713 de *Lecture Notes in Computer Science*, p. 269–278. Springer Berlin Heidelberg.

Acharya, A.; Hruschka, E. R.; Ghosh, J.; Acharyya, S. (2014). An optimization framework for combining ensembles of classifiers and clusterers with applications to nontransductive semisupervised learning and transfer learning. *ACM Transactions on Knowledge Discovery from Data*, v.9, n.1, p.1:1–1:35.

Agarwal, A.; Xie, B.; Vovsha, I.; Rambow, O.; Passonneau, R. (2011). Sentiment analysis of twitter data. *Proceedings of the Workshop on Languages in Social Media*, LSM '11, p. 30–38, Stroudsburg, PA, USA. Association for Computational Linguistics.

Aisopos, F.; Papadakis, G.; Varvarigou, T. (2011). Sentiment analysis of social media content using n-gram graphs. *Proceedings of the 3rd ACM SIGMM International Workshop on Social Media*, WSM '11, p. 9–14, New York, NY, USA. ACM.

Asiaee T., A.; Tepper, M.; Banerjee, A.; Sapiro, G. (2012). If you are happy and you know it... tweet. *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, p. 1602–1606, New York, NY, USA. ACM.

Baccianella, S.; Esuli, A.; Sebastiani, F. (2010a). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. Chair), N. C. C.; Choukri, K.; Maegaard, B.; Mariani, J.; Odijk, J.; Piperidis, S.; Rosner, M.; Tapias, D., editores, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Baccianella, S.; Esuli, A.; Sebastiani, F. (2010b). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. *in Proc. of LREC*.

Bai, R.; Chai, T.; Ma, E. (2007). A novel approach for extraction of fuzzy rules using the neuro-fuzzy network and its application in the blending process of raw slurry. *Advances in Neural Networks, ISNN 2007*, v. 4492 de *Lecture Notes in Computer Science*, p. 362–370. Springer Berlin/Heidelberg.

Bakliwal, A.; Arora, P.; Madhappan, S.; Kapre, N.; Singh, M.; Varma, V. (2012). Mining sentiments from tweets. *Proceedings of the 3rd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA)*, v. 12.

Balahur, A.; Steinberger, R.; Kabadjov, M.; Zavarella, V.; van der Goot, E.; Halkia, M.; Pouliquen, B.; Belyaeva, J. (2010). Sentiment analysis in the news. Chair), N. C. C.; Choukri, K.; Maegaard, B.; Mariani, J.; Odijk, J.; Piperidis, S.; Rosner, M.; Tapias, D., editores, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Barbosa, L.; Feng, J. (2010). Robust sentiment detection on twitter from biased and noisy data. *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, p. 36–44, Stroudsburg, PA, USA. Association for Computational Linguistics.

Basch, D. (2012). Some fresh twitter stats (as of july 2012, dataset included). http://diegobasch.com/some-fresh-twitter-stats-as-of-july-2012. [Online; accessed 25-June-2013].

Basu, S.; Bilenko, M.; Mooney, R. J. (2004). A probabilistic framework for semi-supervised clustering. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, p. 59–68, New York, NY, USA. ACM.

Basu, S.; Davidson, I.; Wagstaff, K. (2008). *Constrained Clustering: Advances in Algorithms, Theory, and Applications*. Chapman & Hall/CRC. Edição 1.

Bauer, E.; Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine Learning*, v. 36, p. 105–139. Kluwer Academic Publishers, Hingham, MA, USA.

Baugh, W. (2013). bwbaugh : Hierarchical sentiment analysis with partial self-training. *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings*

of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), p. 539–542, Atlanta, Georgia, USA. Association for Computational Linguistics.

Becker, L.; Erhart, G.; Skiba, D.; Matula, V. (2013). Avaya: Sentiment analysis on twitter with self-training and polarity lexicon expansion. *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, p. 333–340, Atlanta, Georgia, USA. Association for Computational Linguistics.

Belkin, M.; Matveeva, I.; Niyogi, P. (2004). Regularization and semi-supervised learning on large graphs. *COLT*, v. 3120, p. 624–638. Springer.

Bengio, Y.; Goodfellow, I. J.; Courville, A. (2015). Deep learning. Book in preparation for MIT Press.

Bermingham, A.; Smeaton, A. F. (2010). Classifying sentiment in microblogs: Is brevity an advantage? *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, p. 1833–1836, New York, NY, USA. ACM.

Bezdek, J. C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. Kluwer Academic Publishers, Norwell, MA, USA.

Bifet, A.; Frank, E. (2010). Sentiment knowledge discovery in twitter streaming data. *Proceedings of the 13th International Conference on Discovery Science*, DS'10, p. 1–15, Berlin, Heidelberg. Springer-Verlag.

Bifet, A.; Holmes, G.; Pfahringer, B. (2011). Moa-tweetreader: Real-time analysis in twitter streaming data. *Proceedings of the 14th International Conference on Discovery Science*, DS'11, p. 46–60, Berlin, Heidelberg. Springer-Verlag.

Bishop, C. M. (2007). *Pattern Recognition and Machine Learning - Information Science and Statistics*. Springer. Edição 1.

Blei, D. M.; Ng, A. Y.; Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, v.3, p.993–1022.

Blum, A.; Chawla, S. (2001). Learning from labeled and unlabeled data using graph mincuts. *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*, p. 19–26.

Blum, A.; Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. *Proceedings of the eleventh annual conference on Computational learning theory*, COLT' 98, p. 92–100, New York, NY, USA. ACM.

Bollen, J.; Mao, H.; Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*.

Bora, N. N. (2012). Summarizing public opinions in tweets. New Delhi, India.

Boser, B. E.; Guyon, I. M.; Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *Proceedings of the fifth annual workshop on Computational learning theory*, p. 144–152. ACM.

Bousquet, O.; Boucheron, S.; Lugosi, G. (2003). Introduction to statistical learning theory. Bousquet, O.; von Luxburg, U.; Rätsch, G., editores, *Advanced Lectures on Machine Learning*, v. 3176 de *Lecture Notes in Computer Science*, p. 169–207. Springer.

