UNIVERSIDADE DO VALE DO RIO DOS SINOS — UNISINOS
UNIDADE ACADÊMICA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA
NÍVEL MESTRADO

RODRIGO DE MORAES

UMA INVESTIGAÇÃO EMPÍRICA E COMPARATIVA DA APLICAÇÃO DE RNAS AO PROBLEMA DE MINERAÇÃO DE OPINIÕES E ANÁLISE DE SENTIMENTOS

SÃO LEOPOLDO
2013

Rodrigo de Moraes

UMA INVESTIGAÇÃO EMPÍRICA E COMPARATIVA DA APLICAÇÃO DE RNAS AO PROBLEMA DE MINERAÇÃO DE OPINIÕES E ANÁLISE DE SENTIMENTOS

Dissertação apresentada como requisito parcial
para a obtenção do título de Mestre pelo
Programa de Pós-Graduação em Computação
Aplicada da Universidade do Vale do Rio dos
Sinos — UNISINOS

Orientador:
Prof. Dr. João F. Valiati

São Leopoldo
2013

DADOS INTERNACIONAIS DE CATALOGAÇÃO NA PUBLICAÇÃO (CIP)

Moraes, Rodrigo de

Uma investigação empírica e comparativa da aplicação de RNAs ao problema de Mineração de Opiniões e Análise de Sentimentos / Rodrigo de Moraes — 2013.

113 f.: il.; 30 cm.

Dissertação (mestrado) — Universidade do Vale do Rio dos Sinos, Programa de Pós-Graduação em Computação Aplicada, São Leopoldo, 2013.

"Orientador: Prof. Dr. João F. Valiati, Unidade Acadêmica de Pesquisa e Pós-Graduação".

1. Análise de Sentimentos. 2. Mineração de Opiniões. 3. Aprendizado de Máquina. 4. Classificadores. 5. Redes Neurais Artificiais. 6. *Support Vector Machines*. 7. *Naïve Bayes*. I. Título.

CDU 004

Bibliotecário responsável: Flávio Nunes — CRB 10/1298

AGRADECIMENTOS

Para a realização deste trabalho alguns aspectos foram fundamentais. Dentre eles, algumas pessoas com quem convivi e que, de diferentes maneiras, acabaram sendo essenciais. Sendo assim, presto meus mais sinceros agradecimentos:

Ao professor e amigo Dr. João F. Valiati, orientador deste trabalho, pelos seus conhecimentos, atenção, dedicação e orientações, que foram, e sempre serão, fundamentais para o meu crescimento pessoal e acadêmico;

Aos demais professores do corpo docente do Programa Interdisciplinar de Pós-Graduação em Computação Aplicada da UNISINOS que através de aulas ou conversas informais contribuíram para a aquisição de conhecimento, desenvolvimento de senso crítico e capacidade de abstração;

Ao professor Dr. Wilson P. Gavião Neto, colega de projeto de pesquisa e amigo, pela constante instigação de questões técnicas e conceituais e contribuição para o meu crescimento acadêmico e profissional.

Aos meus colegas de curso que, diretamente ou indiretamente, contribuíram para a realização deste trabalho com apoio, compartilhamento de conhecimento e companheirismo;

A minha namorada, Clara C. Scherer, com quem pude contar com apoio e compreensão, sem restrições, fundamentais em diversas situações, proporcionando-me motivação e tranquilidade para me manter focado na realização deste trabalho.

E aos meus pais que, desde sempre, foram essenciais na minha formação em todos os sentidos e que, com muito esforço reconhecido, me proporcionando um ambiente e condições favoráveis para a realização deste trabalho, considerando tanto questões financeiras quanto familiares.

"*O mal de quase todos nós é que preferimos ser arruinados pelo elogio a ser salvos pela crítica.*".
(Norman Vincent)

RESUMO

A área de Mineração de Opiniões e Análise de Sentimentos surgiu da necessidade de processamento automatizado de informações textuais referentes a opiniões postadas na web. Como principal motivação está o constante crescimento do volume desse tipo de informação, proporcionado pelas tecnologia trazidas pela Web 2.0, que torna inviável o acompanhamento e análise dessas opiniões úteis tanto para usuários com pretensão de compra de novos produtos quanto para empresas para a identificação de demanda de mercado. Atualmente, a maioria dos estudos em Mineração de Opiniões e Análise de Sentimentos que fazem o uso de mineração de dados se voltam para o desenvolvimentos de técnicas que procuram uma melhor representação do conhecimento e acabam utilizando técnicas de classificação comumente aplicadas, não explorando outras que apresentam bons resultados em outros problemas. Sendo assim, este trabalho tem como objetivo uma investigação empírica e comparativa da aplicação do modelo clássico de Redes Neurais Artificiais (RNAs), o *multilayer perceptron*, no problema de Mineração de Opiniões e Análise de Sentimentos. Para isso, bases de dados de opiniões são definidas e técnicas de representação de conhecimento textual são aplicadas sobre essas objetivando uma igual representação dos textos para os classificadores através de unigramas. A partir dessa representação, os classificadores *Support Vector Machines* (SVM), *Naïve Bayes* (NB) e RNAs são aplicados considerandos três diferentes contextos de base de dados: (i) bases de dados balanceadas, (ii) bases com diferentes níveis de desbalanceamento e (iii) bases em que a técnica para o tratamento do desbalanceamento *undersampling* randômico é aplicada. A investigação do contexto desbalanceado e de outros originados dele se mostra relevante uma vez que bases de opiniões disponíveis na web normalmente apresentam mais opiniões positivas do que negativas. Para a avaliação dos classificadores são utilizadas métricas tanto para a mensuração de desempenho de classificação quanto para a de tempo de execução. Os resultados obtidos sobre o contexto balanceado indicam que as RNAs conseguem superar significativamente os resultados dos demais classificadores e, apesar de apresentarem um grande custo computacional para treinamento, proporcionam tempos de classificação significantemente inferiores aos do classificador que apresentou os resultados de classificação mais próximos aos dos resultados das RNAs. Já para o contexto desbalanceado, as RNAs se mostram sensíveis ao aumento de ruído na representação dos dados e ao aumento do desbalanceamento, se destacando nestes experimentos, o classificador NB. Com a aplicação de *undersampling* as RNAs conseguem ser equivalentes aos demais classificadores apresentando resultados competitivos. Porém, podem não ser o classificador mais adequado de se adotar nesse contexto quando considerados os tempos de treinamento e classificação, e também a diferença pouco expressiva de acerto de classificação.

Palavras-chave: Análise de Sentimentos. Mineração de Opiniões. Aprendizado de Máquina. Classificadores. Redes Neurais Artificiais. *Support Vector Machines*. *Naïve Bayes*.

ABSTRACT

The area of Opinion Mining and Sentiment Analysis emerges from the need for automated processing of textual information about reviews posted in the web. The main motivation of this area is the constant volume growth of such information, provided by the technologies brought by Web 2.0, that makes impossible the monitoring and analysis of these reviews that are useful for users, who desire to purchase new products, and for companies to identify market demand as well. Currently, the most studies of Opinion Mining and Sentiment Analysis that make use of data mining aims to the development of techniques that seek a better knowledge representation and using classification techniques commonly applied and they not explore others classifiers that work well in other problems. Thus, this work aims a comparative empirical research of the application of the classical model of Artificial Neural Networks (ANN), the multilayer perceptron, in the Opinion Mining and Sentiment Analysis problem. For this, reviews datasets are defined and techniques for textual knowledge representation applied to these aiming an equal texts representation for the classifiers. From this representation, the classifiers Support Vector Machines (SVM), Naïve Bayes (NB) and ANN are applied considering three data context: (i) balanced datasets, (ii) datasets with different unbalanced ratio and (iii) datasets with the application of random undersampling technique for the unbalanced handling. The unbalanced context investigation and of others originated from it becomes relevant once datasets available in the web ordinarily contain more positive opinions than negative. For the classifiers evaluation, metrics both for the classification perform and for run time are used. The results obtained in the balanced context indicate that ANN outperformed significantly the others classifiers and, although it has a large computation cost for the training fase, the ANN classifier provides classifcation time (real-time) significantly less than the classifier that obtained the results closer than ANN. For the unbalanced context, the ANN are sensitive to the growth of noise representation and the unbalanced growth while the NB classifier stood out. With the undersampling application, the ANN classifier is equivalent to the others classifiers attaining competitive results. However, it can not be the most appropriate classifier to this context when the training and classification time and its little advantage of classification accuracy are considered.

Keywords: Sentiment Analysis. Opinion Mining. Machine Learning. Classifiers. Artificial Neural Networks. Support Vector Machines. Naïve Bayes.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

ANN	*Artificial Neural Networks*
BOW	*Bag-of-Words*
DCBD	Descoberta do Conhecimento em Base de Dados
IG	*Information Gain*
IR	*Information Retrieval*
JST	*Joint Sentiment-Topic*
KDD	*Knowledge Discovery in Databases*
LDA	*Latent Dirichlet Allocation*
LM	*Language Model*
MD	Mineração de Dados
ME	*Maximum Entropy*
ML	*Machine Learning*
MLP	*Multilayer Perceptron*
NB	*Naïve Bayes*
OS	Orientação Semântica
PLN	Processamento de Linguagem Natural
POS	*Part-of-Speech*
RNAs	Redes Neurais Artificiais
SGBD	Sistema de Gerência de Banco de Dados
SVM	*Support Vector Machines*
SVs	*Support Vectors*
TF-IDF	*Term Frequency - Inverse Document Frequency*
TMG	*Text to Matrix Generation*
WEKA	*Waikato Environment for Knowladge Analysis*

SUMÁRIO

1 INTRODUÇÃO

Nos dias de hoje está se tornando cada vez mais comum a compra de produtos através de web sites do tipo *e-commerce*. Devido à distância física dos consumidores aos produtos e a ausência de um atendente, um recurso que está sendo cada vez mais utilizado é a consulta de opiniões de consumidores que já adquiriram o produto desejado. *E-commerce*, como Amazon.com, chegam a ter mais de 1500 opiniões para um mesmo título de livro ou modelo de mp3 *player*. Além disso, web sites como cnet.com, viewpoints.com e gsmarena.com, se dedicam a avaliação de produtos e disponibilizam espaço para os usuários/consumidores postarem suas opiniões e experiências com produtos, armazenando mais de 25000 registros opinativos sobre um mesmo produto. Outros, como o reclameaqui.com.br, que acabam sendo muito úteis para consumidores que não têm suas reclamações atendidas quando feitas diretamente às empresas, servem como base de consulta à reputação de empresas, serviços e produtos.

O grande volume de informações opinativas mencionado evidencia o interesse dos usuários, tanto em compartilhar seus pontos de vista, como em procurar opiniões previamente a uma compra, em forma textual e não somente uma avaliação direta em forma de nota. Tais textos possivelmente exercem uma forte influência na decisão de compra dos consumidores.

Além de serem de grande valor para consumidores, esses textos também são uma fonte importante para empresas fabricantes dos produtos ou prestadoras de serviços. Com a análise desses comentários postados na web, essas empresas podem identificar falhas e oportunidades relacionadas ao seu produto e, com isso, atender melhor seus clientes e até mesmo expandir sua gama de produtos.

Porém, o grande volume de opiniões e o número elevado de postagens constante por parte dos usuários torna árdua a tarefa de acompanhamento e identificação de aspectos importantes para a tomada de decisão de empresas e avaliação dos futuros consumidores. Sendo assim, uma área de estudos que vem recebendo grande atenção nos últimos tempos é a de Mineração de Opiniões e Análise de Sentimentos. Nessa área busca-se, através da aplicação de técnicas de Inteligência Artificial, a identificação automatizada da polaridade de opiniões (PANG; LEE, 2008).

Em Mineração de Opiniões e Análise de Sentimentos o principal objetivo é a identificação de padrões ou principais características que conseguem distinguir textos com polaridades diferentes. Com a evolução dos estudos, outras frentes de pesquisa foram surgindo dentro da Mineração de Opiniões e Análise de Sentimentos, como o aperfeiçoamento de técnicas voltadas para a seleção dos melhores termos para a representação do conhecimento (ABBASI et al., 2011) e a identificação de diferentes tópicos (YU et al., 2011) ou características de produtos ou serviços, mencionados dentro das opiniões.

1.1 Motivação

Apesar da área de Mineração de Opiniões e Análise de Sentimentos apresentar recentes evoluções, alguns aspectos ainda não são muito explorados. Para a identificação de polaridade de opinião pode-se fazer o uso de técnicas de aprendizado de máquina (*Machine Learning* - ML) que identificam padrões em dados de treinamento (BISHOP, 2007). Porém, estudos que fazem o uso de técnicas de ML, mais especificamente classificadores, normalmente, focam-se no desenvolvimento de técnicas que objetivam a seleção de características representativas das opiniões e acabam aplicando classificadores comumente utilizados em Mineração de Opiniões e Análise de Sentimentos, o *Support Vector Machines* (SVM) e o *Naïve Bayes* (NB). Desta forma, classificadores que apresentam bom desempenho em outros problemas, como Redes Neurais Artificias (RNAs), acabam sendo raramente investigados e aplicados em Mineração de Opiniões e Análise de Sentimentos.

Embora alguns estudos realizem a aplicação de RNAs em Mineração de Opiniões e Análise de Sentimentos (CHEN; LIU; CHIU, 2011; BESPALOV et al., 2011), esses não avaliam seu desempenho frente às técnicas de ML já aplicadas na área, SVM e NB, para a investigação de seu potencial de aplicabilidade ao problema em questão. Além disso, a não utilização de bases clássicas da literatura nesses estudos também inviabiliza a comparação com outros trabalhos.

Outra característica da maioria dos estudos da área é que as técnicas de ML são avaliadas ou comparadas considerando bases de dados de treinamento balanceadas. Porém, como evidenciado por Blitzer, Dredze e Pereira (2007) as bases de opiniões presentes na web são desbalanceadas, ou seja, bases que apresentam uma significante diferença entre as quantidades de registros pertencentes a diferentes classes. Tal característica pode influenciar consideravelmente a aplicação de técnicas de ML, já que essas técnicas utilizam recursos estatísticos para a identificação de diferentes padrões, e requerer a utilização de métricas específicas para avaliação dos classificadores.

1.2 Objetivo

Considerando o contexto e a motivação apresentados anteriormente, este trabalho tem como objetivo geral a realização de uma investigação empírica e comparativa entre o modelo clássico de RNAs, *Multilayer Perceptron*, e os modelos clássicos da literatura, SVM e NB, aplicados ao problema de Mineração de Opiniões e Análise de Sentimentos considerando os contextos balanceado e desbalanceado de bases de dados.

Para isso, define-se como objetivos específicos do trabalho as seguintes tarefas:

- Realização de experimentos com a base de dados clássica da literatura consolidada por Pang e Lee (2004) referente a filmes;
- Aplicação da abordagem proposta sobre outras 3 bases referenciadas na literatura rela-

cionadas a *e-commerce*;

- Utilização das métricas: acurácia; *precision* e *recall* por classe; tempo de treinamento; e tempo de teste para a avaliação dos classificadores;
- Uso da técnicas *undersampling* para a comparação com experimentos referentes ao contexto desbalanceado;
- Produção de uma análise crítica dos resultados obtidos apontando indícios de situações em que RNAs se mostram competitivas quando aplicadas a dados textuais opiniativos da web.

1.3 Organização do Trabalho

Este trabalho apresenta no capítulo 2 os conceitos de Mineração Textual e Análise de Sentimentos, assim como técnicas para a descoberta automatizada de conhecimentos em bases de dados textuais. O capítulo 3 relata alguns trabalhos da área de Mineração de Opiniões e Análise de Sentimentos que se relacionam com este, sendo dividido em seções objetivando a análise dos diferentes aspectos. No capítulo 4, a metodologia de avaliação dos classificadores e as configurações das experimentações são apresentados. Além disso, as seções finais do capítulo 4 são dedicadas ao relato dos resultado obtidos, apontamento de características relevantes observadas e a uma discussão considerando conceitos relacionados, técnicas aplicadas em todo o processo de avaliação e resultados. Por fim, no capítulo 5, são apresentadas as conclusões proporcionadas pelo trabalho e investigação realizados, juntamente com as contribuições do trabalho e tópicos para a extensão do estudo.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo tem como objetivo descrever os principais conceitos e técnicas que contribuem para o objetivo deste trabalho. Sendo assim, as seções 2.1 e 2.2 contêm os conceitos da área de Mineração Textual e de Mineração de Opiniões e Análise de Sentimentos, respectivamente.

Já a seção 2.3 descreve o modelo clássico do processo de Descoberta do Conhecimento em Base de Dados (DCBD) que é utilizado como base para a realização dos experimentos e as seções seguintes apresentam técnicas específicas de cada uma das etapas desse processo. Na seção, 2.4, são apresentadas técnicas utilizadas para o pré-processamento textual, tarefa em que é feita a limpeza das bases de dados através da remoção de informações não relevantes para a análise dos dados.

Na seção 2.5 são descritas técnicas utilizadas para adaptar os dados para a aplicação das técnicas de mineração de dados. Essas, por sua vez, são vistas na seção 2.6, que apresenta técnicas voltadas à classificação de dados. Finalizando as etapas do DCBD, na seção 2.7 são apresentadas as metodologias e métricas de avaliação dos resultados dos classificadores.

Por fim, na seção 2.8 é descrito o conceito de bases de dados desbalanceadas e implicações desse tipo de base para a classificação de dados. Além disso, algumas técnicas possíveis de serem aplicadas para o tratamento desse problema são brevemente comentadas.

2.1 Mineração Textual

A Mineração de Textos tem o objetivo de processar uma grande quantidade de dados textuais para que sejam extraídas informações relevantes e para a obtenção de conhecimento. Os resultados gerados pelas técnicas que permitem realizar essa extração podem auxiliar processos de tomada de decisão.

Sendo a área de mineração de textos relativamente nova, não há um conceito único que a define existindo assim várias definições encontradas na literatura:

- processo de extração de padrões interessantes e não-triviais, a partir de documentos textuais (TAN, 1999);
- a descoberta de informações desconhecidas ou novas, através da utilização de ferramentas de extração automática de informação, a partir de documentos de textos não-estruturados (HEARST, 1999);
- o estudo de extração de informações de textos usando os princípios da linguística computacional (SULLIVAN, 2000);

Embora esse tipo de mineração possa ser confundido com Mineração de Dados (MD), a mineração de textos tem fundamentação à extração de informações em documentos em linguagem natural, ou seja, documentos que possuem pouca ou nenhuma estrutura nos dados. Sendo

assim, conhecimentos em base textual, também denominada de *corpus*, necessitam de mecanismos mais elaborados para sua extração. Já em MD a extração ocorre a partir de bases de dados formalmente estruturadas, geralmente armazenadas em Sistemas de Gerência de Banco de Dados (SGBD) (FELDMAN; SANGER, 2007).

2.2 Mineração de Opiniões e Análise de Sentimentos

A área de Mineração de Opiniões e Análise de Sentimentos tem como principal objetivo a mineração de dados textuais para a identificação de polaridade de opinião, ou seja, identificar em um grande volume de dados textuais se o que está sendo dito sobre determinado produto, marca ou fato é expresso positivamente ou negativamente pelos autores dos textos.

Esse tipo de análise se mostra importante tanto para empresas como para consumidores que buscam informações relativas a produtos antes de adquiri-los. Para empresas, ter conhecimento sobre a experiência dos consumidores sobre seus produtos/serviços é de valor estratégico. Para consumidores que pretendem adquirir novos produtos/serviços, a opinião de consumidores que já os adquiriram pode ser útil na decisão da melhor compra.

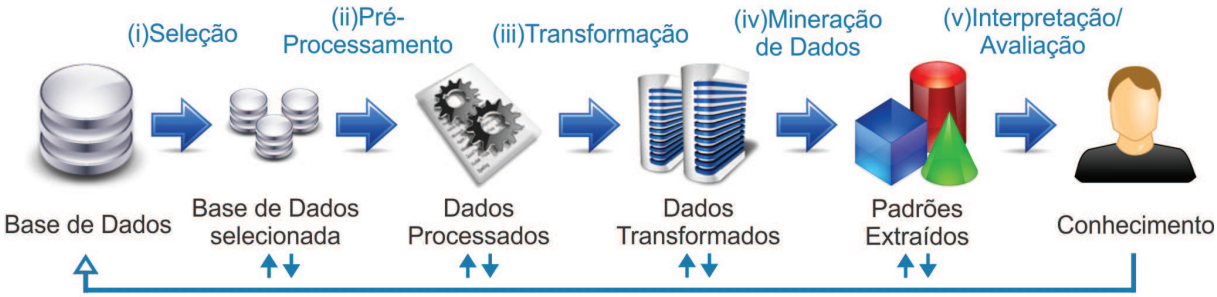
Esta recente área de pesquisa é discutida em detalhes no livro (PANG; LEE, 2008) e recentemente revista no livro (LIU, 2012). Apesar de se tratar de um problema em que documentos textuais são utilizados, algumas particularidades, como oposição de ideias e negação, fazem com que técnicas de mineração utilizadas em mineração textual para categorização de documentos nem sempre apresentem o mesmo desempenho quando aplicadas em Mineração de Opiniões e Análise de Sentimentos (PANG; LEE; VAITHYANATHAN, 2002).