Bousquet, O.; Boucheron, S.; Lugosi, G. (2004). Introduction to Statistical Learning Theory. *Advanced Lectures on Machine Learning*, p. 169–207.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, v. 24, p. 123–140. Kluwer Academic Publishers, Hingham, MA, USA.

Brody, S.; Diakopoulos, N. (2011). Cooooooooooooooollllllllllllll!!!!!!!!!!!!!!: using word lengthening to detect sentiment in microblogs. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, p. 562–570, Stroudsburg, PA, USA. Association for Computational Linguistics.

Brown, G.; Wyatt, J.; Harris, R.; Yao, X. (2005). Diversity creation methods: A survey and categorisation. *Information Fusion*, v.6, p.5–20.

Bullas, J. (2012). 48 significant social media facts, figures and statistics plus 7 infographics. http://www.jeffbullas.com/2012/04/23/48-significant-social-media-facts-figures-and-statistics-plus-7-infographics/#e5fZgwTPL4wacMHt.99. [Online; accessed 25-June-2013].

Cai, W.; Chen, S.; Zhang, D. (2007). Robust fuzzy relational classifier incorporating the soft class labels. *Pattern Recognition Letters*, v.28, p.2250–2263.

Cai, W.; Chen, S.; Zhang, D. (2009). A simultaneous learning framework for clustering and classification. *Pattern Recognition Letters*, v.42, n.7, p.1248–1259.

Cai, W.; Chen, S.; Zhang, D. (2010). A multiobjective simultaneous learning framework for clustering and classification. *Neural Networks, IEEE Transactions*, v.21, n.2, p.185–200.

Caragea, C.; Silvescu, A.; Mitra, P. (2011). Protein sequence classification using feature hashing. Wu, F.-X.; Zaki, M. J.; Morishita, S.; Pan, Y.; Wong, S.; Christianson, A.; Hu, X., editores, *BIBM*, p. 538–543. IEEE.

Carter, D.; Inkpen, D. (2015). Inferring aspect-specific opinion structure in product reviews using co-training. Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing*, v. 9042 de *Lecture Notes in Computer Science*, p. 225–240. Springer International Publishing.

Carvalho, P.; Sarmento, L.; Silva, M. J.; de Oliveira, E. (2009). Clues for detecting irony in user-generated contents: Oh...!! it's "so easy";-). *Proceedings of the 1st International CIKM Workshop on Topic-sentiment Analysis for Mass Opinion*, TSA '09, p. 53–56, New York, NY, USA. ACM.

Chang, C.-C.; Lin, C.-J. (2011). Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, v.2, n.3, p.27:1–27:27.

Chapelle, O.; Scholkopf, B.; Zien, A., editores (2006). *Semi-Supervised Learning*. MIT Press, Cambridge, MA.

Clark, S.; Wicentwoski, R. (2013). Swatcs: Combining simple classifiers with estimated accuracy. *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, p. 425–429, Atlanta, Georgia, USA. Association for Computational Linguistics.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, v.5, n.4, p.559 – 583.

Cohn, D. A.; Ghahramani, Z.; Jordan, M. I. (1995). Active learning with statistical models.

Coletta, L.; Hruschka, E.; Acharya, A.; Ghosh, J. (2013). Towards the use of metaheuristics for optimizing the combination of classifier and cluster ensembles. *Computational Intelligence and 11th Brazilian Congress on Computational Intelligence (BRICS-CCI CBIC), 2013 BRICS Congress on*, p. 483–488.

Coletta, L. F.; Hruschka, E. R.; Acharya, A.; Ghosh, J. (2015a). Using metaheuristics to optimize the combination of classifier and cluster ensembles. *Integrated Computer-Aided Engineering*, v.22, n.3, p.229–242.

Coletta, L. F. S.; Hruschka, E. R.; Acharya, A.; Ghosh, J. (2015b). A differential evolution algorithm to optimise the combination of classifier and cluster ensembles. *International Journal of Bio-Inspired Computation*, v.7, n.2, p.111–124.

Cortes, C.; Vapnik, V. (1995). Support-vector networks. *Machine learning*, v.20, n.3, p.273–297.

Crammer, K.; Kulesza, A.; Dredze, M. (2009). Adaptive regularization of weight vectors. *NIPS*, p. 414–422.

Cui, A.; Zhang, M.; Liu, Y.; Ma, S. (2011). Emotion tokens: Bridging the gap among multilingual twitter sentiment analysis. Salem, M.; Shaalan, K.; Oroumchian, F.; Shakery, A.; Khelalfa, H., editores, *Information Retrieval Technology*, v. 7097 de *Lecture Notes in Computer Science*, p. 238–249. Springer Berlin Heidelberg.

da Silva, N.; Coletta, L.; Hruschka, E.; Hruschka, E. (2014a). Combining classification and clustering for tweet sentiment analysis. *Intelligent Systems (BRACIS), 2014 Brazilian Conference on*, p. 210–215.

da Silva, N. F.; Coletta, L. F.; Hruschka, E. R. (2016a). A survey and comparative study of tweet sentiment analysis via semi-supervised learning. *ACM Computing Surveys*, v.–, n.–, p.000–000.

da Silva, N. F.; Hruschka, E. R.; Jr., E. R. H. (2014b). Biocom usp: Tweet sentiment analysis with adaptive boosting ensemble. *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, p. 129–134, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

da Silva, N. F.; Hruschka, E. R.; Jr., E. R. H. (2014c). Tweet sentiment analysis with classifier ensembles. *Decision Support Systems*, v.66, p.170 – 179.

da Silva, N. F. F.; Coletta, L. F.; Hruschka, E. R.; Jr., E. R. H. (2016b). Using unsupervised information to improve semi-supervised tweet sentiment classification. *Information Sciences*, v., p.–.