Atualmente os estudos da área focam no aperfeiçoamento de técnicas de pré-processamento objetivando a seleção de termos/expressões que melhor conseguem representar o conhecimento para a distinção de textos opinativos positivos de negativos (CHEN; LIU; CHIU, 2011; ABBASI et al., 2011). Outra linha de estudos recentemente explorada dentro da área é a investigação de tópicos mencionados dentro das opiniões. Como será visto no capítulo 3, a quebra das opiniões e identificação de diferentes tópicos mencionados pode refletir em um aumento do acerto de classificação quanto ao sentido de opinião.

2.3 Processo de Descoberta do Conhecimento

O processo de Descoberta do Conhecimento em Base de Dados (DCBD), em inglês *Knowledge Discovery in Databases* (KDD), envolve um conjunto de atividades que compartilham o conhecimento a partir de uma base de dados (DUA; DU, 2011). Segundo Fayyad et al. (1996), tal processo pode ser dividido em cinco etapas: (i) Seleção, (ii) Pré-Processamento, (iii) Transformação, (iv) Mineração de Dados e (v) Interpretação/Avaliação, como pode ser visto na Figura 1.

Figura 1: Processo de Descoberta do Conhecimento.



A etapa de (i) Seleção dos dados objetiva a seleção de uma base de dados de onde será extraído o conhecimento. Nesta etapa normalmente são utilizados critérios de seleção mais gerais, como a quantidade e a natureza dos dados.

O (ii) Pré-Processamento tem como principal objetivo a filtragem e limpeza dos dados, eliminando redundâncias e informações desnecessárias para o conhecimento que se deseja extrair e preparando, no caso deste estudo, os textos para a MD (GONCALVES et al., 2006). Esta etapa é considera a mais custosa no processo, já que não há uma só técnica, ou uma combinação delas, que quando aplicada, obtém uma representação satisfatória em todos os domínios (HAN; KAMBER; PEI, 2006). As principais técnicas aplicadas nesta etapa, considerando bases de dados textuais, são apresentadas na seção 2.4.

Após o Pré-Processamento, a próxima etapa do processo é a de (iii) Transformação dos dados, onde a formatação dos dados é adequada para atender aos requisitos da atividade (iv) Mineração de Dados. Para problemas de mineração textual, o principal objetivo desta etapa é construir uma representação numérica dos textos.

Na etapa de (iv) Mineração de Dados são aplicadas as técnicas de mineração voltadas ao aprendizado de máquina (*Machine Learning* - ML) para a obtenção de novos conhecimentos (WITTEN; FRANK; HALL, 2011). Tais técnicas permitem a extração dos padrões relevantes que os dados apresentam para a consolidação da modelagem computacional. No contexto deste trabalho, essas técnicas devem ser capazes de identificar características que diferenciam documentos pertencentes a diferentes classes e realizar a classificação de documentos não classificados.

Por fim, a aplicação tanto das técnicas de mineração como sua combinação com as demais aplicadas em todo o processo são avaliadas através de métricas calculadas sobre os resultados da etapa de mineração, caracterizando a etapa (v) Validação.

Além disso, o processo de DCBD se caracteriza por ser cíclico. Ao final de cada uma das etapas, os resultados devem ser avaliados individualmente e caso não apresentem resultados satisfatórios ou os dados não estiverem devidamente refinados, deve-se realizar alterações no processo para a realização de um novo ciclo.

2.4 Técnicas de Pré-Processamento

Considerando que o problema abordado neste trabalho se refere a dados textuais, esta seção apresenta brevemente algumas técnicas que podem ser aplicadas na etapa de Pré-Processamento sobre dados textuais para o refinamento de informações e melhor representação do conhecimento.

Técnicas de Pré-Processamento se mostram bastante relevantes para problemas em que dados textuais estão envolvidos, já que esses, normalmente apresentam uma grande quantidade de termos, atributos, para sua representação. Este fato resulta em uma representação denominada esparsa, em que a representação de uma amostra apresenta a grande maioria dos atributos nula.

Desta forma, e considerando que a complexidade da maioria das técnicas de MD está relacionada com a quantidade de atributos, uma boa representação para o problema em questão é aquela que, além de conseguir identificar os atributos que melhor representam o conhecimento, consegue também reduzir consideravelmente a quantidade destes sem perder características importantes da base de dados.

2.4.1 Identificação de *Tokens* e colocações

A Identificação de *Tokens*, ou *Tokenização*, permite a separação de palavras ou termos contidos em um texto, o que é possível pelo fato das palavras contidas nele estarem separadas por espaços ou sinais de pontuação, como por exemplo: “ ”, “.”, “,”, “!” (MANNING; RAGHAVAN; SCHTZE, 2008).

Além disso, a aplicação da *Tokenização* também define a unidade de representação dos textos que será considerada. Para a modelagens dos textos, pode-se considerar a divisão desses em palavras únicas, denominadas de unigramas, ou então em expressões formadas por duas ou mais palavras, birgramas ou multi-gramas, respectivamente. Essa representação é denominada *Bag-of-Words* (BOW) e pode-se considerar mais de uma unidade, como por exemplo unigramas e bigramas, para um mesmo conjunto de dados textuais.

Desta forma, é possível o mapeamento de características potencialmente relevantes para a mineração de opiniões, como negação e oposição de ideias. Porém, a aplicação de modelagens que consideram mais de uma unidade de análise devem ser avaliadas, já que, essas acabam gerando uma quantidade grande de atributos a serem processados (BESPALOV et al., 2011).

2.4.2 Remoção de *Stopwords*

Outra técnica que pode ser aplicada na atividade de Pré-Processamento é a Remoção de *Stopwords* que é responsável pela remoção de palavras que não apresentam conteúdo semântico significativo no contexto, sendo irrelevantes para a análise (WEISS; INDURKHYA; ZHANG, 2004).

Desta maneira, uma lista de *stopwords* normalmente é composta por artigos, advérbios, conjunções e pronomes, como “a”, “o”, “ali”, “pelo” e outros. Porém, para se aplicar essa técnica, deve-se levar em consideração o que se deseja manter do texto original, pois palavras importantes para a aplicação podem ser consideradas *stopwords*, ou dependendo do contexto, palavras que normalmente não compõem uma lista de *stopwords* podem ser adicionadas a ela.

Exemplo disso são as palavras “filme” e “não”. A primeira pode ser incluída na lista, já que é um termo que pode aparecer frequentemente em opiniões referentes a filmes, tanto positivas como negativas não expressando assim qualquer relevância para a distinção das classes. Já a palavra “não”, que tipicamente faz parte de uma lista de *stopwords*, pode se tornar relevante por interferir no conteúdo semântico, por exemplo, dos adjetivos, invertendo o sentido da opinião.

2.4.3 Processamento de Linguagem Natural

A aplicação de técnicas de Processamento de Linguagem Natural (PLN) objetiva a construção de um sistema com melhor precisão e desempenho na aplicação das técnicas de mineração de dados, tais como: classificação, sumarização, agrupamento, etc (MANNING; SCHüTZE, 1999).

Uma das técnicas que podem ser aplicadas em sistemas de classificação automatizados é a utilização de um *tesauro* que contenha informações semânticas de palavras. Através dela é possível expandir a base de termos relevantes para uma análise.

Outra técnica é a aplicação de Algoritmos de *Stemming*, que permitem fazer a extração do radical de palavra. Segundo Russell e Norvig (2009), esses algoritmos de radicalização podem ser divididos em dois tipos: os baseados em regras, que realizam a radicalização geralmente de acordo com as terminações das palavras, e os baseados em dicionários, que realizam uma busca em uma lista para evitar que determinadas palavras que apresentam características específicas sofram alterações incorretas.

Para a língua inglesa, três são os algoritmos mais utilizados na literatura: Porter (PORTER, 1980), *Snowball* (PORTER, 2001) e *Lovins* (LOVINS, 1968).

A técnica de *stemming* apresenta boa eficiência na remoção de informações redundantes, já que consegue concentrar o mesmo significado semântico de diferentes palavras em um único *token* de representação e sua não aplicação pode acabar distorcendo a importância de alguns termos na etapa de transformação de dados.

Outra técnica de PLN aplicada em alguns estudos, como será visto no capítulo 3, é a denominada *Part-of-Speech* (POS). Essa técnica permite a classificação gramatical das palavras considerando as demais palavras próximas à ela. Com isso, é possível distinguir quando determinada palavra está sendo utilizada como, por exemplo, substantivo ou advérbio, ou então simplesmente uma melhor representação do conhecimento associando a classe identificada à representação dos textos. Além disso, pode-se realizar a seleção de palavras representativas dos documentos, selecionando aquelas pertencentes a somente uma ou algumas classes gramaticais.

2.4.4 Abordagens estatísticas de Seleção de Características

Além das técnicas apresentadas anteriormente, um outro tipo de técnica que objetiva a seleção de dados mais representativos eliminando ruídos e reduzindo a dimensionalidade[1] dos dados é a Seleção de Características através de abordagens estatísticas. Essas técnicas procuram calcular a importância dos atributos dos dados considerando sua frequência nos registros e podem também levar em conta o conhecimento supervisionado de dados de treinamento.

Uma dessas técnicas é o Ganho de Informação (*Information Gain* - IG) que, apesar de não ser a que apresenta os melhores resultados de classificação quando associada a classificadores, se mostra competitiva em estudos da área (YANG; PEDERSEN, 1997; FORMAN, 2003; LI et al., 2009; ABBASI et al., 2011; XIA; ZONG; LI, 2011).

O IG constrói uma qualificação dos atributos dos dados através da mensuração de um peso para cada um deles. Este peso se baseia na distribuição dos diferentes valores do atributo em cada uma das classes. Sendo assim, atributos que apresentam cada um de seus valores presentes em somente uma das classes são melhores classificados (LIU, 2011). Considerando a representação de um documento textual através da ausência e presença dos termos dentro dele, o peso definido para cada termo a_i pela técnica IG é definido como:

$$IG(a_i) = \sum_{k=1}^{|C|} P(c_k) \log_2 \frac{1}{P(c_k)} - \left[P(a_i = 1) \sum_{k=1}^{|C|} P(a_i = 1|c_k) \log_2 \frac{1}{P(a_i = 1|c_k)} + P(a_i = 0) \sum_{k=1}^{|C|} P(a_i = 0|c_k) \log_2 \frac{1}{P(a_i = 0|c_k)} \right] \tag{2.1}$$

onde $P(c_k)$ é a probabilidade de ocorrência de um documento na classe c_k, $|C|$ e a quantidade total de classes, $P(a_i = 1)$ é a probabilidade do termo a_i ocorrer em um documento, $P(a_i = 0)$ é a probabilidade do termo a_i **não** ocorrer em um documento, $P(c_k|a_i = 1)$ é a probabilidade do termo a_i ocorrer em um documento pertencente à classe c_k e $P(c_k|a_i = 0)$ é a probabilidade do termo a_i **não** ocorrer em um documento pertencente à classe c_k.

Além do IG, outras técnicas estatísticas como a *Chi-Square*, Informação Mútua (*Mutual Information*, *ODD Ratio*, *Weight ODD Ratio* e *Gain Ratio* também são aplicadas em estudos da área, como mostra os estudos de Mladenic e Grobelnik (1999) e Forman (2003).

2.5 Transformação de Dados

Para a extração dos padrões de dados, a maioria das técnicas de MD utiliza valores numéricos que representam os dados armazenados na base. Para essa extração em conteúdo textual, faz-se necessária a aplicação de alguma técnica que consiga representar cada elemento textual

[1] Redução de dimensionalidade: consiste na eliminação de atributos representativos de amostras, tendo em vista a diminuição da complexidade do problema.

em um número que represente algo significativo ao conhecimento contido na base.

A técnica mais comum que permite fazer essa representação é a de representação binária, na qual um conjunto de atributos é representado por um conjunto de números composto somente por zeros e uns, onde o número zero representa a ausência de um determinado atributo e o número um a presença. Outra técnica básica de aplicação que aprimora a representação binária é a que considera, ao invés de somente a presença das palavras, também sua frequência dentro dos documentos. Neste caso, na representação, o número um é substituído pelo valor da frequência do atributo dentro do texto.

Além dessas técnicas, ainda podem ser aplicas técnicas que atribuem um peso representativo que, além de considerar a informação contida somente dentro de cada documento, considera informações contidas em toda a base de dados, como é o caso da técnica TF-IDF e a Transformação do modelo Bayesiano, apresentadas nas proximas seções.

2.5.1 TF-IDF

A técnica TF-IDF (*Term Frequency – Inverse Document Frequency*) atribui a cada termo da base textual um peso que representa a sua importância considerando todo o *corpus* (MANNING; RAGHAVAN; SCHTZE, 2008). Este peso aumenta proporcionalmente ao número de vezes que o termo aparece no documento e é compensado pela frequência do termo no *corpus* (SALTON; ALLAN, 1994).

A contagem de termos, $TF_{i,j}$, é o número de vezes que um termo t_i aparece em um documento d_j, porém é normalizada para evitar polarização em documentos longos, sendo definida por:

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{2.2}$$

Nesta equação, $n_{i,j}$ é o número de ocorrências do termo t_i no documento d_j, e o denominador é a soma do número de ocorrências de todos os termos k pertencentes ao documento d_j.

A frequência inversa do documento, IDF_i, definida na Equação 2.3, é uma medida que representa a importância geral do termo t_i, obtida pela divisão do número total de documentos D pelo número de documentos que contêm o termo t_i seguida pelo cálculo do logaritmo desse quociente. Dependendo de como será aplicada essa técnica, é comum se usar no denominador $1 + |\{d : t_i \in d\}|$, pois caso o termo t_i não esteja presente em nenhum documento do *corpus*, ocorrerá uma divisão por zero.

$$IDF_i = \log \frac{|D|}{|\{d : t_i \in d\}|} \tag{2.3}$$

Por fim, é calculado o valor TF-IDF para um termo t_i que relaciona as duas medidas, Equação 2.4. Essa técnica, diferente do IG, não considera a classe do problema para o cálculo do peso de cada termo.

$$TFIDF_{i,j} = TF_{i,j} \times IDF_i \tag{2.4}$$

O estudo de Paltoglou e Thelwall (2010) investiga variações desta versão original apresentada comparando diferentes abordagens para a representação tanto da presença dos termos dentro dos documentos, como dentro de todo o *corpus*.

2.5.2 Transformação do modelo Bayesiano

A técnica de classificação *Naïve Bayes*, que é explicada em maiores detalhes na seção 2.6.1, utiliza uma transformação de dados específica para cada classe, ou seja, cada característica a_i, no contexto textual, palavra/termo de um documento, apresenta diferentes valores numéricos para cada classe c. Esses valores, denominados de *likelihood*, $Pr(a_i|c)$, representam a importância da palavra na classe (MANNING; RAGHAVAN; SCHTZE, 2008).

Essa transformação é realizada na etapa de aprendizagem do classificador que, por ser uma técnica de ML, utiliza medidas iniciais em sua aplicação obtidas a partir de dados de aprendizagem. No contexto deste trabalho, esses dados são documentos textuais já classificados nas classes positivo e negativo.

Para problemas em que a representação das amostras é realizada através de atributos com valores não numéricos, ou seja, atributos nominais, o algoritmo de transformação tradicional do classificador *Naïve Bayes*, denominado de binário ou então somente *Naïve Bayes*, não apresenta bom desempenho, uma vez que esse não considera a frequência dos atributos dentro de uma mesma amostra (WITTEN; FRANK; HALL, 2011).

Sendo assim, para problemas com essa característica, como o de Mineração de Opiniões e Análise de Sentimentos, o algoritmo *Multinomial* foi desenvolvido (MCCALLUM; NIGAM, 1998). Nesse algoritmo, assume-se que cada documento é representado pela frequências dos termos nele contidos, ver seção 2.5, e a probabilidade $Pr(a_i|c)$ é definida da seguinte maneira:

$$Pr(a_i|c) = \frac{1 + \sum_{j=1}^{|S|} N_{a_i s_j} Pr(c|s_j)}{|\mathbf{A}| + \sum_{k=1}^{|\mathbf{A}|} \sum_{j=1}^{|\mathbf{S}|} N_{z_k s_j} Pr(c|s_j)}, \tag{2.5}$$

em que $|S|$ é o tamanho do conjunto total de amostras de treinamento, $N_{a_i s_j}$ é o número de vezes que o atributo a_i ocorre na amostra s_j, $Pr(c|s_j) \in \{0, 1\}$ indica se s_j pertence a classe c, $|\mathbf{A}|$ é o número total te atributos contidos no conjunto de treinamento e $N_{z_k s_j}$ é o número de vezes que o atributo $z_k \in \mathbf{A}$ ocorre na amostra s_j.

2.6 Mineração de Dados

Os processos de Mineração de Dados são responsáveis por encontrar padrões, irregularidades e regras em uma base de dados através da aplicação de técnicas de Inteligência Artificial, caracterizando assim a extração do conhecimento (HAN; KAMBER; PEI, 2006).

Segundo Fayyad et al. (1996), para a realização deste processo, há diversos métodos, sendo os mais importantes os de: Classificação, Modelos Relacionais entre variáveis, Análise de Agrupamento (Clusterização), Sumarização, Modelo de Dependência, Regras de Associação e Análise de Séries Temporais. É importante destacar, que muitas dessas técnicas são aplicáveis para transformação de dados, servindo para refinamento dos dados para uso posterior de técnicas de ML.

Os métodos de Classificação associam ou classificam um item em uma ou várias classes pré-definidas normalmente através da aplicação de uma técnica estatística. Essa classificação tem o objetivo de relacionar as descrições gráficas e algébricas das características diferenciais observadas na base de dados para classificá-las em classes (DANTAS et al., 2008).

As técnicas de Classificação e outros métodos de mineração, como Regras de Associação e Análise de Séries Temporais, podem ser divididas em tipos considerando sua necessidade de informação supervisionada dos dados para treinamento. Técnicas consideradas supervisionadas requerem amostras de dados de treinamento que, além de atributos representativos também apresentam um indicativo à qual(is) classe(s) elas pertencem. Com isso, é possível realizar o treinamento da técnica de mineração que utilizará essas amostras para a consolidação de um modelo que irá capturar os padrões existentes nas amostras pertencentes às diferentes classes.

Além das técnicas supervisionadas, há também técnicas não supervisionadas e semi-supervisionadas. Técnicas não supervisionadas, como Redes Neurais do tipo *Self Organinzing Maps* (HAYKIN, 2001) e técnicas de Clusterização (RUSSELL; NORVIG, 2009), apesar de também apresentarem uma etapa de treinamento, não fazem o uso de qualquer informação supervisionada e utilizam métricas de distância para estabelecer similaridades nos dados para a identificação de agrupamentos de amostras. Já técnicas semi-supervisionadas se caracterizam por utilizar um pequeno conjunto de dados supervisionados e realizar um treinamento inicial com ele. Após isso, são utilizados métodos para a integração de amostras sem conhecimento supervisionado na base de treinamento e um novo treinamento é realizado se caracterizando um processo incremental (RUSSELL; NORVIG, 2009).

É importante esclarecer que as definições relatadas acima se referem a técnicas de MD. O processo de extração do conhecimento pode, em outras etapas, como no pré-processamento, aplicar técnicas que requerem conhecimento supervisionado, mas tal fato não implica na aplicação ou não de técnicas de mineração não supervisionadas.

Tendo em vista o problema de classificação de documentos textuais abordado neste trabalho, esta seção apresenta três classificadores supervisionados. A seção 2.6.1 relata o funcionamento do classificador *Naïve Bayes* que é baseado na probabilidade de ocorrência dos termos. Em seguida, na seção 2.6.2, é apresentado o classificador SVM que utiliza um espaço dimensional de atributos para a identificação da fronteira de decisão entre as classes. Por fim, na seção 2.6.3, o classificador RNAs, uma rede de nerônios matemáticos divida em camadas, é apresentado.

2.6.1 *Naïve Bayes*

Os classificadores que utilizam a técnica de estatística *Naïve Bayes* (NB) são os mais utilizados em MD (CHAKRABARTI, 2002). Essa técnica é considerada ingênua (*Naïve*) por levar em consideração que os atributos representativos são condicionalmente independentes um dos outros e, por se tratar de um método supervisionado, sua aplicação se divide numa etapa de aprendizagem e outra de classificação.

Na etapa de aprendizagem é definida uma distribuição geradora $Pr(a_i|c)$, também conhecida como *likelihood*, para cada classe c, que representa a importância de um atributo a_i na classe c a partir dos dados de treinamento. Essa medida também pode ser entendida como uma transformação de dados como foi explicado na seção 2.5.2.

Além disso, cada classe c possui uma probabilidade *a priori* associada $Pr(c)$, tal que $\sum_c Pr(c) = 1$, que representa a probabilidade inicial de uma amostra pertencer à classe c. Normalmente, o cálculo utilizado para a definição de $Pr(c)$ é simplesmente o número de amostras presentes em cada classe dividido pelo número total de amostras do conjunto de treinamento.

Após a obtenção dessas duas medidas, a etapa de aprendizagem é concluída e o algoritmo é capaz de realizar a classificação de novas amostras que não foram utilizadas na etapa de aprendizagem.

Para essa classificação, o classificador *Naïve Bayes* utiliza o cálculo da probabilidade *a posteriori* $Pr(c|\mathbf{x})$ que define qual é a probabilidade da amostra x pertencer à classe c, considerando os atributos, termos, presentes em $\mathbf{x}$. Segundo Manning, Raghavan e Schtze (2008), em problemas de classificação textual em que a matriz de representação dos documentos pode ser muito esparsa, ou seja, conter uma grande quantidade de valores zerados, a maneira clássica de definição da $Pr(c|\mathbf{x})$ (MCCALLUM; NIGAM, 1998) pode ser alterada para:

$$\log Pr(c|\mathbf{x}) = \log Pr(c) + \sum Pr(a_i|\mathbf{x}) \times \log Pr(a_i|c), \tag{2.6}$$

em que $Pr(c)$ é a probabilidade *a priori* da classe c, $Pr(a_i|c)$ é a *likelihood* do atributo a_i pertencente ao conjunto de atributos de treinamento A, utilizado para a definição da *likelihhod* com visto na seção 2.5.2, e $Pr(a_i|\mathbf{x}) \in \{0, 1\}$ indica se a_i está ou não presente em $\mathbf{x}$.

2.6.2 *Support Vector Machines*

O classificador *Support Vector Machines* (SVM) tem como principal objetivo mapear as amostras de dados em um espaço dimensional e identificar as fronteiras entre os conjuntos de amostras pertencentes a diferentes classes. Por necessitar de um conjunto de treinamento com o conhecimento supervisionado, o SVM também é considerado um classificador supervisionado.

Sendo assim, o SVM, na etapa de treinamento, constrói um espaço cujas as dimensões são definidas pelos atributos representativos das amostras de treinamento e os valores desses

atributos definem a posição de cada uma das amostras no espaço construído. Para um melhor entendimento, a Figura 2 ilustra um espaço bidimensional formado pelo atributos "médico"e "computador"de um problema de classificação textual onde quatro documentos, x_1, x_2, x_3, x_4, estão inseridos e cada um deles apresenta diferentes pesos para cada um do atributos conforme a Tabela 1.

Figura 2: Elementos textuais representados em um espaço bidimensional.

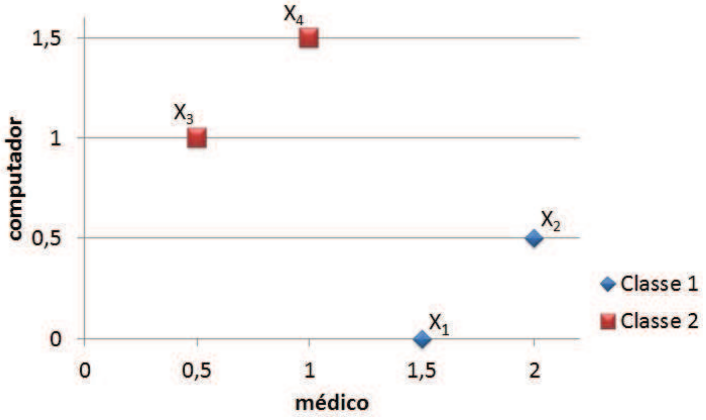


Tabela 1: Elementos textuais representadas por duas palavras.

Documentos	Pesos	
	"médico"	"computador"
x_1	1,5	0
x_2	2	0,5
x_3	0,5	1
x_4	1	1,5

A etapa de aprendizado do SVM tem como principal objetivo encontrar a “linha”[2] que realiza a separação das amostras pertencentes a diferentes classes. Para isso, são utilizadas as amostras de treinamento com o intuito de encontrar a linha que maximiza a distância, d, entre ela e a amostra mais próxima de cada classe, definindo assim a linha ótima, como ilustrado na Figura 3 (BURGES, 1998). Neste caso são utilizadas as classes denominadas Classe1 e Classe2.

A partir dessa definição, o algoritmo do SVM identifica as amostras de cada classe que estão mais próximas à linha divisória e às caracteriza como *Support Vectors* (SVs) como mostra a Figura 4[3] em que essas aparecem destacadas com um círculo em volta. Para que isso seja possível, o SVM utiliza técnicas de otimização, fazendo com que os SVs sejam definidos através

[2]Para espaços tridimensionais o objetivo é encontrar planos. Para casos de espaço com mais de três dimensões, o objetivo é encontrar hiperplanos.

[3]Figura adaptada de (BURGES, 1998).

Figura 3: Separação de classes com a utilização de uma "linha" ótima definida pelo SVM.

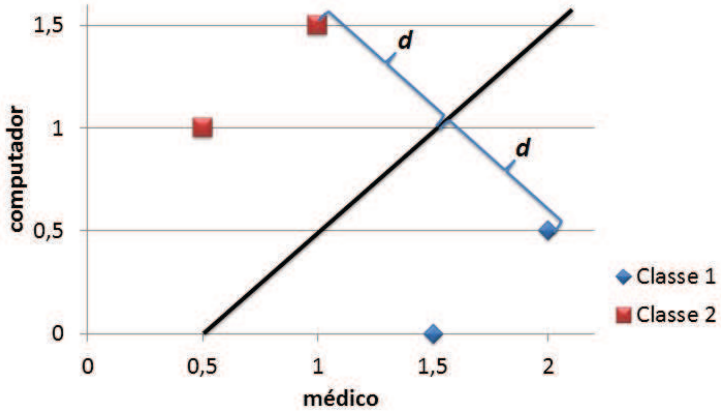
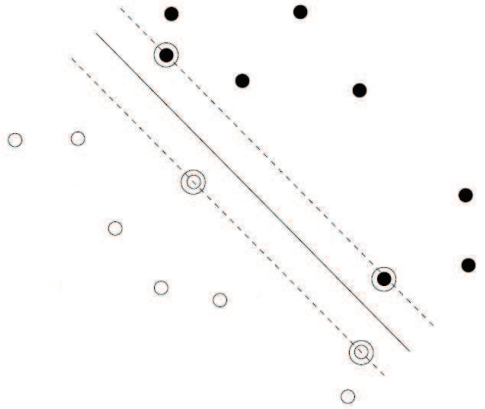


Figura 4: Identificação das amostras *Support Vectors*.



da maximização da seguinte equação:

$$W(\boldsymbol{\alpha}) = \sum_{i=1}^{|N|} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{|N|} y_i y_j \alpha_i \alpha_j (\boldsymbol{x_i} \cdot \boldsymbol{x_j}) \tag{2.7}$$

sujeito a:

$$\sum_{i=1}^{|N|} y_i \alpha_i = 0; \forall_{i=1}^{N} : 0 \leq \boldsymbol{\alpha} \leq C, \tag{2.8}$$

onde $\boldsymbol{x_i}$ é uma amostra de treinamento, y_i sua respectiva classificação correta sujeita a $y_i \in \{+1, -1\}$, $|N|$ é o número total de amostras de treinamento, $\boldsymbol{\alpha}$ é um vetor de variáveis multiplicadoras a serem determinadas pela otimização, denominadas multiplicadores de *Lagrange* (SEMOLINI, 2002), e o parâmetro $C \geq 0$ é definido pelo usuário sendo que seu papel é controlar a relação entre a complexidade do algoritmo e o número de amostras de treinamento classificadas incorretamente.

Com a maximização da Equação 2.7, as amostras de treinamento $\boldsymbol{x_i}$ que apresentarem seu respectivo multiplicador de *Lagrange* α_i maior que zero são identificadas como SVs.

Além dos SVs, o aprendizado do SVM também define o elemento b_0, utilizado na etapa de

classificação, chamado de intercepto. Esse parâmetro tem a mesma função do deslocamento linear da equação de uma reta, ou seja, deslocar a reta em relação à origem em todos os pontos. Para a definição de b_0 a seguinte equação é aplicada:

$$b_0 = -\frac{1}{2}\left[\max_{(i|y_i=-1)}\left(\sum_{j=1}^{|N_{sv}|} y_j\alpha_j(\boldsymbol{x_i}\cdot\boldsymbol{x_j})\right) + \min_{(i|y_i=1)}\left(\sum_{j=1}^{|N_{sv}|} y_j\alpha_j(\boldsymbol{x_i}\cdot\boldsymbol{x_j})\right)\right], \tag{2.9}$$

onde é utilizada a mesma notação das Equações 2.7 e 2.8 e $|N_{sv}|$ é o número de *Support Vectors* definidos pela aplicação da Equação 2.7.

A partir do aprendizado, o algoritmo de classificação do SVM utiliza a seguinte equação para classificar uma amostra $\boldsymbol{x^t}$:

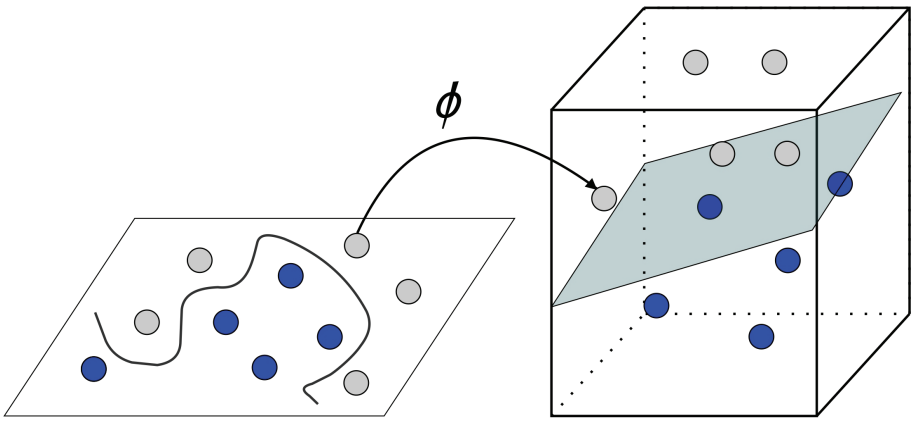
$$d(\boldsymbol{x^t}) = sgn\left(\sum_{i=1}^{|N_{sv}|} y_i\alpha_i(\boldsymbol{x_i^{sv}}\cdot\boldsymbol{x^t}) + b_0\right), \tag{2.10}$$

onde $\boldsymbol{x_i^{sv}}$ são os SVs.

Pode-se perceber que a Equação 2.10 faz o uso da função sinal, fazendo com que só existam dois possíveis resultados, 1 e -1, que correspondem as classes que a amostra $\boldsymbol{x^t}$ pode ser classificada. Desta maneira, o SVM é considerado um classificador binário e, para problemas com mais de duas classes, é necessário o treinamento e definição de mais de um modelo SVM.

O algoritmo do SVM, tanto na aprendizagem como na classificação, ainda é capaz de tratar casos de conjuntos de elementos não-linearmente separáveis[4]. Para isso, ele faz o uso de uma função de *Kernel* ao invés dos produtos utilizados nas equações para a multiplicação dos vetores de características de dois elementos, ($\boldsymbol{x_i}$,$\boldsymbol{x_j}$). Caso o problema não seja resolvido, são utilizadas variáveis de folga, fazendo com que a linha de separação das classes sofra um deslocamento local (SEMOLINI, 2002).

Figura 5: Aplicação ϕ de uma função de *Kernel*.



[4]Conjuntos de dados pertencentes a duas ou mais classes em que não se consegue definir uma linha reta que separa as amostras pertencentes a diferentes classes (HAN; KAMBER; PEI, 2006).

A utilização de uma função de *Kernel*, $K(\boldsymbol{x} \cdot \boldsymbol{x'})$, tem como objetivo mapear as amostras em um novo espaço com a quantidade de dimensões muito maior do que a do espaço original. Com isso, o conjunto de amostras pode se tornar linearmente separável, já que as posições das amostras no novo espaço são diferentes das do espaço original, possibilitando a definição de um hiperplano que separe os elementos de diferentes classes (HAN; KAMBER; PEI, 2006). Porém, a aplicação de uma função de *Kernel* ocasiona a seleção de uma quantidade maior de SVs refletindo em uma maior complexidade de classificação que pode ser evidenciada pela Equação 2.10.

A Figura 5 ilustra a aplicação ϕ de uma função de *Kernel* K em um conjunto de amostras inicialmente mapeadas em um espaço com m dimensões, os re-mapeando em um novo espaço com M dimensões, onde $M \gg m$.

As três funções de *Kernel* mais comuns utilizadas na aplicação do SVM são:

- Função de Base Radial (RBF):

$$K(\boldsymbol{x} \cdot \boldsymbol{x'}) = \exp\left(\frac{-\|\boldsymbol{x'} - \boldsymbol{x}\|^2}{2\sigma^2}\right), \tag{2.11}$$

  onde o parâmetro σ^2 (interpretado como a variância da RBF) é especificado *a priori* pelo usuário;

- Função Polinomial:

$$K(\boldsymbol{x} \cdot \boldsymbol{x'}) = (\boldsymbol{x'} \cdot \boldsymbol{x} + 1)^d, \tag{2.12}$$

  onde o parâmetro d (grau do polinômio) é especificado *a priori* pelo usuário;

- Perceptron:

$$K(\boldsymbol{x} \cdot \boldsymbol{x'}) = \tanh(\beta_0 \boldsymbol{x'} \cdot \boldsymbol{x} + \beta_1), \tag{2.13}$$

  onde β_0 e β_1 são parâmetros ajustados pelo usuário.

Além da função de *Kernel*, caso as classes ainda não sejam separáveis linearmente, o SVM realiza a aplicação de variáveis de folga. Estas, por sua vez, envolvem processos complexos, já que essas variáveis são definidas a partir da minimização de uma medida estatística de erro de classificação utilizando as amostras de treinamento. Sendo assim, o algoritmo de aprendizagem também recebe a função de medir qual é o desvio, denominado de variável de folga, de cada elemento de treinamento que minimiza a medida de erro (BURGES, 1998).

Como pode ser visto, o SVM é uma técnica sofisticada que, para a resolução de problemas de classificação complexos, disponibiliza diversos recursos conseguindo uma interessante flexibilidade. Porém, devido a essa gama de recursos, para sua aplicação são diversos os parâmetros a serem configurados[5]. Para a maioria deles normalmente considera-se os valores padrões da

[5]Para maiores informações ver: http://www.svms.org/parameters/

implementação utilizada. Porém, a alteração de alguns deles reflete significantemente nos resultados de classificação dependendo do problema ao qual o SVM é aplicado.

Como visto anteriormente, o parâmetro C está diretamente relacionado com a complexidade do algoritmo, pois define a distância máxima d mostrada na Figura 3. A definição deste parâmetro não é trivial visto que valores grandes podem ocasionar o problema de *overfitting* em que um grande número SVs é selecionado e o classificador não consegue uma boa generalização do problema. Por outro lado, valores baixos fazem com que o SVM não consiga a definição de uma boa fronteira de decisão entre as classes caracterizando o problema de *underfitting* (ALPAYDIN, 2010).

Outro importante parâmetro é a função de *Kernel*. Segundo Rifkin (2002) não há como se definir previamente qual *Kernel* será mais adequado para o problema a ser resolvido. Sendo assim, uma boa prática é a aplicação de diferentes funções de *Kernel* em testes iniciais para então se realizar experimentos.

2.6.3 Redes Neurais Artificiais

As Redes Neurais Artificiais (RNAs) são modelos matemáticos inspirados na estrutura e aspectos funcionais de redes neurais biológicas. Tais modelos consistem em um grupo interconectado de neurônios artificiais dividido em subgrupos em forma de camadas, e processam a informação utilizando uma abordagem conexionista através de pesos e funções matemáticas. Na maioria dos casos, as RNAs são um sistema adaptativo que modifica sua estrutura baseado em informação interna ou externa, que flui através da rede durante a etapa de aprendizado. Além disso, elas são ferramentas de modelagem de dados estatísticos normalmente não-lineares e geralmente são utilizadas para modelar relacionamentos complexos entre entradas e saídas ou para encontrar padrões em dados (HAYKIN, 2001).

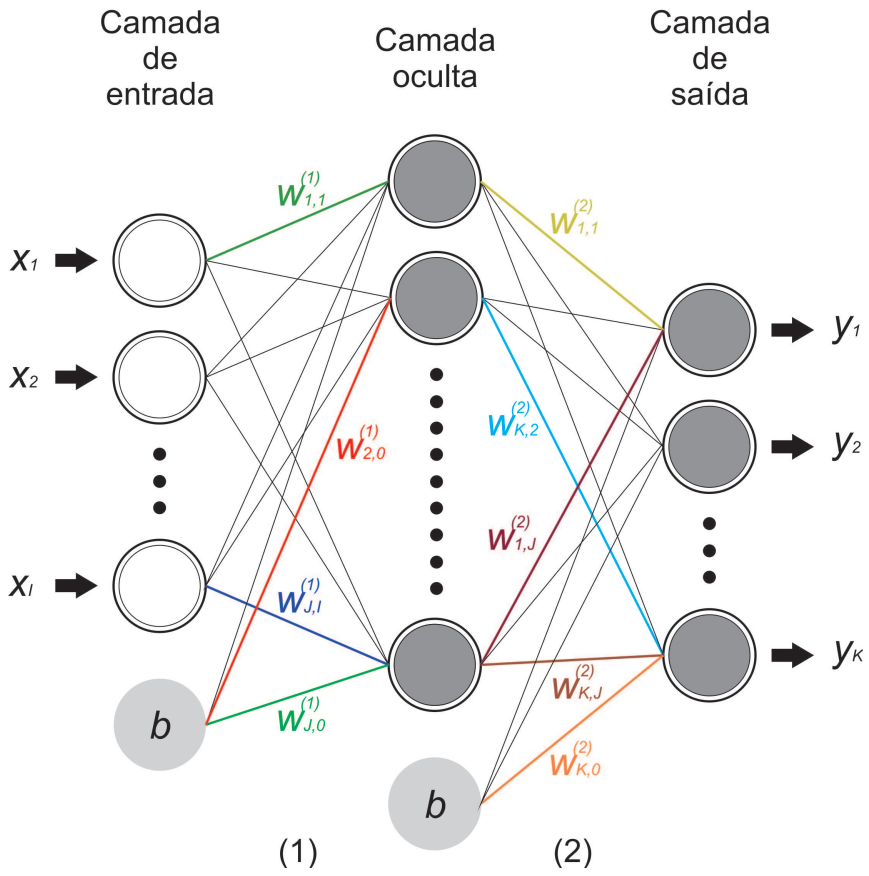
Para problemas de classificação o modelo clássico de RNAs, denominado de *multilayer perceptron* (MLP) (RUMELHART; HINTON; WILLIAMS, 1986), é comumente aplicado. Essas redes se diferenciam do modelo original de discriminação *percetron*, definido por Rosenblatt (1962), por utilizarem funções matemáticas para o processamento, denominadas de funções de ativação, parametrizáveis e não-lineares (BISHOP, 2007). Além disso, essas redes se caracterizam por cada neurônio de sua estrutura se conectar somente com os neurônios da camada seguinte, característica denominada *feedforward*, não havendo assim, conexões entre neurônios da mesma camada ou conexões de um nerônio com ele mesmo.

A Figura 6[6] mostra a estrutura de uma RNA MLP *feedforward* com três camadas: (i) entrada, que não realiza qualquer operação matemática servindo apenas como representação de uma amostra dos dados, (ii) oculta e (iii) saída. Essa duas últimas são responsáveis por realizar os cálculos matemáticos tanto para o processo de treinamento como para a tarefa de classificação, sendo que a saída da rede resulta em diferentes níveis de ativações dos neurônios de saída

[6] As notações presentes na Figura 6 seguem as mesmas definições da Equação 2.14.

que devem ser interpretados para consolidar a classificação.

Figura 6: Estrutura de uma RNA *multilayer perceptron*.



Considerando a quantidade de neurônios e pesos de uma RNA, o número de neurônios na camada de entrada, I, é definido pela quantidade de características representativas das amostras e a quantidade de neurônios da camada de saída, K, está relacionado com o número de classes do problema e sua representação. Já o número de neurônios na(s) camada(s) oculta(s) é definido pelo usuário previamente à aplicação. Por fim, a quantidade de pesos é definida pelo número de neurônios em cada uma das camadas, já que, em redes MLP *feedforward*, cada um dos neurônios se conecta através de pesos $\boldsymbol{W}$ com todos os neurônios da camada seguinte, assim como ilustrado na Figura 6.

As RNAs do tipo MLP são técnicas de aprendizado de máquina supervisionado e seu treinamento pode ser realizado através do algoritmo *back-propagation*, definido inicialmente por Werbos (1974), capaz de reconhecer padrões complexos dos dados e utilizar funções de mapeamento não-lineares (FREEMAN; SKAPURA, 1991). Este algoritmo de treinamento é dividido em duas etapas: (i) propagação e (ii) retropropagação, sendo que essas etapas são repetidas sucessivamente até que a rede atinja um dos critérios de parada do treinamento definidos previamente. Tais critérios podem ser: o número de repetições em que todas as amostras de treinamento são propagadas pela rede, denominadas de épocas; um erro mínimo de classificação

atingido; a variação de seu aprendizado definida pela variação do erro; ou ainda, algum outro método aplicado a técnicas iterativas.

A etapa de propagação das amostras se caracteriza pela passagem das amostras por todos os neurônios de todas as camadas no sentido da camada de entrada para a de saída. Com isso, cada um dos neurônios gera uma saída, fazendo o uso das funções de ativação, que é repassada para cada um dos neurônios da próxima camada até que se chegue na camada de saída da rede, resultando na resposta final de classificação (FREEMAN; SKAPURA, 1991).