Dahal, N.; Abuomar, O.; King, R.; Madani, V. (2015). Event stream processing for improved situational awareness in the smart grid. *Expert Systems with Applications*, v.42, n.20, p.6853 – 6863.

Dave, K.; Lawrence, S.; Pennock, D. M. (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. *Proceedings of the 12th International Conference on World Wide Web*, WWW '03, p. 519–528, New York, NY, USA. ACM.

Davidov, D.; Tsur, O.; Rappoport, A. (2010). Enhanced sentiment learning using twitter hashtags and smileys. *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, p. 241–249, Stroudsburg, PA, USA. Association for Computational Linguistics.

De Castro, L.; Hruschka, E.; Campello, R. (2004). An evolutionary clustering technique with local search to design rbf neural network classifiers. *Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on*, v. 3, p. 2083 – 2088 vol.3.

de Freitas, L. A.; Vanin, A. A.; Hogetop, D. N.; Bochernitsan, M. N.; Vieira, R. (2014). Pathways for irony detection in tweets. *Proceedings of the 29th Annual ACM Symposium on Applied Computing*, SAC '14, p. 628–633, New York, NY, USA. ACM.

Dean, J.; Ghemawat, S. (2008). Mapreduce: simplified data processing on large clusters. *Commun. ACM*, v.51, n.1, p.107–113.

Dempster, A. P.; Laird, N. M.; Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, v.39, n.1, p.1–38.

Diakopoulos, N. A.; Shamma, D. A. (2010). Characterizing debate performance via aggregated twitter sentiment. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, p. 1195–1198, New York, NY, USA. ACM.

Dietterich, T. G. (2000). Ensemble methods in machine learning. *International Workshop on Multiple Classifier Systems*, p. 1–15. Springer-Verlag.

Domingos, P.; Hulten, G. (2000). Mining high-speed data streams. *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '00, p. 71–80, New York, NY, USA. ACM.

Drury, B.; Torgo, L.; Almedia, J. J. (2011). Guided self training for sentiment classification. *Robust Unsupervised and Semi-Supervised Methods in Natural Language Processing*, p. 18–25.

Duval-Poo, M. A.; Sosa-Garcia, J.; Guerra-Gandon, A.; Vega-Pons, S.; Ruiz-Shulcloper, J. (2012). A new classifier combination scheme using clustering ensemble. *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, v. 7441 de *Lecture Notes in Computer Science*, p. 154–161. Springer Berlin Heidelberg.

Efron, B.; Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, New York.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *J. Mach. Learn. Res.*, v.3, p.1289–1305.

Forman, G.; Kirshenbaum, E. (2008). Extremely fast text feature extraction for classification and indexing. *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, p. 1221–1230, New York, NY, USA. ACM.

Freund, Y.; Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting,. *Journal of Computer and System Sciences*, v.55, n.1, p.119 – 139.

Fumera, G.; Roli, F. (2005). A theoretical and experimental analysis of linear combiners for multiple classifier systems. *IEEE Trans. Pattern Anal. Mach. Intell.*, v.27, n.6, p.942–956.

Gan, H.; Sang, N.; Huang, R.; Tong, X.; Dan, Z. (2013). Using clustering analysis to improve semi-supervised classification. *Neurocomputing*, v.101, n.0, p.290–298.

Ganchev, K.; Dredze, M. (2008). Small statistical models by random feature mixing. *Proceedings of the ACL-2008 Workshop on Mobile Language Processing*. Association for Computational Linguistics.

Gao, J.; Liang, F.; Fan, W.; Sun, Y.; Han, J. (2013). A graph-based consensus maximization approach for combining multiple supervised and unsupervised models. *Knowledge and Data Engineering, IEEE Transactions on*, v.25, n.1, p.15–28.

Gao, J.; Liangy, F.; Fanz, W.; Suny, Y.; Han, J. (2009). Graph-based Consensus Maximization among Multiple Supervised and Unsupervised Models. *23rd Annual Conference on Neural Information Processing Systems*.

Ghaemi, R.; Sulaiman, N.; Ibrahim, H.; Mustapha, N. (2009). A Survey: Clustering Ensembles Techniques. *Proceedings of World Academy of Science, Engineering and Technology*, v.38.

Ghiassi, M.; Skinner, J.; Zimbra, D. (2013). Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. *Expert Systems with Applications*, v.40, n.16, p.6266 – 6282.

Go, A.; Bhayani, R.; Huang, L. (2009). Twitter sentiment classification using distant supervision. *Processing*, p. 1–6.

Goldberg, A. (2010). *New directions in semi-supervised learning*. Tese (Doutorado), University of Wisconsin - Madison.

Gonzalez-Ibanez, R.; Muresan, S.; Wacholder, N. (2011). Identifying sarcasm in twitter: A closer look. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, p. 581–586, Stroudsburg, PA, USA. Association for Computational Linguistics.

Günther, T.; Furrer, L. (2013). Gu-mlt-lt: Sentiment analysis of short messages using linguistic features and stochastic gradient descent. *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, p. 328–332, Atlanta, Georgia, USA. Association for Computational Linguistics.

Hagen, M.; Potthast, M.; Büchner, M.; Stein, B. (2015). Twitter sentiment detection via ensemble classification using averaged confidence scores. *Advances in Information Retrieval - 37th European Conference on IR Research, ECIR 2015, Vienna, Austria, March 29 - April 2, 2015. Proceedings*, p. 741–754.

Hajmohammadi, M. S.; Ibrahim, R.; Selamat, A.; Fujita, H. (2015). Combination of active learning and self-training for cross-lingual sentiment classification with density analysis of unlabelled samples. *Information Sciences*, v.317, n.0, p.67 – 77.

Hamdan, H.; Béchet, F.; Bellot, P. (2013). Experiments with dbpedia, wordnet and sentiwordnet as resources for sentiment analysis in micro-blogging. *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, p. 455–459, Atlanta, Georgia, USA. Association for Computational Linguistics.