Já a etapa de retropropagação inicia após a propagação de uma amostra, ou conjunto de amostras, de treinamento com a comparação da saída obtida com a saída desejada computando-se o erro da rede. Considerando que a etapa de aprendizagem tem como objetivo encontrar o conjunto de pesos $\boldsymbol{W}$ que minimiza o erro de classificação da RNA, para a aplicação do algoritmo *back-propagation*, pode-se utilizar o método de otimização denominado de Gradiente Descendente. Tal método, realizará o mapeamento do erro no espaço de atributos e procurará a combinação dos valores dos pesos que resultará no menor erro (HAYKIN, 2001). A partir do cálculo do gradiente δ, é então definida a atualização dos pesos da rede que ocorrerá da camada de saída para a de entrada considerando as saídas dos neurônios e os valores atuais dos pesos da rede (BISHOP, 2007).

A etapa de propagação para uma RNAs com uma camada oculta, como a ilustrada na Figura 6, pode ser formalmente definida como:

$$y_k(\boldsymbol{X}, \boldsymbol{W}) = \sigma\left(\sum_{j=1}^{|M|} w_{k,j}^{(2)}\theta\left(\sum_{i=1}^{I} w_{j,i}^{(1)}x_i + w_{j,0}^{(1)}\right) + w_{k,0}^{(2)}\right), \tag{2.14}$$

onde $y_k(\boldsymbol{X}, \boldsymbol{W})$ representa a saída do neurônio da camada de saída y_k, $\boldsymbol{X}$ é o vetor de atributos da amostra a ser propagada, $\boldsymbol{W}$ é o conjunto de pesos da rede, x_i é o atributo i de $\boldsymbol{X}$, $w_{j,i}^{(1)}$ é o peso entre o neurônio i de entrada e o neurônio j da camada oculta, I é o número total de atributos representativos das amostras, $w_{j,0}^{(1)}$ é o peso do bias[7] da camada de entrada, θ representa a função de ativação do neurônio j da camada oculta definida inicialmente, sendo que o resultado de θ dá origem à saída do neurônio. Os termos $w_{k,j}^{(2)}$ e $w_{k,0}^{(2)}$ representam o peso entre o neurônio j da camada oculta e o neurônio k da camada de saída e o peso do bias da camada de oculta, respectivamente. Já os termos subescritos $^{(1)}$ e $^{(2)}$ são utilizados para diferenciar os dois conectores existentes entre a camada de entrada e a oculta, $^{(1)}$, e entre a camada oculta e a de saída, $^{(2)}$, ilustrados na Figura 6, e σ representa a função de ativação da camada de saída.

Já a etapa de retropropagação é definida da seguinte maneira:

$$w_{k,j}^{(l)}(n+1) = w_{k,j}^{(l)}(n) + \eta\delta_j^{(l)}(n)y_j^{(l-1)}(n), \tag{2.15}$$

onde $w_{k,j}^{(l)}(n+1)$ é o peso entre os neurônios j e k do conector de camadas l da próxima época

[7] O bias tem o mesmo conceito do elemento b_0 do SVM como descrito na seção 2.6.2.

$(n+1)$, $w_{k,j}^{(l)}(n)$ é o valor do peso na época atual (n), η é o parâmetro de aprendizagem da rede que está relacionado com o tempo e eficiência de treinamento e é definido previamente ao treinamento, $\delta_k^{(l)}(n)$ é o gradiente local e $y_j^{(l-1)}(n)$ é o valor da saída do neurônio y_j do conector de camadas l na época n.

O resultado final do treinamento são os valores dos pesos da rede atualizados de forma que, ao se propagar uma amostra, da mesma maneira realizada pelo treinamento na etapa de propagação, a rede seja capaz de indicar a qual classe essa amostra pertence. Para isso, os diferentes índices de ativação dos neurônios da camada de saída devem ser interpretados considerando a representação da informação supervisionada das amostras de treinamento.

Para a aplicação do treinamento de uma RNA, são dois os principais parâmetros a serem definidos. O primeiro deles é a função de ativação a ser utilizada nos neurônios que realizam os cálculos de classificação. Essa definição está relacionada com a forma de representação do problema e sua complexidade. O outro parâmetro é o número de neurônios em cada uma das camadas ocultas, que pode estar relacionado com a quantidade de neurônios presentes na camada de entrada e na de saída. Porém, não há como definir previamente com exatidão qual é a quantidade de neurônios ocultos que proporcionará melhores resultados de classificação, já que isso também está relacionado com a quantidade de regiões separáveis no espaço euclidiano, que, para problemas complexos não-linearmente separáveis, é difícil de se determinar (ZURADA, 1992).

2.6.4 Comparação teórica entre SVM e RNAs

Segundo Romero e Toppo (2007), SVM e RNAs do tipo *feedforward* apresentam suas estruturas muito similares já que ambos realizam a classificação através de uma função resultante que é expressa como uma combinação linear de simples funções. Sendo assim, de forma abstrata e genérica, as funções resultantes de classificação de ambos os classificadores podem ser definidas em uma só da seguinte maneira:

$$f(\boldsymbol{x^t}) = b + \sum_{k=1}^{M} \boldsymbol{\lambda_k} h(\boldsymbol{\omega_k}, \boldsymbol{x^t}). \tag{2.16}$$

Como mencionado anteriormente, o bias b é comum para o SVM e as RNAs (HAYKIN, 2001) e a Tabela 2 relaciona o significado dos termos da Equação 2.16 com os elementos do SVM e das RNAs, considerando uma RNA *multilayer perceptron feedforward* com somente uma camada oculta e neurônios na camada de saída com função linear (ROMERO; ALQUÉZAR, 2012).

Apesar dessa similaridade na função resultante, os dois classificadores se diferem na forma com que a solução é obtida. O número M de *support vectors* é normalmente um resultado do problema de otimização do SVM, e os *support vectors* $\{\boldsymbol{\omega_k}\}_{k=1}^{M}$ são sempre parte das amostras

Tabela 2: Correspondência dos termos da função resultante do SVM e das RNAs na Equação 2.16.

Elementos da Eq. 2.16	Elementos da estrutura do SVM	Elementos da estrutura das RNAs
$\boldsymbol{x^t}$	amostra a ser classificada	
M	num. de *support vectors*	num. de neurônios na camada oculta
h	função de *Kernel*	função de ativação
$\{\boldsymbol{\omega_k}\}_{k=1}^{M}$	*support vectors*	pesos dos neurônio da camada oculta
$\{\boldsymbol{\lambda_k}\}_{k=1}^{M}$	coeficientes definidos pela otimização	pesos dos neurônio da camada de saída

de treinamento (ROMERO; TOPPO, 2007). Já para o classificador neural, o número M de neurônios na camada oculta é um parâmetro a ser fixado/definido anteriormente ao treinamento.

Desta forma, ao contrário das RNAs, o algorítimo do SVM é capaz de automaticamente conseguir definir o tamanho M de seu modelo, selecionando os *support vectors* como uma fração do conjunto de treinamento. Entretanto, tal fato caracteriza uma estrutura rígida e um grande conjunto de *support vectors* pode ser necessário para a construção da função resultante, fazendo com que o SVM seja computacionalmente lento na etapa de classificação de amostras, sendo custoso para aplicações em tempo real (*real-time*) (BISHOP, 2007). Embora não seja trivial a tarefa de definição do número de neurônios na camada oculta de uma RNA, a complexidade do modelo pode ser controlada pela escolha de quantidades baixas para essa definição (HAYKIN, 2001).

Uma importante vantagem do SVM frente as RNAs está na abordagem de otimização. O SVM consegue definir seus *support vectors* através de um problema de otimização convexo, que sempre encontra o mínimo global e uma solução única, enquanto as RNAs utilizam métodos de gradiente descendente que podem não convergir para a solução ótima/global (HASTIE; TIBSHIRANI; FRIEDMAN, 2001; HAYKIN, 2001). Porém, algumas técnicas mostram avanços na minimização da chance de uma conversão para um mínimo local, como é o caso do Gradiente Conjugado Escalado que, além disso, tem como objetivo a aceleração do processo de convergência (MüLLER, 1993).

Neste trabalho não é realizada uma comparação mais detalhada entre os classificadores RNAs e NB considerando a diferença das abordagens, otimização e estatística, respectivamente, para a realização do treinamento dos classificadores apresentadas em seções anteriores.

2.7 Avaliação de Classificação

Após o treinamento de técnicas de classificação é possível a avaliação dos resultados obtidos através de métodos e métricas que objetivam a identificação de diversas características do classificador (LIU, 2011).

Dentre os métodos de avaliação de classificadores, os mais aplicados são o *k-fold cross-validation* e o *percentage split*. O método de validação cruzada, *cross-validation*, divide o conjunto de amostras de treinamento em k partes iguais e realiza o treinamento do classificador k vezes, sendo que em cada uma delas, são utilizadas $k-1$ partes da base para o treinamento e

a fração restante para teste classificatório. Dessa forma, o resultado de cada uma das métricas de avaliação é a média das k repetições e quanto maior for o valor do parâmetro k, maior será o tamanho da base de treinamento em cada uma das k repetições (WITTEN; FRANK; HALL, 2011). Já o método *percentage split* se caracteriza por fazer apenas uma divisão da base de dados. A fração dessa divisão é parâmetro na aplicação desse método e define diretamente qual será a proporção da base que será utilizada para treinamento, sendo o restante utilizado para teste.

Quanto as métricas de avaliação de classificação, a maioria delas são definidas com base em uma matriz que contém quantidades de amostras classificadas corretamente e incorretamente, denominada matriz de confusão. Essa matriz considera amostras positivas e negativas de uma das classes, ou seja, amostras positivas são pertencentes a uma das classes e amostras negativas são todas as outras pertencentes a outras classes. Dessa forma, essa matriz pode ser construída para cada uma das classes do problema a ser avaliada. É importante destacar que o problema de Mineração de Opiniões e Análise de Sentimentos apresenta as classes positivo e negativo, porém essas não se equivalem, necessariamente, aos conceitos utilizados na matriz de confusão.

Tabela 3: Matriz de confusão.

Classe correta	Classe predita	
	Positiva	Negativa
Positiva	# Verdadeiras Positivas (VP)	# Falsas Negativas (FN)
Negativa	# Falsas Positivas (FP)	# Verdadeiras Negativas (VN)

Na matriz de confusão representada na Tabela 3, VP representa o número de amostras positivas classificadas corretamente, FP as amostras de outras classes classificadas na classe positiva, FN a quantidade de amostras da classe positiva classificada em qualquer outra classe e VN o número de amostras das outras classes classificadas corretamente.

A partir da matriz de confusão as métricas acurácia, *precision* e *recall*, comumente utilizadas na avaliação de classificadores (LIU, 2011), podem ser definidas. A acurácia, Equação 2.17, permite uma avaliação geral de acerto de classificação considerando amostras negativas e positivas e todas as classes. Já a métrica *precision* representa a capacidade do classificador de rejeitar amostras de outras classes, Equação 2.18. Por último, a Equação 2.19 define a métrica *recall* que avalia a capacidade de classificação de amostras positivas, ou seja, da classe considerada para a criação da matriz de confusão.

$$\text{acurácia} = \frac{VP + VN}{VP + VN + FP + FN} \tag{2.17}$$

$$\textit{precision} = \frac{VP}{VP + FP} \tag{2.18}$$

$$\textit{recall} = \frac{VP}{VP + FN} \tag{2.19}$$

Além das métricas acima descritas, em experimentos de comparação entre dois ou mais classificadores, se mostra relevante uma análise que considere a diferença entre os resultados gerados por cada um deles. Para isso, há vários testes estatísticos possíveis de serem aplicados dependendo do método de avaliação adotado e alguns desses são descritos em (ALPAYDIN, 2010).

Considerando o método de avaliação *k-fold cross-validation* o teste que melhor se aplica para a comparação de desempenho entre dois classificadores é o teste de significância *t* (WITTEN; FRANK; HALL, 2011). Nesse teste, cada uma das k repetições de cada um dos classificadores é comparada considerando que o conjunto de amostras de teste e treinamento é o mesmo para a realização das repetições para ambos os classificadores.

Para essa comparação, o resultado da métrica de avaliação que se deseja avaliar, normalmente a acurácia, de cada um dos classificadores é subtraído um do outro gerando a diferença de acerto entre eles. A partir disso, assume-se a hipótese de que essa diferença é nula e, caso o teste de significância rejeite essa hipótese inicial, pode-se afirmar que a diferença é estatisticamente significante considerando uma determinada significância α (HAN; KAMBER; PEI, 2006). O teste *t* é definido da seguinte maneira:

$$t_{k-1} = \frac{\sqrt{k} \times m}{S}, \tag{2.20}$$

onde m representa a média das k diferenças de resultados e S o desvio padrão.

Com isso, o teste *t* para o método *k-fold cross-validation* rejeita a hipótese de que o resultado, da métrica que está sendo testada, é igual para os dois classificadores se o valor de t_{k-1} não estiver entre o intervalo $\left(-t_{\frac{\alpha}{2},k-1}, t_{\frac{\alpha}{2},k-1}\right)$ onde $-t_{\frac{\alpha}{2},k-1}$ e $t_{\frac{\alpha}{2},k-1}$ são definidos a partir da tabela de distribuição *t* disponível em (LARSON; FARBER, 2010).

2.8 Bases de Dados Desbalanceadas

Bases de dados são caracterizadas como desbalanceadas quando essas apresentam quantidades de amostras significantemente diferentes em cada uma das classes. Tratando-se de bases que contenham somente duas classes, normalmente a classe que apresenta menos amostras, denominada minoritária, é a classe de interesse e isso deve ser considerado na aplicação do treinamento de classificadores (KOTSIANTIS; KANELLOPOULOS; PINTELAS, 2006).

Bases desbalanceadas acabam por dificultar o aprendizado de técnicas supervisionadas de classificação, o que consequentemente gera resultados de classificação não satisfatórios para a classe minoritária (VAN HULSE; KHOSHGOFTAAR; NAPOLITANO, 2007). Tal fato ocorre devido aos classificadores se basearem em técnicas estatísticas que, dada a presença de mais amostras na classe majoritária, acabam por ignorar a diferença dos padrões que distinguem as classes, dificultando a identificação das fronteiras de decisão das classes, e apenas aumentam seu conhecimento dos padrões contidos na classe majoritária.

Porém, bases de dados desbalanceadas são encontradas em diversos problemas de classifi-

cação presentes no mundo real, como identificação de possíveis clientes inadimplentes no ramo de telecomunicações, classificação textual, reconhecimento de fala, detecção de vazamento de petróleo em imagens de satélites (CHAWLA, 2010) e outros, incluindo a classificação de opiniões evidenciado no estudo de Blitzer, Dredze e Pereira (2007). Para o problema de Mineração de Opiniões e Análise de Sentimentos, algumas possíveis causas do desbalanceamento podem ser apontadas: (a) as pessoas tendem a publicar suas opiniões somente sobre produtos que atenderam ou superaram suas necessidades e expectativas e (b) existem algumas opiniões que são de autoria dos próprios fabricantes dos produtos ou então de vendedores (LI et al., 2011).

Outra questão importante a se ressaltar é que, dependendo do domínio do problema, o custo do erro de classificação pode ser alto. Como exemplo, pode-se citar o problema de identificação de pacientes com câncer. Tal problema apresenta uma grande base de dados de pacientes em que possivelmente somente uma pequena fração desses dados pertence a pacientes com câncer. Desta forma, um classificador treinado com essa base de forma desbalanceada pode ocasionar a não identificação de pacientes que correm risco de vida (LING; SHENG, 2008).

Sendo assim, várias são as técnicas propostas para o tratamento do desbalanceamento de bases de dados e podem ser divididas em dois tipos: (i) a nível de base de dados, que são técnicas que atuam sobre os dados, e (ii) a nível de classificador, que se referem a forma de aplicação dos classificadores.

As técnicas aplicadas sobre bases de dados ainda podem ser dividias em outros dois tipos: *undersampling* e *oversampling*. As técnicas de *undersampling* se caracterizam por realizar o balanceamento da base de dados igualando a quantidade de amostras da classe majoritária à quantidade da classe minoritária, sendo que o critério mais comum de escolha de amostras a permanecerem na base de treinamento é o randômico. Já as técnicas de *oversampling* têm como objetivo o balanceamento através da geração ou cópia de amostras da classe minoritária até que a quantidade de amostras de ambas as classes seja a mesma (CHAWLA, 2010).

Embora ambos esses tipos consigam o balanceamento da base de dados, técnicas de *undersampling* podem descartar amostras da classe majoritária que contenham padrões dos dados importantes, assim como, dependendo da quantidade de amostras presentes na classe minoritária, gerar um volume pequeno de dados de treinamento. Já técnicas de *oversampling* podem criar, ou então, aumentar o ruído dos dados para os classificadores (CHAWLA, 2010).

Considerando as técnicas a nível de classificador, três são os tipos existentes. O primeiro deles é aplicado sobre classificadores que apresentam em sua saída algum tipo de métrica, como a probabilidade do classificador NB ou nível de ativação de neurônios em RNAs. A partir disso, é definido um limiar de corte para essa métrica que servirá para a consideração ou não da resposta do classificador (KOTSIANTIS; KANELLOPOULOS; PINTELAS, 2006).

Os outros dois tipos são o *one-class* e o *cost-sensitive*. O *one-class*, aplicado a classificadores binários como o SVM, considera somente a classificação de uma das classes conseguindo assim um melhor aprendizado dos padrões da classe minoritária (JUSZCZAK; DUIN, 2003). Já o *cost-sensitive* adota diferentes pesos para o erro de classificação de amostras pertencentes

a classe majoritária e amostras da minoritária. Com isso, na etapa de treinamento do classificador, é possível considerar um erro maior quando amostras da classe minoritária são classificadas incorretamente (LING; SHENG, 2008).

Porém, quando não há a aplicação de técnicas para o tratamento do problema de desbalanceamento, a aplicação das métricas de avaliação dos classificadores devem ser revistas (WEISS; PROVOST, 2003). Neste contexto, a métrica mais comum, a acurácia, não permite uma boa avaliação já que, os altos índices de acurácia podem estar relacionados apenas à classificação correta da maioria das amostras de classe majoritária. Por exemplo, considerando uma base de dados que apresente 90 amostras de uma classe e 10 de outra, a classificação de todas as 100 amostra para a classe majoritária representa 90% de acurácia, o que qualifica um bom classificador. Sendo assim, a análise das métricas *recall* e *precision* considerando individualmente as classes se mostra mais relevante, uma vez que pode-se identificar o nível de acerto e a capacidade de rejeição de cada uma das classes.

2.9 Considerações

Este capítulo abordou os conceitos das áreas de Mineração Textual e Análise de Sentimentos que colaboram para a compreensão do problema ao qual este trabalho está inserido. Além disso, o processo de DCBD, utilizado como base para a realização dos experimentos deste trabalho, como será visto no capítulo 4, foi apresentado e, em seguida, técnicas possíveis de aplicação nas etapas desse processo são relatadas.

Considerando o objetivo principal deste trabalho, a comparação entre técnicas de ML ao problema de Mineração de Opiniões e Análise de Sentimentos, os três classificadores utilizados neste trabalho foram apresentados de forma teórica através dos principais conceitos envolvidos e formulações matemáticas. Por fim, as principais métricas e métodos de avaliação e comparação de classificadores foram formalizados, concluindo assim, a revisão bibliográfica deste trabalho.

A revisão realizada tem como objetivo permitir uma melhor compreensão dos métodos e técnicas aplicados tanto nos trabalhos relacionados, apresentados no próximo capítulo, como os aplicados neste trabalho.

3 TRABALHOS RELACIONADOS

Neste capítulo são apresentados seis trabalhos que se relacionam com este trabalho. O estudo de Pang, Lee e Vaithyanathan (2002) se destaca por ser um dos primeiros na área de Mineração de Opiniões e Análise de Sentimentos e nele foram realizados experimentos considerando algumas técnicas tanto para a modelagem das opiniões como para a classificação. Já He, Lin e Alani (2011) apresentam os resultados do estado-da-arte em bases consolidadas da área obtidos através da aplicação de uma modelagem de dados proposta no estudo e de um classificador clássico. Esses dois trabalhos são revisados na seção 3.1 com a intenção de apresentar uma visão geral da área em termos de tratamento de dados, técnicas de classificação, assim como, bases de dados comumente utilizadas em experimentos da literatura.

Já na seção 3.2 são relatados dois trabalhos recentes que aplicam RNAs ao problema de Mineração de Opiniões e Análise de Sentimentos de maneiras bem particulares. Em (CHEN; LIU; CHIU, 2011) RNAs foram aplicadas associadas a técnicas para o mapeamento de Orientação Semântica (OS) (TANG; TAN; CHENG, 2009) para a modelagem dos dados. Já Bespalov et al. (2011) realizaram a aplicação de RNAs incorporando o classificador em uma nova técnicas para a modelagem dos dados objetivando a redução da dimensionalidade.

Por fim, na seção 3.3 são apresentados dois trabalhos que abordam o problema de desbalanceamento de dados em Mineração de Opiniões e Análise de Sentimentos. Em (LI et al., 2011) o desbalanceamento foi tratado com a aplicação de técnicas clássicas para o problema e os experimentos foram comparados com um novo método proposto pelos autores. Já Burns et al. (2011) compararam o desempenho de dois classificadores sobre bases desbalanceadas de opiniões.

3.1 Visão Geral

O trabalho de Pang, Lee e Vaithyanathan (2002) é um dos primeiros trabalhos da área de Mineração de Opiniões e Análise de Sentimentos e seu objetivo foi realizar uma investigação empírica de técnicas, tanto de representação do conhecimento, como de classificação, já aplicadas ao problema de sumarização de documentos, tratado em (JOACHIMS, 1998), para a classificação de polaridade de opinião.