Hansen, L.; Salamon, P. (1990). Neural network ensembles. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, v.12, n.10, p.993 –1001.

Hassan, A.; Abbasi, A.; Zeng, D. (2013). Twitter sentiment analysis: A bootstrap ensemble framework. *SocialCom*, p. 357–364. IEEE.

Hassan Saif, Miriam Fernandez, Y. H.; Alani, H. (2013). Evaluation datasets for twitter sentiment analysis: A survey and a new dataset, the sts-gold. *first ESSEM workshop*.

Hastie, T. J.; Tibshirani, R. J.; Friedman, J. H. (2009). *The elements of statistical learning : data mining, inference, and prediction*. Springer series in statistics. Springer, New York.

Hatfield, E.; Cacioppo, J. T.; Rapson, R. L. (1993). Emotional contagion.

He, Y.; Lin, C.; Gao, W.; Wong, K.-F. (2012). Tracking sentiment and topic dynamics from social media. *ICWSM*.

He, Y.; Zhou, D. (2011). Self-training from labeled features for sentiment analysis. *Information Processing & Management*, v.47, n.4, p.606 – 616.

Hong, S.; Lee, J.; Lee, J.-H. (2014). Competitive self-training technique for sentiment analysis in mass social media. *Soft Computing and Intelligent Systems (SCIS), 2014 Joint 7th International Conference on and Advanced Intelligent Systems (ISIS), 15th International Symposium on*, p. 9–12.

Hsu, C.-W.; Chang, C.-C.; Lin, C.-J.; others (2010). A practical guide to support vector classification, 2003. *Paper available at http://www.csie.ntu.edu.tw/ cjlin/papers/guide/guide.pdf*.

Hu, M.; Liu, B. (2004). Mining and summarizing customer reviews. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, p. 168–177, New York, NY, USA. ACM.

Hu, X.; Tang, J.; Gao, H.; Liu, H. (2013a). Unsupervised sentiment analysis with emotional signals. *Proceedings of the 22nd international conference on World Wide Web*, WWW'13. ACM.

Hu, X.; Tang, L.; Tang, J.; Liu, H. (2013b). Exploiting social relations for sentiment analysis in microblogging. *Proceedings of the sixth ACM international conference on Web search and data mining*.

Jaakkola, M. S. T.; Szummer, M. (2002). Partially labeled classification with markov random walks. *Advances in neural information processing systems (NIPS)*, v.14, p.945–952.

Jiang, L.; Yu, M.; Zhou, M.; Liu, X.; Zhao, T. (2011). Target-dependent twitter sentiment classification. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, p. 151–160, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jiliang, T.; Chikashi, N.; Anlei, D.; Yi, C.; Huan, L. (2015). Propagation-based sentiment analysis for microblogging data. *Proceedings of the 2015 SIAM International Conference on Data Mining*, p. 577–585.

Jo, Y.; Oh, A. H. (2011). Aspect and sentiment unification model for online review analysis. *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, p. 815–824, New York, NY, USA. ACM.

Joachims, T. (1999a). Making large-scale SVM learning practical. Schölkopf, B.; Burges, C.; Smola, A., editores, *Advances in Kernel Methods - Support Vector Learning*, chapter 11, p. 169–184. MIT Press, Cambridge, MA.

Joachims, T. (1999b). Transductive inference for text classification using support vector machines. *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99, p. 200–209, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Johnson, C.; Shukla, P.; Shukla, S. (2011). On classifying the political sentiment of tweets.

Jones, K. (1994). Natural language processing: A historical review. Zampolli, A.; Calzolari, N.; Palmer, M., editores, *Current Issues in Computational Linguistics: In Honour of Don Walker*, v. 9 de *Linguistica Computazionale*, p. 3–16. Springer Netherlands.

Kanakaraj, M.; Guddeti, R. (2015). Performance analysis of ensemble methods on twitter sentiment analysis using nlp techniques. *Semantic Computing (ICSC), 2015 IEEE International Conference on*, p. 169–170.

Kane, B. (2012). Twitter stats in 2012. http://marketingland.com/social-network-demographics-pew-study-shows-who-uses-facebook-twitter-pinterest-others-21594. [Online; accessed 25-June-2013].

Kaufman, L.; Rousseeuw, P. (1987). *Clustering by means of medoids*. North-Holland.

Kim, H.-G.; Lee, S.; Kyeong, S. (2013). Discovering hot topics using twitter streaming data: Social topic detection and geographic clustering. *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '13, p. 1215–1220, New York, NY, USA. ACM.

Kim, S.-M.; Hovy, E. (2006). Automatic identification of pro and con reasons in online reviews. *Proceedings of the COLING/ACL on Main Conference Poster Sessions*, COLING-ACL '06, p. 483–490, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kiritchenko, S.; Zhu, X.; Mohammad, S. (2014). Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, v.50, p.724–762.

Kollios, G.; Gunopulos, D.; Koudas, N.; Berchtold, S. (2003). Efficient biased sampling for approximate clustering and outlier detection in large data sets. *Knowledge and Data Engineering, IEEE Transactions on*, v.15, n.5, p.1170 – 1187.

Kouloumpis, E.; Wilson, T.; Moore, J. (2011). Twitter sentiment analysis: The good the bad and the omg! *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media - ICWSM'11*.

Krogh, A.; Vedelsby, J. (1995). Neural network ensembles, cross validation, and active learning. *Advances in Neural Information Processing Systems*, p. 231–238. MIT Press.

Kulis, B.; Basu, S.; Dhillon, I.; Mooney, R. (2005). Semi-supervised graph clustering: a kernel approach. *Proceedings of the 22nd international conference on Machine learning*, ICML '05, p. 457–464, New York, NY, USA. ACM.

Kuncheva, L.; Hadjitodorov, S.; Todorova, L. (2006). Experimental comparison of cluster ensemble methods. *Information Fusion, 2006 9th International Conference on*, p. 1–7.

Kuncheva, L.; Whitaker, C. (2003a). Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, v.51, n.2, p.181–207.

Kuncheva, L. I. (2004). *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience.

Kuncheva, L. I.; Hadjitodorov, S. T. (2004). Using diversity in cluster ensembles. *Systems, man and cybernetics, 2004 IEEE international conference on*, v. 2, p. 1214–1219. IEEE.

Kuncheva, L. I.; Whitaker, C. J. (2003b). Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Mach. Learn.*, v.51, n.2, p.181–207.

Lazarsfeld, P. F.; Merton, R. K. (1954). Friendship as a social process: A substantive and methodological analysis. Berger, M.; Abel, T.; Page, C., editores, *Freedom and Control in Modern Society*, p. 18–66. Van Nostrand, New York.

Li, S.; Wang, Z.; Zhou, G.; Lee, S. Y. M. (2011). Semi-supervised learning for imbalanced sentiment classification. *IJCAI*, p. 1826–1831.

Li, T.; Sindhwani, V.; Ding, C. H. Q.; 0005, Y. Z. (2010). Bridging domains with words: Opinion analysis with matrix tri-factorizations. *Proceedings of the SIAM International Conference on Data Mining, SDM 2010, April 29 - May 1, 2010, Columbus, Ohio, USA*, p. 293–302. SIAM.

Li, X.; Ye, N. (2006). A supervised clustering and classification algorithm for mining data with mixed variables. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, v.36, n.2, p.396 – 406.

Lin, J.; Kolcz, A. (2012). Large-scale machine learning at twitter. *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD '12, p. 793–804, New York, NY, USA. ACM.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Liu, B.; Zhang, L. (2012). A survey of opinion mining and sentiment analysis. Aggarwal, C. C.; Zhai, C., editores, *Mining Text Data*, p. 415–463. Springer US.

Liu, K.-L.; Li, W.-J.; Guo, M. (2012). Emoticon smoothed language models for twitter sentiment analysis. *AAAI'12*, p. –1–1.

Liu, S.; Cheng, X.; Li, F.; Li, F. (2015). Tasc:topic-adaptive sentiment classification on dynamic tweets. *Knowledge and Data Engineering, IEEE Transactions on*, v.27, n.6, p.1696–1709.

Liu, S.; Li, F.; Li, F.; Cheng, X.; Shen, H. (2013a). Adaptive co-training SVM for sentiment classification on tweets. *Proceedings of the 22Nd ACM International Conference on Conference on Information Knowledge Management*, CIKM '13, p. 2079–2088, New York, NY, USA. ACM.

Liu, S.; Zhu, W.; Xu, N.; Li, F.; Cheng, X.-q.; Liu, Y.; Wang, Y. (2013b). Co-training and visualizing sentiment evolvement for tweet events. *Proceedings of the 22Nd International Conference on World Wide Web Companion*, WWW '13 Companion, p. 105–106, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Lourenco Jr., R.; Veloso, A.; Pereira, A.; Meira Jr., W.; Ferreira, R.; Parthasarathy, S. (2014). Economically-efficient sentiment stream analysis. *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, p. 637–646, New York, NY, USA. ACM.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. Cam, L. M. L.; Neyman, J., editores, *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, v. 1, p. 281–297. University of California Press.

Maimon, O.; Rokach, L. (2004). Ensemble of decision trees for mining manufacturing data sets. *Machine Engineering*, v.4, n.1-2.

Masud, M.; Chen, Q.; Khan, L.; Aggarwal, C.; Gao, J.; Han, J.; Srivastava, A.; Oza, N. (2013). Classification and adaptive novel class detection of feature-evolving data streams. *Knowledge and Data Engineering, IEEE Transactions on*, v.25, n.7, p.1484–1497.

Masud, M.; Chen, Q.; Khan, L.; Aggarwal, C.; Gao, J.; Han, J.; Thuraisingham, B. (2010a). Addressing concept-evolution in concept-drifting data streams. *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, p. 929–934.

Masud, M.; Gao, J.; Khan, L.; Han, J.; Thuraisingham, B. (2010b). Classification and novel class detection in data streams with active mining. *Advances in Knowledge Discovery and Data Mining*, v. 6119 de *Lecture Notes in Computer Science*, p. 311–324. Springer Berlin/Heidelberg.

Matsuno, P. I.; Rossi, G. R.; Marcani, M. R.; Rezede, O. S. (2015). Análise de sentimentos baseada em aspectos usando aprendizado semissupervisionado em redes heterogêneas. *Symposium on Knowledge Discovery, Mining and Learning, KDMILE 2015.*

McGee, M. (2012). Social network demographics: Pew study shows who uses facebook, twitter, pinterest and others. http://marketingland.com/social-network-demographics-pew-study-shows-who-uses-facebook-twitter-pinterest-others-21594. [Online; accessed 25-June-2013].

McPherson, M.; Smith-Lovin, L.; Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, v.27, n.1, p.415–444.

Mei, Q.; Ling, X.; Wondra, M.; Su, H.; Zhai, C. (2007). Topic sentiment mixture: Modeling facets and opinions in weblogs. *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, p. 171–180, New York, NY, USA. ACM.

Mejova, Y.; Srinivasan, P. (2012). Political speech in social media streams: Youtube comments and twitter posts. *Proceedings of the 4th Annual ACM Web Science Conference*, WebSci '12, p. 205–208, New York, NY, USA. ACM.

Miller, D.; Browning, J. (2003). A mixture model and em-based algorithm for class discovery, robust classification, and outlier rejection in mixed labeled/unlabeled data sets. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, v.25, n.11, p.1468 – 1483.

Miller, G. A. (1995a). Wordnet: A lexical database for english. *Commun. ACM*, v.38, n.11, p.39–41.

Miller, G. A. (1995b). Wordnet: A lexical database for english. *Commun. ACM*, v.38, n.11, p.39–41.