Para a realização dos experimentos, foi definida uma base de dados contendo 1400 opiniões de usuários, balanceadas entre as classes positivo e negativo, referentes ao domínio de filmes. Tal domínio foi escolhido considerando a dificuldade de classificação de opiniões pertencentes a esse domínio conforme relatado no estudo de Turney (2002). Outro motivo para a construção desta base foi a disponibilidade de opiniões já classificadas na Web. Nesta base não foram aplicadas quaisquer técnicas para a redução da dimensionalidade textual, como remoção de *Stopwords* e *Stemmer*.

Para a representação das opiniões na entrada dos classificadores foram consideradas as mo-

delagens por unigramas e bigramas, sendo que na primeira delas foi adicionada a tag "not_" nas palavras consecutivas a uma expressão de negação como: "*not*" e "*didn't*". Além disso, para essa mesma modelagem, foram consideradas somente expressões que se repetiam pelo menos quatro vezes no *corpus* e para a de bigramas, as 16165 expressões mais frequentes, a mesma quantidade de unigramas selecionada.

Além disso, com o objetivo de diferenciação de palavras ambíguas e agregação de informação, foi aplicada a técnica de *Part-of-Speech* (POS) para a identificação da classe gramatical das palavras, dando origem a outras duas representações. A primeira adicionando a classe gramatical às palavras e a segunda selecionando apenas adjetivos.

Devido a aplicação do classificador *Maximum Entropy* (ME) (BERGER; PIETRA; PIETRA, 1996), na maioria dos experimentos, para a transformação das expressões foi utilizada a representação binária. Somente em um dos experimentos a frequência foi utilizada, mas em nenhum deles técnicas mais sofisticadas, como a TF-IDF, foram empregadas.

Para a aplicação dos classificadores, os autores utilizaram o método de avaliação 3-*fold cross validation*. Além do classificador ME, os classificadores SVM e NB também foram aplicados.

Dentre os experimentos que foram realizados, o classificador SVM foi o que apresentou o melhor resultado associado a modelagem com unigramas obtendo uma taxa de classificação correta de 82,9%. Já o classificador NB e o ME alcançaram seus melhores índices de classificação com a modelagem por unigramas e com os 2633 unigramas mais frequentes dentro da base, 78,7% e 80,0%, respectivamente. A aplicação da técnica POS, tanto para a associação das palavras com as classes gramaticais como para a representação dos textos utilizando somente adjetivos não resultou em ganhos para a tarefa de classificação.

A partir dos resultados obtidos, os autores concluíram que as técnicas comumente aplicadas na área de sumarização de textos apresentam desempenho inferior quando aplicadas ao problema de mineração de opiniões realizando a comparação de seus resultados com resultados de experimentos de classificação manual. Além disso, observaram que abordagens de representação do conhecimento simplistas, como unigramas, conseguem atingir melhores resultados de classificação correta. Apesar disso, os autores mencionam que uma técnica que consiga mapear a questão da posição das palavras pode resultar em melhores resultados de classificação devido a grande presença de oposição de ideias nas opiniões.

Considerando a atividade de representação das opiniões, He, Lin e Alani (2011) destacam-se por proporem um novo método que, associado à aplicação de classificadores clássicos, consegue atingir o resultado do estado-da-arte sobre a base construída por Pang e Lee (2004).

O método proposto é uma alteração do método *Joint Sentiment-Topic* (JST) (LIN; HE, 2009) que se caracteriza por fazer um mapeamento dos documentos em um novo espaço dimensional baseado na coocorrência dos termos com a utilização da técnica *Latent Dirichlet Allocation* (LDA) (BLEI; NG; JORDAN, 2003).

O JST original se trata de um método para representação não-supervisionado, considerando a necessidade de documentos previamente classificados, que faz o uso somente de um dicionário

associado à informação de polaridade de cada uma das palavras. Ao se realizar o mapeamento dos documentos no novo espaço gerado pela LDA, os documentos podem ser agrupados formando assim diferentes tópicos do domínio. A partir dessa nova representação, a escolha de N tópicos pode ser utilizada como a representação dos documentos para técnicas de classificação.

Já o método de He, Lin e Alani (2011) se diferencia por incluir o conhecimento supervisionado dos documentos, ou seja, a informação correta de classe, positivo ou negativo, na geração do novo espaço de representação. Essa alteração mantém a característica do JST de ser independente do domínio/assunto das opiniões por utilizar somente o conhecimento de um dicionário genérico de sentimentos, e agrega um nível de informação relevante na geração da nova representação dos documentos evidenciada pelos resultados alcançados.

Para a realização dos experimentos foram utilizadas cinco bases de dados. Além da segunda versão da base do estudo (PANG; LEE; VAITHYANATHAN, 2002) contendo opiniões referentes a filmes consolidada em (PANG; LEE, 2004), também foram utilizadas parte das quatro bases construídas por Blitzer, Dredze e Pereira (2007) contendo opiniões referentes a livros, DVDs, eletrônicos e produtos de cozinha, sendo que todas as cinco continham 1000 opiniões da classe positivo e 1000 da classe negativo.

A partir da nova representação gerada pelo método proposto, tendo como entrada as opiniões modeladas através de *Bag-of-Words* (BOW) e a remoção de pontuação, caracteres não numéricos e *stopwords*, os classificadores NB, SVM e ME foram aplicados através do método de avaliação 5-*fold cross-validation*. Porém, somente os resultados do classificador ME são relatados no estudo, visto que ele apresentou os melhores índices de classificação correta.

Os resultados demonstram que, para todas as bases, a nova representação associada ao ME conseguiu atingir melhores resultados que os resultados de base de comparação utilizados. Para o domínio de filmes, a classificação correta atingiu 94,98% apresentando uma diferença de mais de 10% para o JST original. Os domínios de livros, DVDs, eletrônicos e produtos de cozinha atingiram uma acurácia de 89,95%, 91,7% 88,25% e 89,85%, respectivamente.

Além disso, os autores também realizaram experimentos considerando o problema de *cross-domain* em que se deseja construir um classificador de opiniões independente do domínio ao qual as opiniões estão inseridas. Nesse contexto, o método proposto também conseguiu superar os resultados do estado-da-arte sobre as bases de Blitzer, Dredze e Pereira (2007), mas os experimentos não são aqui relatados vista a fraca relação com este trabalho.

3.2 Aplicação de RNAs em Análise de Sentimentos

Em (CHEN; LIU; CHIU, 2011), os autores propõem uma nova abordagem para a classificação de opiniões unindo técnicas de ML e de Recuperação da Informação (*Information Retrieval* - IR) (HEARST, 1992). A motivação desse trabalho é que apesar de técnicas de ML conseguirem índices satisfatórios de classificação correta, o tempo gasto para seu treinamento é grande o que, em alguns casos, pode inviabilizar sua aplicação. Por outro lado, as técnicas de IR não

requerem tanto tempo para o treinamento, mas apresentam índices inferiores de classificação aos das técnicas de ML.

Sendo assim, a proposta geral do estudo foi utilizar a classificação das técnicas de IR como entrada de técnicas de ML, mais precisamente de RNAs, diminuindo a complexidade da rede e, consequentemente, seu tempo de treinamento. O objetivo foi verificar a eficiência da combinação dos dois tipos de técnicas quanto à classificação correta e o tempo de execução.

Como base de dados, o estudo define 5 *corpora* de opiniões sendo dois deles sobre a temática filmes compostos por parte do *corpus* construído por Pang e Lee (2004). Já os outros três *corpora* são formados por opiniões das temáticas: *e-commerce*, MP3 *players* e blogs.

Baseado em trabalhos como (TURNEY, 2002) e (CHAOVALIT; ZHOU, 2005), para a representação dessas opiniões foram selecionadas somente frases contendo adjetivos e advérbios com o auxílio da técnica POS. A partir disso, foram criadas regras gramaticais para a extração das palavras e expressões consideradas importantes na representação do conhecimento. Após essa seleção, a técnica TF-IDF foi aplicada com o intuito de agregar informação à representação atribuindo pesos às palavras.

A Tabela 4 contém informações detalhadas das bases utilizadas nos experimentos, sendo que a segunda coluna contém a quantidade de expressões definidas para a representação de cada um dos *corpora*.

Tabela 4: Informações sobre as bases utilizadas por Chen, Liu e Chiu (2011).

Domínio	Número de termos	Número de opiniões	Distribuição nas classes (Positivo : Negativo)
blogs	35	353	157 : 186
MP3	48	579	235 : 344
e-commerce	66	386	249 : 137
movies-1	81	500	250 : 250
movies-2	78	1000	500 : 500

Com a representação numérica das bases textuais foi possível a aplicação de quatro técnicas de IR que conseguem mapear a OS de um texto. A primeira delas foi a (i) *Semantic Association* que define um grau de associação entre palavras considerando o sentido de opinião. Outra técnica aplicada foi a *Pointwise Mutual Information* (CHURCH; HANKS, 1990) que consegue calcular a OS de uma frase considerando a coocorrências de duas palavras. Para a aplicação dessa última foram consideradas duas abordagens, uma delas (ii) considerando a coocorrências de duas palavras consecutivas, e outra, (iii) a de duas palavras próximas (TURNEY; LITTMAN, 2003). A última técnica de IR aplicada foi a (iv) *Latentic Semantic Analysis* (DEERWESTER et al., 1990) que calcula a relação entre as palavras utilizando a técnica *Singular Value Decomposition* (SVD).

Cada uma das técnicas de OS é responsável por gerar um valor para cada um dos documentos. Se tratando de valores de grandezas consideravelmente diferentes, foram aplicados dois

métodos para a normalização destes valores resultantes das técnicas de OS. O primeiro deles foi denominado de Quantitativo e normaliza os valores dentro do intervalo [0,1]. Já no segundo, Qualitativo, os valores são transformados de forma discreta, podendo assumir os valores 1, 0 e -1.

Após a normalização dos valores de OS foi aplicado o treinamento do classificador RNAs do tipo *Multilayer Perceptron* através do algoritmo *back-propagation*. A rede modelada apresentava quatro neurônio na camada de entrada, um para cada valor de OS, e sua camada de saída composta por dois neurônios que representam as classes positivo e negativo. Neste treinamento, também foi considerado o método *percentage split* variando de 50% à 90% a base de treinamento.

Além da variação quanto à entrada considerando os métodos de normalização, também foram realizados experimentos com a abordagem clássica de representação de documentos textuais, BOW, com a aplicação da transformação TF-IDF. Nesses experimentos, as opiniões foram representadas somente pelas palavras selecionadas por regras gramaticais. Nesta abordagem a entrada da rede é composta por N neurônios, onde N representa o número de termos no *corpus* relatado na Tabela 4.

Os resultados de classificação obtidos revelam que os métodos propostos no trabalho conseguem ser equivalente e em alguns casos mais eficientes, em até 6%, que a abordagem clássica BOW também considerando a aplicação de RNAs. Os resultados obtidos pelos métodos do estudo apresentam em média, considerando as 5 bases, 69,6% de acerto de classificação enquanto os com BOW 65,9% de classificação correta.

Já os resultados de tempo de execução evidenciam que os métodos de Chen, Liu e Chiu (2011) conseguiram uma redução de tempo de treinamento de até 73,7% em relação ao utilizado pela modelagem BOW.

Além dos experimentos relatados acima, no estudo foi aplicada uma técnica para a identificação de entradas do classificador RNAs que pouco contribuem para a classificação objetivando a eliminação dessas. Porém, a técnica aplicada não conseguiu identificar uma mesma saída para as cinco diferentes bases, evidenciando a particularidade da aplicação tanto de técnicas de representação como de classificação em relação ao conjunto de treinamento.

Ainda considerando a aplicação do classificador RNAs ao problema de Mineração de Opiniões e Análise de Sentimentos, o trabalho de Bespalov et al. (2011) propôs uma composição de técnicas para a redução de dimensionalidade da representação dos dados. Os autores afirmam que a modelagem clássica BOW considerando somente unigramas não é a melhor abordagem uma vez que diversas expressões relevantes para o objetivo de identificação de polaridade deixam de ser mapeadas por conterem duas ou três palavras, como é o caso de *"not good"* e *"not very good"*. Além disso, no estudo é ressaltado que além do não mapeamento da expressão, a modelagem via unigramas acaba por distorcer o real sentido da opinião, pois desta forma, perde-se a conexão de negação de algumas expressões.

Porém, a modelagem textual via bigramas ou trigramas ocasiona um aumento considerável

da dimensionalidade do problema. Por exemplo, considerando uma base de dados que contém N palavras distintas (unigramas), esta mesma base geraria N^2 bigramas e N^3 trigramas. Sendo assim, Bespalov et al. (2011) propuseram uma nova abordagem para a representação e classificação dos documentos opinativos em que a relação entre termos é considerada através do treinamento de um classificador RNAs que é capaz de aprender a função de mapeamento da coocorrência das palavras considerando as classes do problema de forma não-linear. Uma característica a ser destacada é que esse método não distingue as tarefas de representação e de classificação, se tratando então de um método *embedded* (LAL et al., 2006).

Os autores salientam que apesar do aumento da quantidade de parâmetros para a aplicação de seu método, ele consegue mapear melhor o conhecimento contido nos textos opinativos sem aumentar o custo computacional das técnicas de classificação.

Além do método proposto, nos experimentos realizados pelo estudo, a fim de comparação, foram aplicadas as modelagens BOW com unigramas, BOW com bigramas, as técnicas de transformação binária e DELTA-IDF (PALTOGLOU; THELWALL, 2010) e ainda, a técnica estatística de seleção de características *Mutual Information* (MI), que nos experimentos realizados com as modelagens de BOW, selecionou as 10000 melhores características representativas (unigramas ou bigramas).

Como base de dados foram coletadas opiniões dos sites Trip Advisor[1], referentes a hotéis, e Amazon.com[2], de diferentes segmentos de produtos. Após a coleta, as bases foram balanceadas entre as classes positivo e negativo, totalizando 96352 e 384900 opiniões de hotéis e produtos, respectivamente.

Para a aplicação dos classificadores, tanto do método proposto, como para as demais modelagens, foi utilizado o método de avaliação *percentage split* definindo 70% da base para treinamento e restante para teste.

Os resultados obtidos considerando a classificação de opiniões entre somente as classes positivo e negativo revelaram que a abordagem BOW com bigramas, associada ao classificador SVM, conseguiu a mais alta média de acerto considerando as duas bases, 92,54%, seguido pela mesma modelagem com o classificador RNAs, 92,52%. Já a técnica proposta pelo artigo atingiu uma média de 92,28% de acerto na classificação das opiniões. É importante destacar que as bases de dados utilizadas continham uma grande quantidade de opiniões, fato que pode estar fortemente associado às baixas diferenças entre os resultados.

No trabalho ainda são relatados experimentos em que o objetivo também é a identificação do sentido de opinião, porém não considerando apenas as classes positivo e negativo, mas sim, cinco níveis diferentes de intensidade de polaridade representados por uma avaliação em estrelas. Nesses experimentos a técnica proposta por Bespalov et al. (2011) conseguiu superar as demais com uma diferença de até 9,6% de classificação correta. Porém, considerando o objetivo deste trabalho e as técnicas aplicadas, a modelagem e os experimentos referentes a esse objetivo

[1] www.tripadvisor.com
[2] www.amazon.com

não são relatados aqui.

3.3 Contexto Desbalanceado

Tendo em vista o contexto em que as quantidades de amostras de dados de treinamento são significantemente diferentes entre as classes, o trabalho de Li et al. (2011) aborda esse problema na área de Mineração de Opiniões e Análise de Sentimentos. Para isso, os autores variam a aplicação de diferentes técnicas clássicas tanto para o tratamento do desbalanceamento como para a classificação. Além disso, também foi proposto um método semi-supervisionado através de re-treinamento de classificadores.

O estudo aponta que, apesar de outros estudos explorarem técnicas supervisionadas para a classificação de opiniões, essas normalmente dependem de grandes bases de dados previamente classificados. Embora já existam outros estudos considerando métodos semi-supervisionados que necessitam de uma menor quantidade de dados supervisionados, como (DASGUPTA; NG, 2009) e (LI et al., 2010), nenhum deles considera base de dados desbalanceadas, disposição normalmente encontrada nos dados disponíveis na web.

Sendo assim, foi proposto um método semi-supervisionado que utiliza uma pequena quantidade de dados supervisionados e dados que se desconhece a classe. A partir de um treinamento inicial, as amostras não classificadas são classificadas e então é realizado um re-treinamento do classificador incluindo algumas dessa amostra na base de treinamento. O diferencial do algoritmo proposto é que para a classificação de uma amostra são considerados dois classificadores do algorítimo ME treinados com as mesmas amostras, porém, com conjuntos diferentes de características representativas selecionadas randomicamente, ou seja, uma amostra é classificada considerando as duas classificações geradas pelos classificadores (*co-training*) (HALL et al., 2009). A partir disso, as amostras classificadas pelos classificadores ME com as mais altas confianças de pertencerem a uma das classes são adicionadas no conjunto de treinamento e os classificadores são re-treinados, caracterizando uma iteração.

A fim de comparação, os autores aplicaram outras cinco abordagens supervisionadas. A primeira delas é a classificação através do classificador ME sem qualquer tratamento da base de dados, ou seja, desbalanceada, que foi denominada de *Full-training*. A segunda também aplica o classificador ME e utiliza um *oversampling* randômico para a seleção de amostras da classe minoritária a serem reaplicadas. Já a terceira se caracteriza por realizar uma técnica randômica para o balanceamento das classes via *undersampling* e aplica o mesmo classificador das anteriores. As demais abordagens utilizam o classificador SVM, sendo uma delas, a *One-class* e a outra a *Cost-Sensitive* (ver seção 2.8).

Como base de dados foram coletadas opiniões referentes a quatro domínios diferentes, livros, DVDs, eletrônicos e produtos de cozinha. As quantidades de opiniões em cada base foram definidas tendo como base a quantidade de 100 opiniões na classe negativa e diferentes proporções de desbalanceamento, uma para cada domínio, considerando o estudo de Blitzer, Dredze e

Pereira (2007). A Tabela 5 contém as quantidades de opiniões utilizadas em cada base.

Tabela 5: Informações sobre as bases utilizadas por Li et al. (2011).

Domínio	Quantidade de opiniões positivas	Quantidade de opiniões negativas
livros	729	100
DVDs	608	100
eletrônicos	357	100
produtos de cozinha	378	100

Para a avaliação dos resultados, os autores utilizaram a métrica geométrica *G-mean* (KUBAT; MATWIN, 1997) que é capaz de avaliar o desempenho dos classificadores considerando a acurácia de cada uma das classes e o desbalanceamento da base.

Os resultados da aplicação das técnicas supervisionadas mostram que a aplicação da técnica *undersampling* randômica associada ao classificador ME obteve o melhor desempenho de classificação correta seguida pela aplicação da *oversampling*. O classificador SVM conseguiu seus melhores resultados através da abordagem *One-class*, mas apresentou diferenças superiores a 15% da métrica *G-mean* dos resultados alcançados pelo *undersampling* com o ME. A abordagem *Full-training* variou bastante em relação aos melhores resultados obtidos com *undersampling* apresentando diferenças entre 14% e 40%.

Já o método proposto foi comparado diretamente com a abordagem *undersampling*. Os resultados apresentados pelo método se dão em função do número de re-treinamentos realizados. Estes resultados mostram que o algoritmo proposto pode superar em até 6% os melhores resultados dos classificadores supervisionados.

O último trabalho apresentado é o de Burns et al. (2011) que se assemelha com este por realizar uma comparação entre dois algoritmos de classificação no contexto de Mineração de Opiniões e Análise de Sentimentos considerando bases de dados desbalanceadas. Apesar do estudo não propor nenhum novo método, a comparação de classificadores considerando a acurácia, *recall* e *precision* para bases de dados desbalanceadas é rara na literatura, fato evidenciado pelo ano de publicação do estudo e sua simplicidade.

Os classificadores comparados por Burns et al. (2011) foram o NB, apresentado na Seção 2.6, e um baseado em *Language Models* (LM) (MANNING; RAGHAVAN; SCHTZE, 2008). Este último também se baseia em probabilidades de ocorrências, mas se diferencia do NB por considerar não somente a ocorrência das palavras de forma independente mas também a coocorrência com outras em seu treinamento.

Para a realização dos experimentos, foram definidas três bases de dados sendo elas dos domínios de: TV, câmeras e produtos de cozinha. Essas bases originalmente são desbalanceadas contendo uma quantidade maior de opiniões positivas do que negativas. A partir disso, diversas proporções do conjunto original, de 30% a 100%, foram consideradas nos experimentos. Além desses conjuntos desbalanceados, cada um deles originou um conjunto balanceado através da

aplicação de *undersampling* randômico. A Tabela 6 apresenta as quantidades máximas e mínimas de opiniões utilizadas e a relação entre quantidade de opiniões negativas pela quantidade de opiniões positivas.

Tabela 6: Informações sobre as bases utilizadas por Burns et al. (2011).