Miura, Y.; Sakaki, S.; Hattori, K.; Ohkuma, T. (2014). Teamx: A sentiment analyzer with enhanced lexicon mapping and weighting scheme for unbalanced data. *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, p. 628–632, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Mohammad, S. M.; Kiritchenko, S.; Zhu, X. (2013). Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*, Atlanta, Georgia, USA.

Mostafa, M. (2013). An emotional polarity analysis of consumers airline service tweets. *Social Network Analysis and Mining*, p. 1–15.

Nakov, P.; Rosenthal, S.; Kozareva, Z.; Stoyanov, V.; Ritter, A.; Wilson, T. (2013). Semeval-2013 task 2: Sentiment analysis in twitter. *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, p. 312–320, Atlanta, Georgia, USA. Association for Computational Linguistics.

Naldi, M. C.; Carvalho, A.; Campello, R. J. (2013). Cluster ensemble selection based on relative validity indexes. *Data Mining and Knowledge Discovery*, v.27, n.2, p.259–289.

Narr, S.; Hulfenhaus, M.; Albayrak, S. (2012). Language-independent twitter sentiment analysis. *Knowledge Discovery and Machine Learning (KDML), LWA*, p. 12–14.

Nielsen, F. Å. (2011a). A new anew: Evaluation of a word list for sentiment analysis in microblogs. *CoRR*, v.abs/1103.2903.

Nielsen, F. Å. (2011b). A new ANEW: evaluation of a word list for sentiment analysis in microblogs. *CoRR*, v.abs/1103.2903.

Nigam, K.; Mccallum, A.; Thrun, S.; Mitchell, T. (2000). Text classification from labeled and unlabeled documents using em. *Machine Learning*, v.39, n.2-3, p.103–134.

O'Connor, B.; Balasubramanyan, R.; Routledge, B. R.; Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM'10*, p. 1–1.

Ortega Bueno, R.; Fonseca Bruzón, A.; Gutiérrez, Y.; Montoyo, A. (2013). Ssa-uo: Unsupervised sentiment analysis in twitter. *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, p. 501–507, Atlanta, Georgia, USA. Association for Computational Linguistics.

Owoputi, O.; Dyer, C.; Gimpel, K.; Schneider, N.; Smith, N. A. (2013a). Improved part-of-speech tagging for online conversational text with word clusters. *In Proceedings of NAACL.*

Owoputi, O.; Dyer, C.; Gimpel, K.; Schneider, N.; Smith, N. A. (2013b). Improved part-of-speech tagging for online conversational text with word clusters. *In Proceedings of NAACL.*

Oza, N. C.; Tumer, K. (2008). Classifier ensembles: Select real-world applications. *Information Fusion*, v.9, n.1, p.4–20.

Pak, A.; Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Pang, B.; Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, v.2, n.1-2, p.1–135.

Pang, B.; Lee, L.; Vaithyanathan, S. (2002). Thumbs up? sentiment classification using machine learning techniques. *Proceedings of EMNLP*, p. 79–86.

Papadimitriou, S.; Mavroudi, S.; Vladutu, L.; Bezerianos, A. (2001). Ischemia detection with a self-organizing map supplemented by supervised learning. *Neural Networks, IEEE Transactions on*, v.12, n.3, p.503 –515.

Pedrycz, W.; Vukovich, G. (2004). Fuzzy clustering with supervision. *Pattern Recognition*, v.37, n.7, p.1339 – 1349.

Perrone, M. P.; Cooper, L. N. (1993). When networks disagree: Ensemble methods for hybrid neural networks. p. 126–142. Chapman and Hall.

Polikar, R. (2009). Ensemble learning. *Scholarpedia*, v.4, n.1, p.2776.

Pozzi, F.; Maccagnola, D.; Fersini, E.; Messina, E. (2013). Enhance user-level sentiment analysis on microblogs with approval relations. Baldoni, M.; Baroglio, C.; Boella, G.; Micalizio, R., editores, *AI*IA 2013: Advances in Artificial Intelligence*, v. 8249 de *Lecture Notes in Computer Science*, p. 133–144. Springer International Publishing.

Price, K. (1996). Differential evolution: a fast and simple numerical optimizer. *Fuzzy Information Processing Society, 1996. NAFIPS. 1996 Biennial Conference of the North American*, p. 524–527. IEEE.

Proisl, T.; Greiner, P.; Evert, S.; Kabashi, B. (2013). Klue: simple and robust methods for polarity classification. *Seventh International Workshop on Semantic Evaluation (SemEval 2013*, v. 2, p. 395–401. Association for Computational Linguistics.

Punera, K.; Ghosh, J. (2008). Consensus-based ensembles of soft clusterings. *Appl. Artif. Intell.*, v.22, p.780–810.

Qian, Q.; Chen, S.; Cai, W. (2012). Simultaneous clustering and classification over cluster structure representation. *Pattern Recognition*, v.45, n.6, p.2227 – 2236.

Qiu, L.; Zhang, W.; Hu, C.; Zhao, K. (2009). Selc: A self-supervised model for sentiment classification. *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, p. 929–936, New York, NY, USA. ACM.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Rahman, A.; Verma, B. (2013). Cluster-based ensemble of classifiers. *Expert Systems*, v.30, n.3, p.270–282.

Ranzato, M.; Hinton, G. E.; LeCun, Y. (2015). Guest editorial: Deep learning. *International Journal of Computer Vision*, v.113, n.1, p.1–2.

Rodrıguez-Penagos, C.; Atserias, J.; Codina-Filba, J.; Garcıa-Narbona, D.; Grivolla, J.; Lambert, P.; Saurı, R. (2013). Fbm: Combining lexicon-based ml and heuristics for social media polarities. *Proceedings of SemEval-2013 – International Workshop on Semantic Evaluation Co-located with *Sem and NAACL*, Atlanta, Georgia. Url date at 2013-10-10.

Rokach, L. (2010). *Pattern Classification Using Ensemble Methods*. World Scientific Publishing Co., Inc., River Edge, NJ, USA.