Qnt.	TV			câmeras			cozinha		
	Desbalanceadas								
	Pos.	Neg.	Prop.	Pos.	Neg.	Prop.	Pos.	Neg.	Prop.
Min.	3712	622	0,16	1893	330	0,17	4721	1236	0,26
Máx.	12374	2076	0,16	6309	1099	0,17	15737	4119	0,26
	Balanceadas								
Min.	622		1	330		1	1236		1
Máx.	2076		1	1099		1	4119		1

Observação: as abreviaturas "Pos."e "Neg."representam as classes Positivo e Negativo, respectivamente, e a abreviatura "Prop."refere-se a palavra "proporção".

Quanto a preparação das opiniões para a aplicação dos classificadores, o estudo não menciona a aplicação de qualquer técnica de pré-processamento ou seleção de características e realiza a representação através da abordagem clássica BOW unigramas. Além disso, as quantidades de unigramas não são relatadas.

A partir dessa representação, os dois classificadores foram aplicados através do método de avaliação 10-*fold cross-validation* e as métrica acurácia, *precision* e *recall* foram utilizadas.

Os resultados obtidos revelam que os altos índices gerais de classificação correta obtidos com os conjuntos de opiniões desbalanceados estão fortemente relacionados com o *recall* da classe positiva que contém mais opiniões, já que, em média, o classificador NB obteve um *recall* de 49% para a classe negativa e 98% para a positiva e o classificador LM 44% para negativa e 98% para a positiva. Isso demonstra que ambos os classificadores apresentam dificuldade em generalizar o problema e tendem a classificar a grande maioria das amostras para a classe que apresenta a maior quantidade de amostras no conjunto de treinamento.

Além disso, considerando as diferentes quantidades de opiniões, o estudo revela que para as bases desbalanceadas as métricas *precison* e *recall* tendem a cair conforme o aumento do número de amostras de treinamento para os três domínios de opiniões. Já nos conjuntos balanceados, ainda em relação a variação da quantidade de opiniões, o comportamento das métricas varia de acordo com o domínio e o classificador.

Nos experimentos com conjuntos balanceados, o classificador NB obteve em média 88% de *recall* para a classe negativa e 92% para a positiva. Já o LM apresentou o mesmo índice para a classe negativa e 2% abaixo do NB para a classe positiva. Este fato demostra que, assim como nos conjuntos não balanceados, o classificador LM apresentou uma maior dificuldade de generalização que o NB por favorecer mais uma classe do que a outra. Porém, para ambos os classificadores a classe negativa continuou sendo a que apresentou maior dificuldade de classificação.

Em resumo, Burns et al. (2011) concluíram que o classificador NB é o mais indicado para o contexto de Mineração de Opiniões e Análise de Sentimentos em bases de dados desbalanceadas, considerando a variação do tamanho da base de treinamento e o domínio das opiniões. Porém, os baixos índices de *recall* da classe minoritária abrem um precedente para a investigação de outros classificadores, como SVM, também amplamente utilizado na área de classificação de opiniões, em conjuntos de dados desbalanceados.

3.4 Discussão

Apesar de todos os trabalhos relatados nas seções anteriores estarem parcialmente relacionados com este trabalho proposto, nenhum deles realiza uma investigação aprofundada e comparativa da aplicação do classificador RNAs no problema de Mineração de Opiniões e Análise de Sentimentos frente aos classificadores comumente aplicados (SVM e NB).

O trabalho de Bespalov et al. (2011) é o único que realiza a comparação de RNAs com SVM, porém o estudo é direcionado ao tratamento da alta dimensionalidade dos dados e, por isso, torna difícil a comparação entre os classificadores, não apresentando detalhes importantes da aplicação de ambos classificadores. Além disso, os resultados apresentados se referem somente a métrica de erro de classificação e o contexto de bases de dados desbalanceadas não é considerado.

Em (CHEN; LIU; CHIU, 2011) também é realizada a aplicação de RNAs para a classificação de opiniões e é feita a medição e análise de métricas interessantes, como o tempo gasto para o treinamento do classificador. Porém, em seus experimentos não foram aplicados outros classificadores, assim como as bases de dados utilizadas não são por totalidade as mesmas utilizadas em outros estudos, dificultando a comparação com outros métodos e experimentos. Outro fator bastante particular do estudo é o método utilizado para a seleção dos termos/expressões utilizadas para a representação das opiniões. Como mostrado na Tabela 4, a quantidade de atributos representativos é pequena se comparada com outros trabalhos relatados neste capítulo.

Em (PANG; LEE; VAITHYANATHAN, 2002) e (HE; LIN; ALANI, 2011) o classificador RNAs não foi aplicado e a análise dos resultados apresentada é simples, já que, consideram somente a métrica acurácia. Porém ambos servem como referências no sentido da evolução de técnicas voltadas à representação do conhecimento de opiniões. Apesar deste trabalho não fazer o uso de técnicas sofisticadas de representação, o trabalho de He, Lin e Alani (2011) é usado como evidência do conhecimento de possíveis técnicas que podem aumentar a taxa de acerto de classificação dos classificadores.

Embora ambos os estudos de Li et al. (2011) e Burns et al. (2011) considerarem o contexto real de bases de opiniões desbalanceadas e fazerem o emprego de técnicas clássicas para este problema de desbalanceamento, nenhum deles faz a aplicação de RNAs e nem uma análise detalhada dos resultados considerando o desbalanceamento. Além disso, o método proposto por Li et al. (2011) é considerado custoso por ser baseado em re-treinamentos e as bases de

dados utilizadas apresentam poucas amostras da classe negativo, como mostrado na Tabela 5, e o trabalho de Burns et al. (2011) não apresenta detalhes das técnicas aplicadas para a representação das opiniões.

Com a revisão dos trabalhos encontrados na literatura de Mineração de Opiniões e Análise de Sentimentos verificou-se que são raros aqueles que utilizam RNAs para a classificação, sendo que, a maioria deles focam na investigação e desenvolvimento de técnicas para a seleção de características e redução da dimensionalidade (O'KEEFE; KOPRINSKA, 2009; LI et al., 2009; DANG; ZHANG; CHEN, 2010; ABBASI, 2010; ABBASI et al., 2011; HE; LIN; ALANI, 2011).

Apesar dos estudos relatados neste capítulo, a literatura também carece de uma investigação que realiza uma comparação direta entre classificadores clássicos na área de mineração de opiniões (SVM e NB) e RNAs, considerando a aplicação das mesmas técnicas na modelagem dos dados, assim como, o uso de dados clássicos. Além disso, questões que se mostram importantes no desenvolvimento de aplicações reais, como tempo de treinamento e de classificação, são pouco exploradas nos trabalhos investigados.

O real cenário de desbalanceamento em bases de opiniões (BLITZER; DREDZE; PEREIRA, 2007; LI et al., 2011) também evidencia a necessidade de comparações de classificadores através de métricas adequadas considerando bases com essa característica, o que pode influenciar a obtenção de classificadores com pouca capacidade de generalização do problema.

4 ABORDAGEM EXPERIMENTAL

Tendo como objetivo a avaliação empírica da aplicação de RNAs no problema de Mineração de Opiniões e Análise de Sentimentos através da comparação desse classificador com SVM e NB, esse trabalho não busca alcançar os resultados de classificação do estado da arte, assim como também não propõe um novo método de classificação ou modelagem de dados.

A partir daqui, adota-se a nomenclatura Redes Neurais Artificiais (RNAs) para se referir ao modelo clássico de Redes Neurais Artificiais, o *multilayer perceptron*.

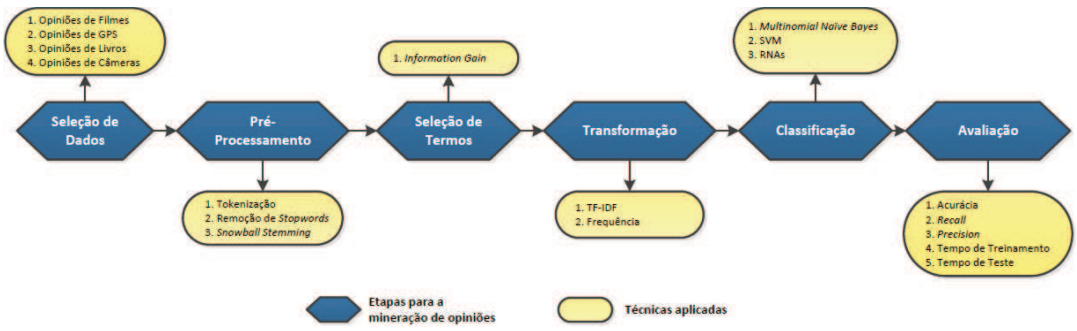
Sendo assim, este capítulo descreve a metodologia empregada para a comparação entre os classificadores tendo como base o processo de DCBD visto na seção 2.3. Nesta metodologia foram realizadas tarefas que vão desde a coleta de dados até a aplicação de métricas para a avaliação dos classificadores, objetivando um melhor conhecimento e controle do processo aplicado, assim como, uma igual representação das opiniões para os diferentes classificadores.

Além disso, na seção 4.2, são apresentadas as configurações dos experimentos de classificação realizados com os diferentes classificadores. Após, nas seções 4.3, 4.4 e 4.5 são relatados os resultados obtidos com as bases de opiniões balanceadas, desbalanceadas e com a aplicação de *undersampling*, respectivamente, considerando métricas de avaliação pertinentes para uma investigação detalhada dos classificadores. Por último, na seção 4.6 é apresentada uma discussão geral desses resultados.

4.1 Metodologia Aplicada

A metodologia aplicada para a comparação dos classificadores tem como base o processo de DCBD que originalmente é dividido em cinco etapas. Porém, considerando as particularidades apresentadas por problemas de Mineração Textual, o processo original de DCBD foi adaptado. A Figura 7 ilustra a metodologia aplicada neste trabalho que é dividida em seis etapas, tendo como diferencial a separação da aplicação de técnicas de seleção de atributos da etapa de pré-processamento. Além das etapas, a figura contém os nomes das principais técnicas aplicadas em cada uma das etapas , assim como os domínios de opiniões considerados.

Figura 7: Bases de dados, etapas e técnicas da modelagem computacional utilizadas para a avaliação dos classificadores.



Objetivando a utilização de bases e temáticas clássicas da área de Mineração de Opiniões e Análise de Sentimentos, na etapa de seleção de dados quatro bases de dados foram definidas para o estudo. Uma delas é referente ao estudo de Pang e Lee (2004) contendo opiniões sobre filmes[1] e, como mencionado no capítulo 3, amplamente utilizada para experimentos em Mineração de Opiniões e Análise de Sentimentos. Já as demais bases foram coletadas a partir do *e-commerce* Amazon.com.

Para a coleta dessas bases foi desenvolvido um programa que permitiu a extração automatizada de comentários referentes a diferentes produtos e modelos comercializados pelo site. Ao final da coleta, foram adquiridas mais de 41.000 opiniões.

Considerando a necessidade de informação supervisionada, além do texto opinativo, foi coletada também uma informação indicativa da classe a qual cada opinião pertence. Essa informação foi considerada através da avaliação direta geral que os autores das opiniões devem manifestar sobre o produto através da marcação de estrelas. Nesta avaliação, cada autor deve indicar de uma à cinco estrelas conforme sua satisfação com o produto.

Sendo assim, opiniões que apresentavam menos de três estrelas foram consideradas negativas e as opiniões associadas a uma avaliação acima de três estrelas foram consideradas positivas. Já os textos com três estrelas foram descartados visto que o objetivo deste trabalho é a classificação de opiniões pertencentes às classes positivo e negativo e tais textos podem conter informações ambíguas, dificultando o treinamento das técnicas de classificação. Esse critério de identificação de classe é o mesmo que foi utilizado por Pang e Lee (2004) para a consolidação da base utilizada naquele estudo.

Com isso, foi possível então a definição das quatro bases de dados utilizadas neste trabalho. A base do estudo de Pang e Lee (2004) é composta por 2000 opiniões referentes a filmes distribuídas igualmente entre as duas classes. Para a definição das demais bases a partir da opiniões coletadas, considerou-se a mesma quantidade e distribuição da base referentes a filmes objetivando uma melhor comparação entre elas. Sendo assim, foram selecionadas aleatoriamente opiniões referentes a livros, aparelhos de GPS e câmeras digitais consolidando as três outras bases. Tais domínios foram definido tendo como motivação as bases do estudo de Blitzer, Dredze e Pereira (2007) e a disponibilidade das opiniões no site da Amazon. É importante ressaltar que as quatro bases são compostas por opiniões escritas na língua inglesa.

Os quatro primeiros Apêndices A, B, C e D trazem exemplos de opiniões contidas nas bases definidas. Em cada um desses apêndices são expostos 4 exemplos de opiniões das bases utilizadas neste trabalho, sendo que os dois primeiros pertencem à classe negativo e os dois últimos à classe positivo.

Após a seleção e definição das bases, técnicas para o pré-processamento dos textos foram aplicadas, que têm como principal objetivo a representação dos textos com a remoção de possíveis ruídos, ou seja, informações que podem atrapalhar a tarefa de classificação, e redundância de informação. Para isso, foram consideradas as técnicas de *tokenização*, *stemming* e remoção

[1] Disponível em: http://www.cs.cornell.edu/people/pabo/movie-review-data/

de *stopwords*. Na aplicação da *tokenização* considerou-se somente a divisão dos textos por *tokens* compostos por somente uma palavra, ou seja, adotou-se abordagem de representação dos textos BOW com somente unigramas, sendo que *tokens* formados por menos de dois caracteres foram removidos. Outras unidades de análise, como bigramas, apesar do possibilitarem um melhor mapeamento do conhecimento como afirmado por Bespalov et al. (2011), não foram consideradas visto o aumento da dimensionalidade do problema.

Para a aplicação da técnica de *stemming*, optou-se pelo algoritmo *Snowball* (PORTER, 2001), uma evolução do algoritmo original de Porter (PORTER, 1980). Já para a remoção de *stopwords*, uma lista composta por 285 palavras foi utilizada (ver Apêndice E). Esta lista foi construída através da junção de uma lista básica de *stopwords*[2] com palavras que não representavam conhecimento algum para o objetivo da pesquisa. Além da remoção de *stopwords*, *tokens* não alfabéticos, como números, sinais de pontuação ou símbolos, também foram removidos.

A aplicação dessas três técnicas foi realizada através da utilização do pacote de ferramentas *Text to Matrix Generation* (TMG) desenvolvido por Zeimpekis e Gallopoulos (2005) para o ambiente de programação funcional MATLAB. Tal pacote oferece diversas técnicas para a importação e processamento de informações textuais dentro do MATLAB.

A Tabela 7 contém as informações características das bases de dados antes e após a realização do pré-processamento e as Figuras 8-11 ilustram as nuvens de palavras de cada uma das bases e cada uma das classes após a remoção de *stopwords* e aplicação de *stemming*. Para a criação das nuvens de palavras é considerada a frequência de cada uma das palavras sendo o tamanho de cada uma delas proporcional a sua frequência e as das demais.

Tabela 7: Características das bases de dados utilizadas nos experimentos.

Domínio	Antes do pré-processamento		Após o pré-processamento	
	Quantidade de termos distintos	Média de termos por documento	Quantidade de termos distintos	Média de termos por documento
filmes	39059	665,6	25456	323,2
GPS	10349	171,5	6880	75,2
livros	16155	189,9	10422	82,6
câmeras	8679	122,6	5996	53,9

A Tabela 7 evidencia a importância da aplicação de técnicas de pré-processamento. A quantidade de termos distintos foi reduzida para cerca de 65% da quantidade original em cada uma das bases, resultado adquirido não só pela eliminação de palavras, mas também pela concentração de conhecimento em um único *token* proporcionada pela aplicação do *stemming*.

Com uma breve análise das Figuras 8-11 é possível constatar que as palavras de maior frequência, maiores nas figuras, são o próprio nome do domínio, sinônimos dele, ou então referentes a características e tópicos particulares de cada domínio. Palavras relacionadas a sentimentos aparecem em tamanho médio, indicando que, apesar de sua frequência ser inferior

[2]Disponível em: http://www.ranks.nl/resources/stopwords.html

a algumas outras, elas aparecem com uma frequência relativamente grande, já que em cada nuvem, somente as 150 palavras de maior frequência dentro da base são mostradas.

Figura 8: Nuvem de palavras para a classe (a) negativo e para a classe (b) positivo da base de opiniões referentes a filmes.

(a) Classe negativo. (b) Classe positivo.

Figura 9: Nuvem de palavras para a classe (a) negativo e para a classe (b) positivo da base de opiniões referentes a GPS.

(a) Classe negativo. (b) Classe positivo.

Figura 10: Nuvem de palavras para a classe (a) negativo e para a classe (b) positivo da base de opiniões referentes a livros.

(a) Classe negativo. (b) Classe positivo.

Figura 11: Nuvem de palavras para a classe (a) negativo e para a classe (b) positivo da base de opiniões referentes a câmeras.

(a) Classe negativo.

(b) Classe positivo.

Além disso, nota-se que as nuvens de palavras da Figura 11 se diferenciam das demais por destacar em expressivamente a palavra *camera*. Este fato pode ser resultante da existência de quase nenhum sinônimos para o objeto alvo das opiniões. Nos demais domínios facilmente se visualiza pelo menos 2 palavras sinônimas que identificam o domínio, com é o caso de *movie* e *film* para o domínio de filmes, *gps*, *garmin* e *tomtom* para o domínio de GPS e *book* e *stories* para o domínio de livros. Outros fatores que pode estar relacionado com essa particularidade da Figura 11 é tamanho do vocabulário de termos distintos e média de termos por opiniões presentes na base de câmeras. Conforme apresentado na Tabela 7, a base de câmera é a que apresenta as menores quantidades absolutas destas duas características.

Objetivando a redução da dimensionalidade e a seleção dos melhores termos representativos do conhecimento, capazes de gerar padrões para a distinção das classes, na etapa de seleção de termos foi utilizada a técnica *Information Gain* (IG) apresentada na seção 2.4.4 através da implementação desta dentro do MATLAB. Visto que não há como se definir previamente qual a quantidade e o conjunto de termos que melhor representa o conhecimento, foram definidas sete quantidades de termos selecionados para cada uma das bases.

Além da representação do conhecimento, essas diferentes quantidades também permitem a avaliação dos classificadores quanto a relação que eles apresentam entre o número de atributos representativos das amostras e seu tempo de treinamento e de classificação. As quantidades de termos definidas são 50, 100, 500, 1000, 3000, 4000, 5000. Além dessas, outras quantidades foram inicialmente utilizadas, porém, para melhor visualização e relato dos experimentos e resultados, essas foram desconsideradas, visto que as possíveis constatações sobre os resultados não sofrem influência.

Após a aplicação do IG, utilizou-se a técnica de transformação dos dados TF-IDF apresentada na seção 2.5.1. Tal técnica foi aplicada sobre cada uma das bases geradas pela aplicação da técnica IG, já que, a TF-IDF utiliza informações referentes a toda base para a definição dos pesos dos termos. Com a utilização da técnica TF-IDF, foi possível então a realização dos experimentos de classificação, apresentados na seção seguinte, e a aplicação de métricas de avaliação a fim de comparação dos classificadores.

4.2 Experimentos

Os experimentos realizados neste trabalho foram executados através do ambiente MATLAB. Tal ambiente foi escolhido devido à disponibilidade dos algoritmos de classificação, facilidade de manipulação dos dados provida pela utilização da ferramenta TMG e melhor controle das etapas da metodologia adotada. Outras ferramentas, como o software WEKA[3], devido a sua estrutura própria para a manipulação dos dados e aplicação de técnicas, tornam pouco flexível o processo de mineração como um todo.

Sendo assim, foram aplicados os classificadores *Naïve Bayes* (NB), *Support Vector Machines* (SVM) e Redes Neurais Artificiais (RNAs). Para isso, utilizou-se o método de avaliação k*-fold cross-validation* onde k foi definido como 10.

O classificador NB foi aplicado utilizando o algoritmo de treinamento *Multinomial*, ver seção 2.5.2, ideal para problemas de dados textuais através da implementação de código próprio. Sendo assim, neste código foram desenvolvidas as funções de cálculo das probabilidades *likelihood* e *a priori* sobre a matriz de frequência dos termos nos documentos resultante da aplicação da TMG, assim como, a função de classificação. É importante destacar, que para o treinamento do classificador NB foi utilizada a representação dos documentos via a frequência dos termos em cada documento, diferentemente de como foi utilizada para o treinamento do SVM e das RNAs, em que o valor da transformação TF-IDF foi utilizado.

Para a aplicação do classificador SVM, utilizou-se a ferramenta LIBSVM (CHANG; LIN, 2011) implementada para MATLAB. Essa ferramenta permite a configuração de diversos parâmetros de aplicação do SVM. Porém, somente o parâmetro C foi alterado na configuração padrão da implementação que utiliza a função Base Radial como *kernel* para o tratamento da não-linearidade do problema. Desta forma, para a definição do parâmetro C foram testados os valores 0,1; 0,3; 0,5; 0,7; 1; 2; 3; 4; 5; 6; 8; 10; 12; 20; 28; 30; 32; 80; 90; 100; 120; 150; 300; 500; 700 e 1000 com base em experimentos preliminares. A realização de testes com diferentes valores do parâmetro C é necessária já que o valor deste capaz de gerar o melhor resultado de classificação é muito particular ao conjunto de treinamento utilizado e de difícil definição anteriormente ao treinamento do classificador (ALPAYDIN, 2010).