Rosenthal, S.; Nakov, P.; Ritter, A.; Stoyanov, V. (2014). SemEval-2014 Task 9: Sentiment Analysis in Twitter. Nakov, P.; Zesch, T., editores, *Proceedings of the 8th International Workshop on Semantic Evaluation*, SemEval 14, Dublin, Ireland.

Saif, H.; Fernandez, M.; He, Y.; Alani, H. (2014). On stopwords, filtering and data sparsity for sentiment analysis of twitter. Chair), N. C. C.; Choukri, K.; Declerck, T.; Loftsson, H.; Maegaard, B.; Mariani, J.; Moreno, A.; Odijk, J.; Piperidis, S., editores, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Saif, H.; He, Y.; Alani, H. (2012a). Alleviating data sparsity for twitter sentiment analysis. *Workshop of Making Sense of Microposts co-located with WWW 2012*.

Saif, H.; He, Y.; Alani, H. (2012b). Semantic sentiment analysis of twitter. *Proceedings of the 11th international conference on The Semantic Web - Volume Part I*, ISWC'12, p. 508–524, Berlin, Heidelberg. Springer-Verlag.

Schapire, R. E.; Freund, Y. (2012). *Boosting: Foundations and Algorithms*. The MIT Press.

Scudder, H. J. (1965). Probability of error of some adaptive pattern-recognition machines. *Information Theory, IEEE Transactions on*, v.11, n.3, p.363–371.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Comput. Surv.*, v.34, n.1, p.1–47.

Setnes, M.; Babuska, R. (1999). Fuzzy relational classifier trained by fuzzy clustering. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, v.29, n.5, p.619 –625.

Settles, B. (2012). Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, v.6, n.1, p.1–114.

Shi, Q.; Petterson, J.; Dror, G.; Langford, J.; Smola, A.; Vishwanathan, S. (2009). I. *J. Mach. Learn. Res.*, v.10, p.2615–2637.

Si, J.; Mukherjee, A.; Liu, B.; Li, Q.; Li, H.; Deng, X. (2013). Exploiting topic based twitter sentiment for stock prediction. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 2: Short Papers*, p. 24–29.

Sindhwani, V.; Keerthi, S. S. (2006). Large scale semi-supervised linear svms. *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, p. 477–484. ACM.

Smailović, J.; Grčar, M.; Lavrač, N.; Žnidaršič, M. (2014). Stream-based active learning for sentiment analysis in the financial domain. *Information Sciences*, v.285, p.181–203.

Soares, R. G. F.; Chen, H.; Yao, X. (2012). Semisupervised classification with cluster regularization. *Neural Networks and Learning Systems, IEEE Transactions on*, v.23, n.11, p.1779–1792.

Speriosu, M.; Sudan, N.; Upadhyay, S.; Baldridge, J. (2011). Twitter polarity classification with label propagation over lexical links and the follower graph. *Proceedings of the First Workshop on Unsupervised Learning in NLP*, EMNLP '11, p. 53–63, Stroudsburg, PA, USA. Association for Computational Linguistics.

Storn, R.; Price, K. (1997). Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces. *Jornal of Global Optimization*, v.11, n.4, p.341–359.

Svetlana Kiritchenko, X. Z.; Mohammad, S. M. (2014). Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research (JAIR)*, v.50, p.723–762.

Taboada, M.; Brooke, J.; Tofiloski, M.; Voll, K.; Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Comput. Linguist.*, v.37, n.2, p.267–307.

Tan, A. C.; Gilbert, D.; David, T.; Deville, Y. (2003). Multi-class protein fold classification using a new ensemble machine learning approach.

Tan, C.; Lee, L.; Tang, J.; Jiang, L.; Zhou, M.; Li, P. (2011). User-level sentiment analysis incorporating social networks. *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, p. 1397–1405, New York, NY, USA. ACM.

Tang, D. (2015). Sentiment-specific representation learning for document-level sentiment analysis. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, p. 447–452. ACM.

Tang, D.; Qin, B.; Liu, T. (2015a). Learning semantic representations of users and products for document level sentiment classification. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, p. 1014–1023, Beijing, China. Association for Computational Linguistics.

Tang, D.; Qin, B.; Liu, T.; Yang, Y. (2015b). User modeling with neural network for review rating prediction. *Proc. IJCAI*, p. 1340–1346.

Teh, Y.; Jordan, M.; Beal, M.; Blei, D. (2006). Hierarchical dirichlet processes. *Journal of the American Statistical Association*, v.101, n.476, p.1566–1581.

Thelwall, M. (2010). Emotion homophily in social network site messages. *First Monday*, v.15, n.4.

Thelwall, M.; Buckley, K.; Paltoglou, G.; Cai, D.; Kappas, A. (2010). Sentiment in short strength detection informal text. *J. Am. Soc. Inf. Sci. Technol.*, v.61, n.12, p.2544–2558.

Timm, H. (2001). Fuzzy cluster analysis of classified data. *IFSA World Congress and 20th NAFIPS International Conference, 2001. Joint 9th*, v. 3, p. 1431 –1436 vol.3.

Tong, S.; Koller, D. (2001). Support vector machine active learning with applications to text classification. *JOURNAL OF MACHINE LEARNING RESEARCH*, p. 45–66.

Tsytsarau, M.; Palpanas, T. (2012). Survey on mining subjective data on the web. *Data Min. Knowl. Discov.*, v.24, n.3, p.478–514.

Tumer, K.; Ghosh, J. (1996). Error correlation and error reduction in ensemble classifiers. *Connection Science*, v.8, n.3-4, p.385–403.

Tumer, K.; Ghosh, J. (2003). Bayes error rate estimation using classifier ensembles. *International Journal of Smart Engineering System Design*, v.5, n.2, p.95–110.

Turney, P. D. (2002). Thumbs up or thumbs down: semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL 02, p. 417–424, Stroudsburg, PA, USA. Association for Computational Linguistics.

Turney, P. D.; Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Trans. Inf. Syst.*, v.21, n.4, p.315–346.

Twitter, C. (2014). About twitter. https://about.twitter.com/company. [Online; accessed 9-April-2015].