Já para a aplicação das RNAs foi utilizada uma rede *feedforward* com somente uma camada escondida treinada através do algoritmo *back-propagation*. Porém, para o treinamento, o algoritmo tradicional de otimização Gradiente Descendente não foi utilizado sendo adotado o Gradiente Conjugado Escalado (*Scaled Conjugated Gradient* - SCG) (MüLLER, 1993) implementado pelo MATLAB que tem como objetivo acelerar o tempo para a convergência da otimização. Além disso, ainda com o intuito de reduzir o tempo de treinamento da rede, a técnica *early stopping* (BISHOP, 2007) foi utilizada. Tal técnica se caracteriza por, a cada época de treinamento, realizar um teste de classificação com um pequeno número de amostras e, caso todas forem classificadas corretamente, o treinamento da rede é encerrado. Já como função de

[3]Disponível em: http://www.cs.waikato.ac.nz/ml/weka/

ativação dos neurônios das camadas oculta e de saída, optou-se pela função Tangente Hiperbólica.

Para a definição do número de neurônios contidos na camada oculta os valores 15; 27; 30; 35; 39; 47 e 55 foram escolhidos com base em experimentos preliminares e então testados. Outro aspecto da aplicação de RNAs nos experimentos realizados neste trabalho, é que, devido à definição inicial dos pesos da rede ser aleatória, cada um dos modelos definidos pelas diferentes quantidades de entradas e neurônios da camada oculta foi treinado três vezes. Tal estratégia foi adotada tanto para se evitar possíveis mínimos locais da otimização do gradiente de erro, como encerramentos precipitados causados pelo *early stopping*.

Como pode ser visto, a realização do treinamento tanto do SVM, com a variação do parâmetro C, como das RNAs, com a variação do número de neurônios na camada escondida e repetições de treinamento, acaba gerando mais de um modelo treinado para cada uma das bases utilizadas. Sendo assim, para a mensuração dos resultados apresentados nas seções 4.3, 4.4 e 4.5, somente as configurações dos classificadores que proporcionaram os melhores resultados de teste foram consideradas.

Destaca-se que a maioria dos gráficos contidos nas seções em que os resultado são relatados, seções 4.3, 4.4 e 4.5, exibem números em seus eixos em que a divisão do número inteiro e da casas decimais é realizada por um ponto e não por uma vírgula. Tal fato ocorre devido ao software utilizado para a geração desse gráficos, que utiliza a formatação numérica da língua inglesa. Esse software foi utilizado pois disponibiliza recursos de customização dos gráficos mais completos.

4.2.1 Configuração das bases de dados no contexto balanceado

Para a realização dos experimentos com bases de dados balanceadas foram consideradas a variação das sete quantidades de termos em cada uma das quatro bases de dados.

A aplicação do método de validação *10-fold cross-validation* originou, para cada quantidade de termos, 10 conjuntos de dados sendo cada um deles contendo 1800 opiniões destinadas ao treinamento dos classificadores e 200 a teste. Ambas essas partes apresentam o mesmo número de opiniões pertencentes a cada uma das classes.

Sendo assim, a aplicação tanto do cálculo de frequência dos termos e TF-IDF, quanto a da técnica IG, relevantes para o treinamento do classificadores, foi realizada somente sobre as opiniões destinadas a treinamento em cada um dos 10 conjuntos.

4.2.2 Configuração das bases de dados no contexto desbalanceado

Já para os experimentos que objetivam a avaliação dos classificadores em um contexto em que os dados estão desbalanceados, considerando as quantidades de amostras nas diferentes classes, foram criados dois conjuntos de dados.

O primeiro conjunto tem como principal objetivo a avaliação do desempenho dos classificadores conforme a variação da proporção de desbalanceamento entre as duas classes. Para esses experimentos, visando um conjunto de experimentos mais compacto, somente foram consideradas as quantidade de termos em que cada um dos classificadores apresentou seu melhor resultado de acurácia sobre as bases balanceadas.

Para a definição das bases desbalanceadas desse primeiro conjunto, os mesmos conjuntos originados definidos na subseção anterior pela aplicação do *10-fold cross-validation* foram utilizados. Porém, antes da aplicação do algoritmo IG de seleção de termos, a quantidade de opiniões da classe negativa destinadas ao treinamento dos classificadores foi reduzida em diferente proporções em cada um dos 10 conjuntos.

Buscando uma melhor avaliação das técnicas de classificação, foram definidas 4 proporções de desbalanceamento entre as classes. Sendo assim, as proporções de opiniões da classe negativa (quantidade de opiniões negativas/quantidade de opiniões positivas) foram: $0,8; 0,6; 0,4$ e $0,2$, resultando na seleção de 720, 540, 360 e 180 opiniões negativas, respectivamente.

Apesar dos conjuntos de opiniões destinadas ao treinamento terem sido alterados quanto a quantidade de opiniões pertencentes à classe negativo, as opiniões utilizadas para o teste dos classificadores foram mantidas as mesmas em cada um dos 10 conjuntos. Desta forma, é possível uma melhor avaliação entre os experimentos com bases balanceadas e desbalanceadas.

Já o segundo conjunto de dados do contexto desbalanceado tem como objetivo a avaliação do desempenho dos classificadores em um cenário com uma diferença expressiva entre as quantidades de documentos de cada uma das classes conforme a variação da quantidade de termos selecionados pelo algoritmo IG.

Sendo assim, para a menor proporção de desbalanceamento, $0,2$, foram gerados conjuntos de dados para todas as quantidades de termos definidas para os conjuntos do contexto balanceado de opiniões da subseção anterior (50, 100, 500, 1000, 3000, 4000, 5000). Porém, devido ao fato da redução da quantidade de opiniões, as base de GPS e câmeras não apresentam quantidade total de termos suficiente para a geração de conjuntos para todas as quantidades de termos.

4.2.3 Configuração das bases de dados com a aplicação de *undersampling*

Além dos dois contextos descritos nas seções anteriores, considerando trabalhos da literatura, como: Li et al. (2011); Burns et al. (2011) e Mountassir, Benbrahim e Berrada (2012), torna-se interessante a avaliação dos classificadores em um contexto balanceado, porém que apresenta uma baixa quantidade de amostras para o treinamento, originado da aplicação da técnica para o tratamento do desbalanceamento denominada *undersampling* randômico. Para a realização dessa avaliação, o cenário de desbalanceamento com a menor proporção de opiniões negativas, $0,2$ foi selecionado.

Sendo assim, os conjuntos de treinamento definidos na subseção anterior para o segundo conjunto de dados que, considerando a aplicação do *10-fold cross-validation*, contêm 900 opi-

niões pertencentes à classe positiva e 180 à negativa, tiveram a quantidade de opiniões de cada classe igualada através do sorteio aleatório de 180 opiniões da classe positiva. Assim como nos conjuntos desbalanceados definidos, as opiniões destinadas a testes foram mantidas às mesmas dos conjuntos balanceados completos da seção 4.2.1.

Após a definição dos conjuntos, a técnica TF-IDF e o algoritmo de seleção de termos IG para a seleção das mesmas quantidade de termos dos conjuntos balanceados completos (50, 100, 500, 1000, 3000, 4000, 5000), também foram aplicados. Porém, assim como nos conjuntos desbalanceados de opiniões de GPS e câmeras com taxa de desbalanceamento $0,2$, as bases de GPS, livros e câmeras definidas com a aplicação do *undersampling* não apresentam quantidade total de termos suficiente para a geração dos conjuntos para todas as quantidades de termos definidas.

4.3 Resultados no Contexto Balanceado

Os três classificadores aplicados às bases balanceadas de opiniões foram avaliados considerando as métricas acurácia, *recall* e *precison*, sendo que a mensuração dessas duas últimas foi realizada considerando as classes individualmente. Além dessas métricas, também foi computado o tempo gasto por cada um dos classificadores para o treinamento e para a realização dos testes de classificação.

Os resultados apresentados nesta seção consideram a variação do número de termos selecionados para a representação das opiniões. Sendo assim, cada uma das quatro bases definidas na subseção 4.2.1 gerou sete versões diferentes da base original, uma para cada quantidade de termos definida.

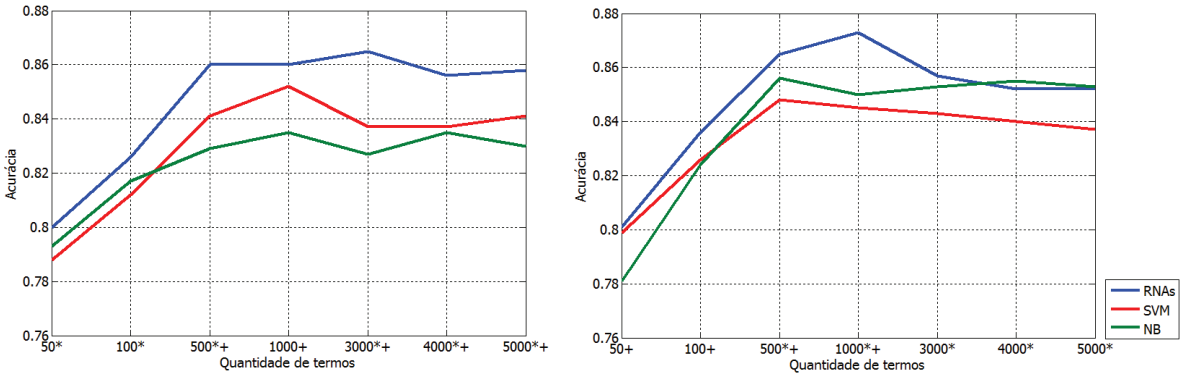
Os Apêndices F-I contêm tabelas que apresentam os valores absolutos dos resultados obtidos pela aplicação dos classificadores. Todos os valores apresentados se referem à média dos valores obtidos com a aplicação do método *cross-validation*. Nessas tabelas são utilizadas as abreviaturas POS e NEG para indicar os resultados referentes às classes positivo e negativo, respectivamente. Considerando a discussão realizada na subseção 2.6.4, a Tabela 8 contém características dos modelos de classificação das RNAs e do SVM que originaram os resultados relatados nesta seção.

Já as Figuras 12 e 13 sumarizam graficamente os resultados referentes à métrica acurácia para uma melhor comparação de desempenho de classificação. Nos gráficos apresentados nessas figuras, as quantidades de termos presentes no eixo x destacadas com os símbolos "*" e "+" representam, respectivamente, os experimentos em que a diferença de acurácia entre os classificadores SVM e RNAs e entre NB e RNAs foi estatisticamente significativa. Para a definição dessa diferença, o teste de significância *t* foi aplicado considerando uma significância de 5%. É importante destacar que os diferentes gráficos não apresentam o mesmo intervalo de taxa de acerto presente no eixo y, já que o objetivo principal é comparar o desempenho entre os classificadores e não entre as diferentes bases de dados.

Tabela 8: Média da quantidade de *support vectors* e de neurênios na camada escondida considerando o método *cross-validation*.

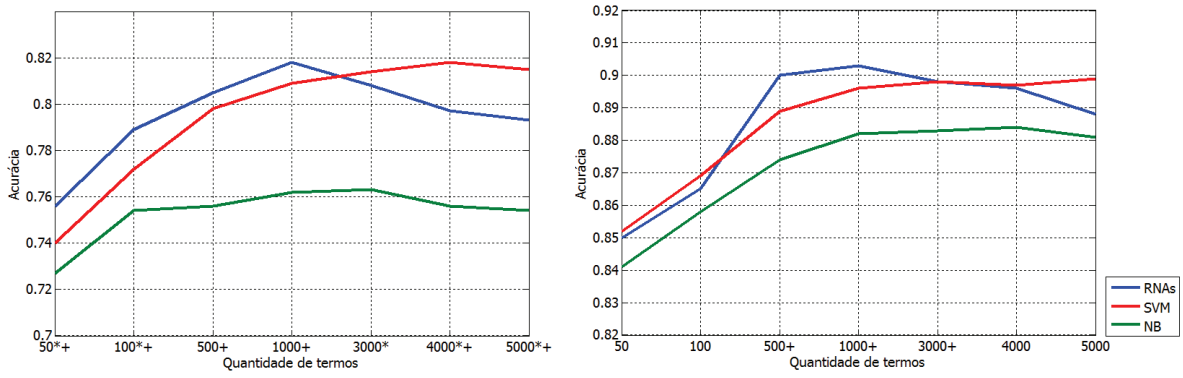
Número de termos	Média de *support vectors*	Média de neurônios escondidos
50	1097.1	30.5
100	1075.7	31.8
500	1051.3	31.8
1000	1143.1	29.4
3000	1210.0	24.7
4000	1220.8	24.2
5000	1233.1	25.6

Figura 12: Média de acurácia de classificação em função da quantidade de termos no contexto balanceado. (a) base de opiniões referentes a filmes. (b) base de opiniões referentes a GPS.



(a) (b)

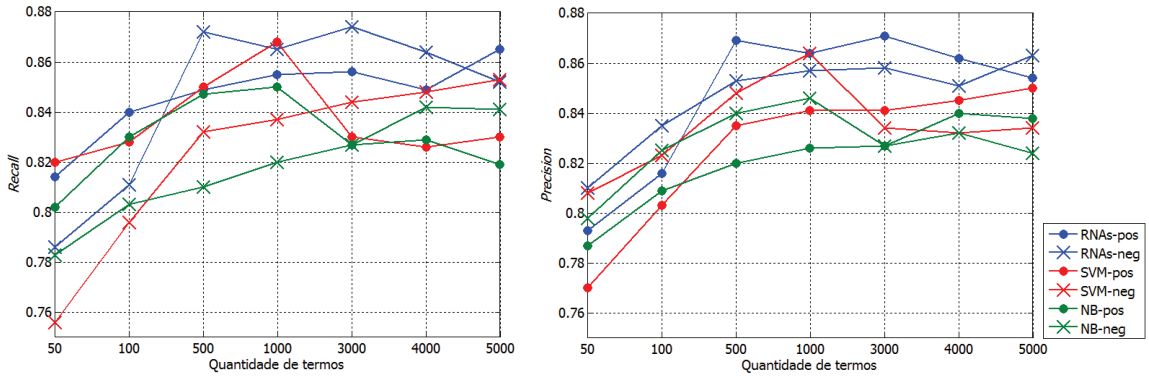
Figura 13: Média de acurácia de classificação em função da quantidade de termos no contexto balanceado. (a) base de opiniões referentes a livros. (b) base de opiniões referentes a câmeras.



(a) (b)

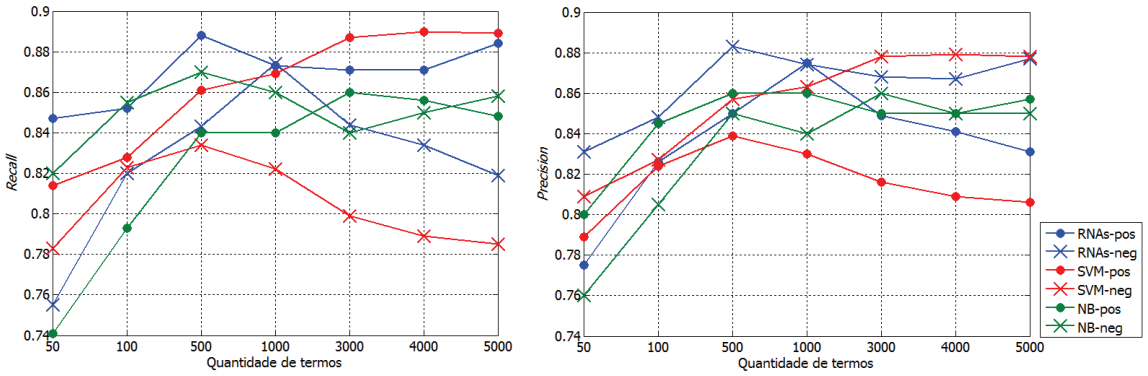
Para uma melhor avaliação dos resultados, as métricas *precision* e *recall* também foram sumarizadas graficamente nas Figuras 14 - 17 para cada um das bases. Nessas figuras, cada classificador está representado por uma cor e ambas as métricas contêm um valor para cada uma das classes que são diferenciadas pelos símbolos $\bullet$ e $\times$ para a classe positivo e negativo, respectivamente.

Figura 14: Média de *recall* (a) e *precision* (b) em função da quantidade de termos no contexto balanceado para a base de filmes.



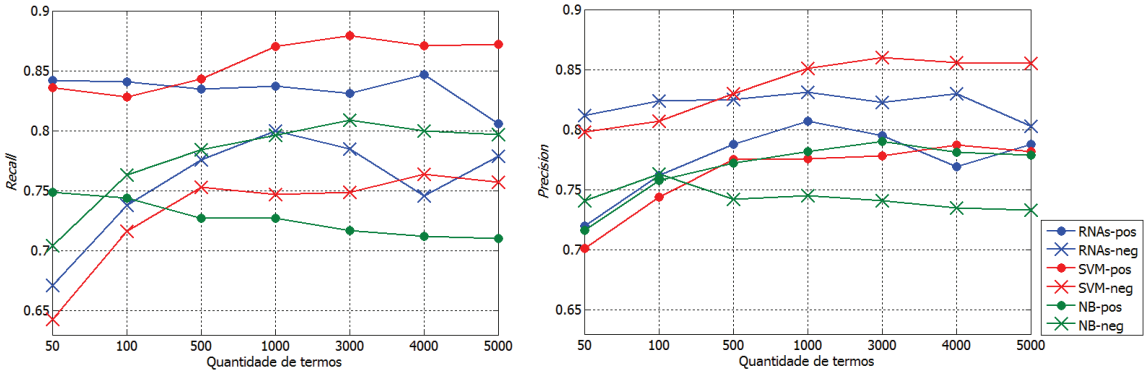
(a) (b)

Figura 15: Média de *recall* (a) e *precision* (b) em função da quantidade de termos no contexto balanceado para a base de GPS.



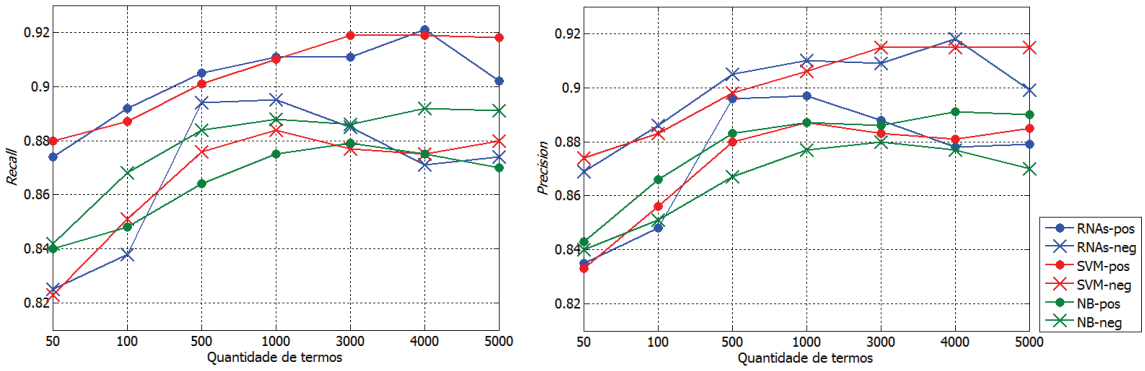
(a) (b)

Figura 16: Média de *recall* (a) e *precision* (b) em função da quantidade de termos no contexto balanceado para a base de livros.



(a) (b)

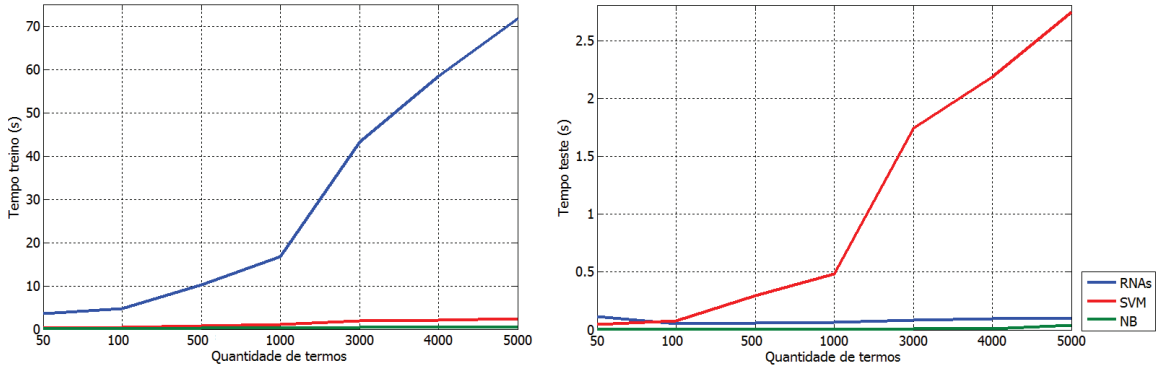
Figura 17: Média de *recall* (a) e *precision* (b) em função da quantidade de termos no contexto balanceado para a base de câmeras.



(a) (b)

Por fim, a Figura 18 ilustra graficamente o tempo gasto para treinamento e teste dos classificadores em função das diferentes quantidades de termos representativos dos textos. Como pode ser visto nos Apêndices F-I, o tempo de treinamento do classificador NB é consideravelmente menor que o dos demais, dificultando a visualização na Figura 18(a). Esse mesmo fato ocorre no tempo de teste, representado na Figura 18(b).

Figura 18: Média de tempo de execução de treino (a) e de teste (b) em função da quantidade de termos no contexto balanceado.