Vanin, A. A.; Freitas, L. A.; Vieira, R.; Bochernitsan, M. (2013). Some clues on irony detection in tweets. *Proceedings of the 22Nd International Conference on World Wide Web Companion*, WWW '13 Companion, p. 635–636, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Verma, B.; Rahman, A. (2012). Cluster-oriented ensemble classifier: Impact of multicluster characterization on ensemble classifier learning. *Knowledge and Data Engineering, IEEE Transactions on*, v.24, n.4, p.605–618.

Wan, X. (2009). Co-training for cross-lingual sentiment classification. *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, p. 235–243, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wang, W.; Zhou, Z.-H. (2010). A new analysis of co-training. *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, p. 1135–1142.

Wang, X.; Wei, F.; Liu, X.; Zhou, M.; Zhang, M. (2011). Topic sentiment analysis in twitter: A graph-based hashtag sentiment classification approach. *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, p. 1031–1040, New York, NY, USA. ACM.

Weinberger, K. Q.; Dasgupta, A.; Langford, J.; Smola, A. J.; Attenberg, J. (2009). Feature hashing for large scale multitask learning. Danyluk, A. P.; Bottou, L.; Littman, M. L., editores, *ICML*, v. 382 de *ACM International Conference Proceeding Series*, p. 140. ACM.

Wiebe, J.; Wilson, T.; Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, v.1, n.2, p.0.

Wilson, T.; Wiebe, J.; Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, p. 347–354, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wilson, T.; Wiebe, J.; Hwa, R. (2004). Just how mad are you? finding strong and weak opinion clauses. *Proceedings of the 19th National Conference on Artifical Intelligence*, AAAI'04, p. 761–767. AAAI Press.

Windeatt, T. (2005). Diversity measures for multiple classifier system analysis and design. *Information Fusion*, v.6, n.1, p.21 – 36. Diversity in Multiple Classifier Systems.

Witten, I. H.; Frank, E. (2005). *Data mining: practical machine learning tools and techniques*. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, 2nd Edition.

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, v.5, p.241–259.

Xiang, B.; Zhou, L. (2014). Improving twitter sentiment analysis with topic-based mixture modeling and semi-supervised training. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, p. 434–439. Association for Computational Linguistics.

Xu, Z.; Jin, R.; Zhu, J.; King, I.; Lyu, M. (2008). Efficient convex relaxation for transductive support vector machine. *Advances in neural information processing systems*, p. 1641–1648.

Yang, M.; Tu, W.; Lu, Z.; Yin, W.; Chow, K.-P. (2015). Lcct: A semi-supervised model for sentiment classification. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, p. 546–555, Denver, Colorado. Association for Computational Linguistics.

Yong, Q.; Zong-yi, X.; Li-min, J.; Ying-ying, W. (2009). Study on interpretable fuzzy classification system based on neural networks. *ICCAS-SICE, 2009*, p. 5318 –5321.

Yu, N. (2013). Domain adaptation for opinion classification: A self- training approach. *Journal of Information Science Theory and Practice*, v.1, n.1.

Yu, N. (2014). Exploring co-training strategies for opinion detection. *Journal of the Association for Information Science and Technology*, p. n/a–n/a.

Zhang, C.; Ma, Y. (2012). *Ensemble Machine Learning: Methods and Applications*. Springer.

Zhang, S.; Neagu, D.; Balescu, C. (2005). Refinement of clustering solutions using a multi-label voting algorithm for neuro-fuzzy ensembles. Wang, L.; Chen, K.; Ong, Y., editores, *Advances in Natural Computation*, v. 3612 de *Lecture Notes in Computer Science*, p. 1300–1303. Springer Berlin Heidelberg.

Zhang, X.; LeCun, Y. (2015). Text understanding from scratch. *CoRR*, v.abs/1502.01710.

Zhao, J.; Lan, M.; Zhu, T. (2014). Ecnu: Expression- and message-level sentiment orientation classification in twitter using multiple effective features. *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, p. 259–264, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Zhou, D.; Bousquet, O.; Lal, T. N.; Weston, J.; Schölkopf, B. (2004). Learning with local and global consistency. *Advances in neural information processing systems*, v.16, n.16, p.321–328.

Zhou, Z. (2012). *Ensemble Methods: Foundations and Algorithms*. Chapman & Hall/CRC Data Mining and Knowledge Discovery Serie. Taylor & Francis.

Zhou, Z.-H.; Li, M. (2005). Tri-training: Exploiting unlabeled data using three classifiers. *Knowledge and Data Engineering, IEEE Transactions on*, v.17, n.11, p.1529–1541.

Zhu, X. (2005). *Semi-supervised Learning with Graphs*. Tese (Doutorado), Pittsburgh, PA, USA. AAI3179046.

Zhu, X. (2008). Semi-supervised learning literature survey.

Zhu, X.; Ghahramani, Z.; Lafferty, J.; others (2003a). Semi-supervised learning using gaussian fields and harmonic functions. *ICML*, v. 3, p. 912–919.

Zhu, X.; Goldberg, A. B. (2009). *Introduction to Semi-Supervised Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan and Claypool Publishers.

Zhu, X.; Kiritchenko, S.; Mohammad, S. M. (2014). Nrc-canada-2014: Recent improvements in the sentiment analysis of tweets. *SemEval 2014*, p. 443.

Zhu, X.; Lafferty, J.; Ghahramani, Z. (2003b). Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. *ICML 2003 workshop on The Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining*, p. 58–65.

Ziegelmayer, D.; Schrader, R. (2012). Sentiment polarity classification using statistical data compression models. *ICDM Workshops*, p. 731–738.

Zimmermann, M.; Ntoutsi, E.; Spiliopoulou, M. (2014). Adaptive semi supervised opinion classifier with forgetting mechanism. *Proceedings of the 29th Annual ACM Symposium on Applied Computing*, SAC '14, p. 805–812, New York, NY, USA. ACM.