(a) (b)

Considerando os resultados apresentados nesta seção, alguns aspectos podem ser ressaltados:

- A seleção de mais de 1000 termos pela técnica IG não resulta em um significante aumento de desempenho dos classificares e, de maneira geral, quantidades de termos em torno de 500 e 1000 podem ser uma boa escolha para a obtenção de uma boa relação entre acurácia e tempo de execução, como pode ser visto nas Figuras 12, 13 e 18. Apesar disso, o classificador NB apresentou seus melhores resultados absolutos de acurácia com grandes quantidades de termos.

- Embora a diferença de acurácia entre os classificadores SVM e RNAs, em nenhum dos experimentos realizados ter sido superior a 3%, as RNAs tiveram melhores índices de classificação estatisticamente significantes em 13 dos 28 experimentos realizados e o SVM conseguiu superar as RNAs somente em 2 dos experimentos. Além disso, os melhores resultados de acurácia atingidos em cada uma das bases foi com a aplicação das RNAs.

- Considerando a diferença de acurácia para RNAs e o NB, o classificador RNAs foi superior com diferença estatisticamente significativa em 19 experimentos, sendo que o NB apresentou em apenas 2 experimentos uma taxa de acurácia maior que as RNAs.

- A Figura 18(a) mostra a forte relação que as RNAs apresentam entre o número de termos selecionados e seu tempo de treinamento. Porém, na Figura 18(b) essa relação entre o número de termos e tempo de execução é apresentada pelo SVM para a classificação de documentos. Já a simplicidade do algoritmo NB é observada tanto no tempo de treinamento, quanto no tempo de teste.

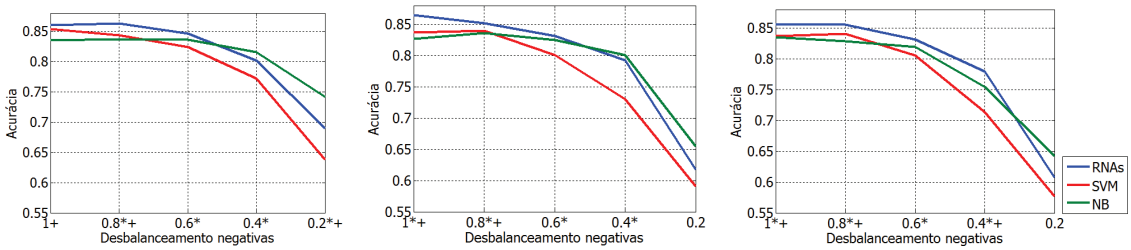
- Os resultados apresentados nas Figuras 14-17 indicam que os classificadores SVM e RNAs produzem resultados equivalentes de *recall* e *precision*, sendo que na maioria dos experimentos as taxas de *recall* obtidas são maiores para a classe positivo do que para a classe negativo. Já o classificador NB apresenta um comportamento diferenciado dos demais resultando em taxas de *recall*, para a maioria dos experimentos, mais altas para a classe negativo. Para a avaliação dessas métricas, *recall* e *precision*, deve-se considerar valores altos e valores de classes distintas próximos, o que caracteriza um modelo de classificação equilibrado entre as classes. Embora esses resultados apresentem variações nas diferentes bases, eles indicam que o classificador NB tende a ser mais equilibrado do que os demais.

4.4 Resultados no Contexto Desbalanceado

Para a avaliação dos classificadores no contexto desbalanceado foram utilizadas somente as métricas acurácia, *recall* e *precison*, visando uma análise mais geral. Esta seção apresenta os resultados dos experimentos com as bases definidas na subseção 4.2.2 de forma gráfica tanto para o primeiro, quanto para o segundo conjunto de bases.

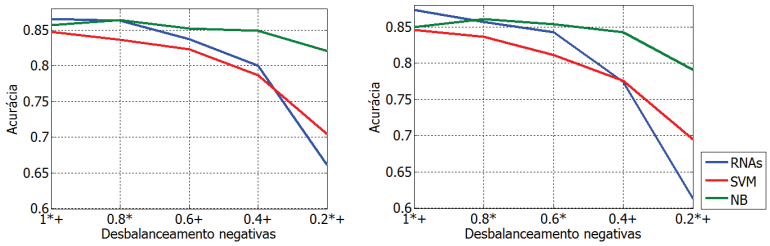
Sendo assim, as Figuras 19-22 apresentam, para cada uma das bases, os resultados de acurácia dos classificadores RNAs, SVM e NB considerando a variação da proporção do desbalanceamento nas quantidades de termos em que cada um dos classificadores apresentou seu melhor resultado de acurácia no contexto balanceado (primeiro conjunto de dados). Para uma melhor avaliação da acurácia, nesses experimentos também foi aplicado o teste de significância t. Nas Figuras 19-22 os símbolos "*" e "+" representam, respectivamente, os experimentos em que a diferença de acurácia entre os classificadores SVM e RNAs e entre NB e RNAs foi estatisticamente significativa.

Figura 19: Resultados de acurácia sobre a base de filmes no contextos desbalanceado considerando as diferentes proporções de desbalanceamento e as quantidades de termos em que (a) RNAs (1000 termos), (b) SVM (3000 termos) e (c) NB (4000 termos) apresentaram seus melhores resultados nos experimentos balanceados.



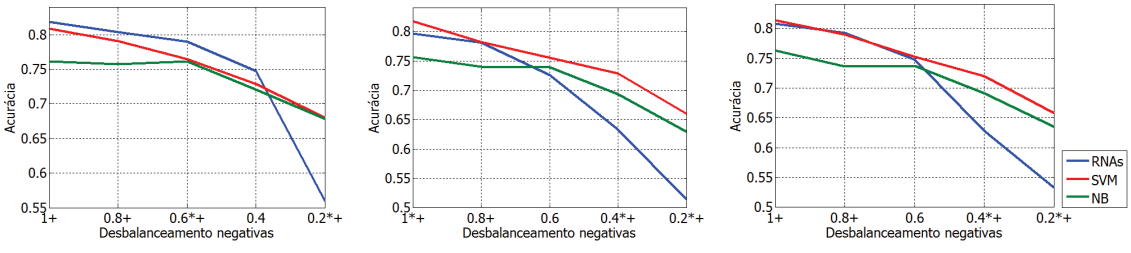
(a) (b) (c)

Figura 20: Resultados de acurácia sobre a base de GPS no contextos desbalanceado considerando as diferentes proporções de desbalanceamento e as quantidades de termos em que (a) RNAs e NB (500 termos) e (b) SVM (1000 termos) apresentaram seus melhores resultados nos experimentos balanceados.



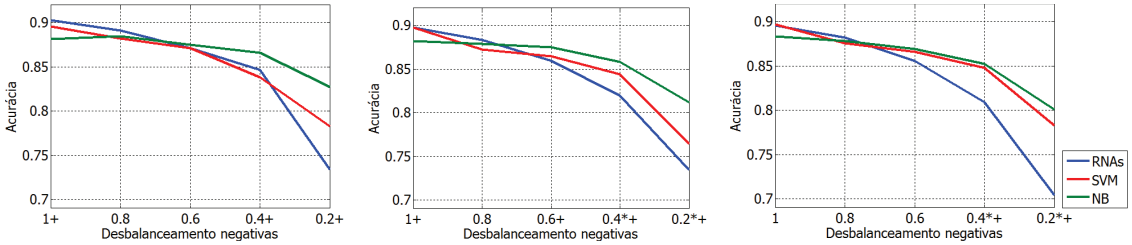
(a) (b)

Figura 21: Resultados de acurácia sobre a base de livros no contextos desbalanceado considerando as diferentes proporções de desbalanceamento e as quantidades de termos em que (a) RNAs (1000 termos), (b) SVM (4000 termos) e (c) NB (3000 termos) apresentaram seus melhores resultados nos experimentos balanceados.



(a) (b) (c)

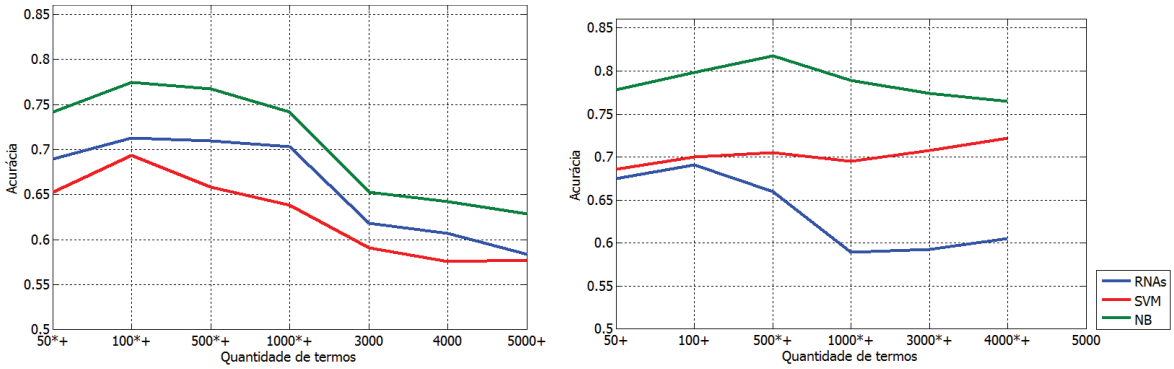
Figura 22: Resultados de acurácia sobre a base de câmeras no contextos desbalanceado considerando as diferentes proporções de desbalanceamento e as quantidades de termos em que (a) RNAs (1000 termos), (b) SVM (3000 termos) e (c) NB (4000 termos) apresentaram seus melhores resultados nos experimentos balanceados.



(a) (b) (c)

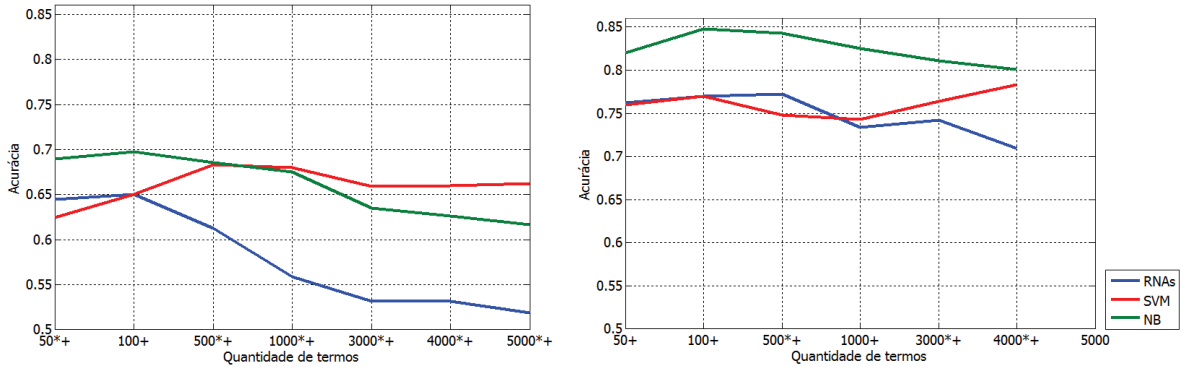
Os resultados dos experimentos com o segundo conjunto de dados do contexto desbalanceado são apresentados nas figuras seguintes, sendo a métrica acurácia representada graficamente nas Figuras 23 e 24 em função da variação da quantidade de termos e as métricas *recall* e *precision* nas Figuras 25-28 seguindo a mesma representação dos resultados dos experimentos com a base balanceada apresentados na seção anterior. As Figuras 23 e 24 também apresentam as indicações de em qual experimentos a diferença dos resultados das RNAs foi estatisticamente significativa.

Figura 23: Média de acurácia de classificação em função da quantidade de termos no contexto desbalanceado. (a) base de opiniões referentes a filmes. (b) base de opiniões referentes a GPS.



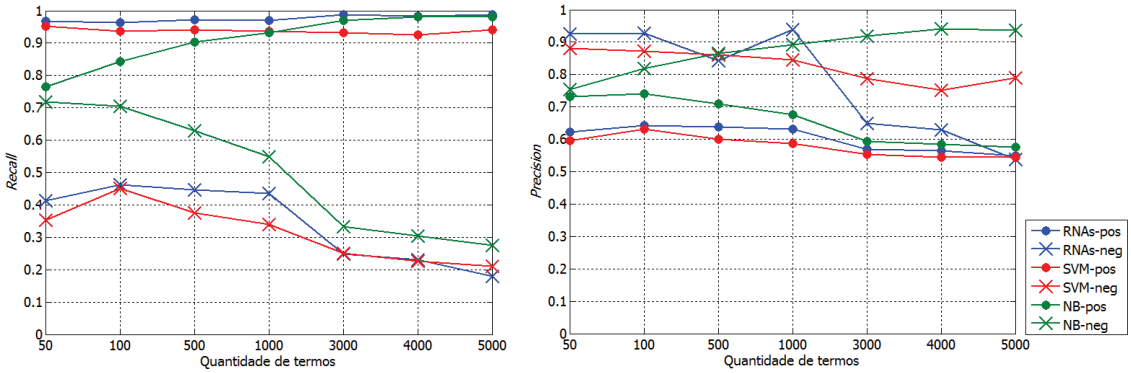
(a) (b)

Figura 24: Média de acurácia de classificação em função da quantidade de termos no contexto desbalanceado. (a) base de opiniões referentes a livros. (b) base de opiniões referentes a câmeras.



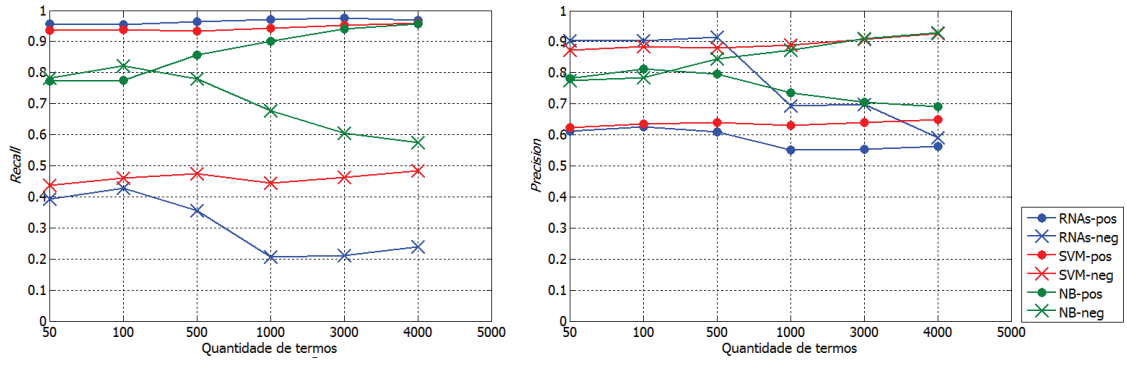
(a) (b)

Figura 25: Média de *recall* (a) e *precision* (b) em função da quantidade de termos no contexto desbalanceado para a base de filmes.



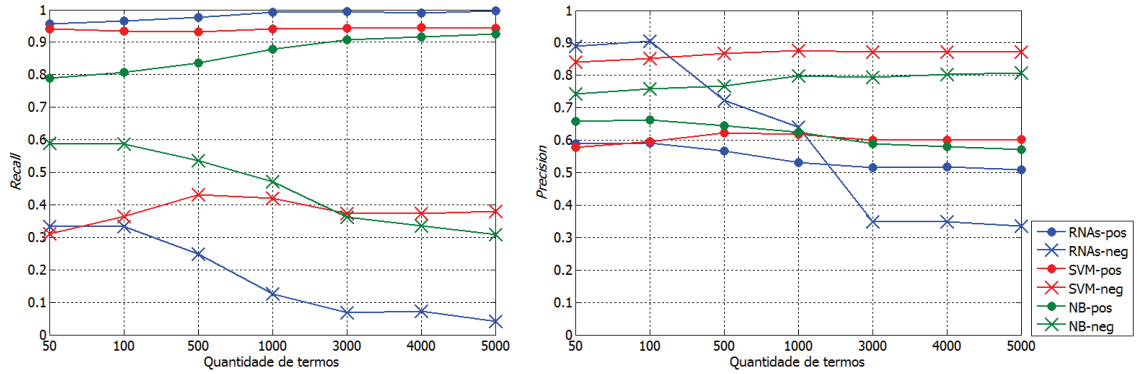
(a) (b)

Figura 26: Média de *recall* (a) e *precision* (b) em função da quantidade de termos no contexto desbalanceado para a base de GPS.



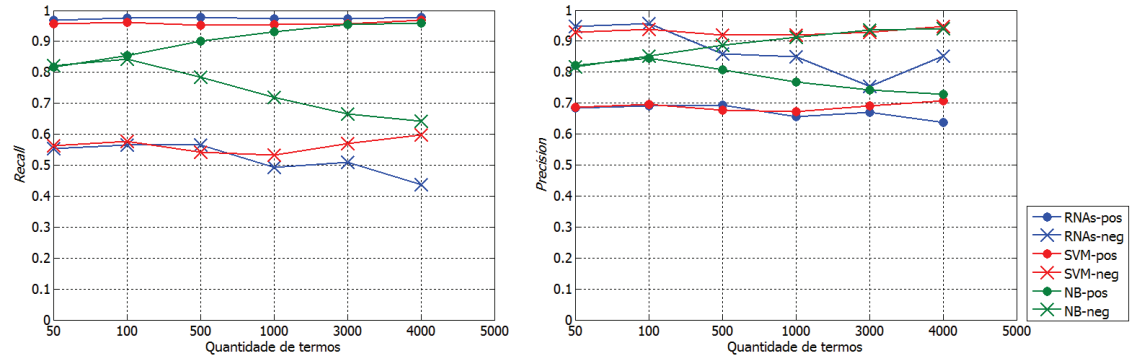
(a) (b)

Figura 27: Média de *recall* (a) e *precision* (b) em função da quantidade de termos no contexto desbalanceado para a base de livros.



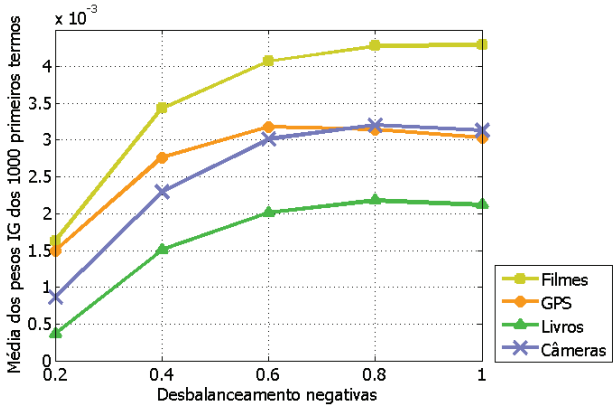
(a) (b)

Figura 28: Média de *recall* (a) e *precision* (b) em função da quantidade de termos no contexto desbalanceado para a base de câmeras.



(a) (b)

Figura 29: Média dos valores IGs dos 1000 termos selecionados em função da taxa de desbalanceamento.



A Figura 29 apresenta a média dos pesos atribuídos pelo algoritmo IG dos 1000 termos selecionados em cada um dos 10 conjuntos de treinamento para cada uma das bases. Tal informação

pode ser interpretada como um indicador da capacidade dos termos na discriminação das classes, já que, como visto na subseção 2.4.4 os termos que aparecem com uma maior frequência em somente uma das classes são melhores classificados pelo IG.

Considerando os resultados da aplicação dos classificadores no contexto desbalanceado de bases de dados, algumas questões podem ser destacadas:

- Como esperado, na grande maioria dos experimentos, os resultados de acurácia dos três classificadores diminuem juntamente com o declínio da taxa de desbalanceamento.

- O classificador RNAs é o mais sensível ao desbalanceamento e também a dados ruidosos comparados os demais classificadores aplicados. Considerando que a seleção de termos é realizada pela classificação destes de acordo com o algorítimo IG, é possível afirmar que conjuntos maiores de termos selecionados têm mais chances de conterem termos menos importantes.

- Apesar da expressiva variação do desempenho das RNAs com a alteração da taxa de desbalanceamento observada na maioria dos experimentos, nos resultados apresentados na Figura 19 é evidenciado que as RNAs conseguem ser superiores ao SVM e competitivas com o NB. Com a análise da Tabela 7, subseção 4.1, e da Figura 29, pode-se perceber que a base de filmes apresenta características mais expressivas, como uma maior média de quantidade de termos e maiores médias de pesos atribuídos aos termos pelo algorítimo IG.

- Considerando os experimentos realizados com o segundo conjunto de bases, as Figuras 23 e 24 revelam que somente em 6 deles o classificador RNAs superou com significância estatística o classificador SVM e em nenhum deles o NB quanto a acurácia. Já o SVM apresentou melhores taxas significativas em relação as RNAs em 11 experimentos, enquanto o NB superou o classificador neural em 24 experimentos.

- Com a análise das Figuras 25-28 é possível constatar que as baixas taxas de acurácia apresentadas pelas RNAs estão associadas a uma classificação priorizada com taxas de *recall* da classe positivo altas e *precision* da classe negativo altas. Por outro lado, a alta taxa de acurácia do classificador NB é resultante de taxas de *recall* e *precision* de valores absolutos mais baixos, quando comparadas com as taxas de *recall* da classe positivo e *precision* da classe negativo dos demais classificadores, mas mais equilibrados entre as classes.

- O classificador SVM também apresenta uma classificação mais balanceada entre as classes quando comparado às RNAs apresentando linhas nos gráficos pertencentes a diferentes classes mais próximas. Porém, tal fato não é tão expressivo como o que ocorre com o classificador NB.

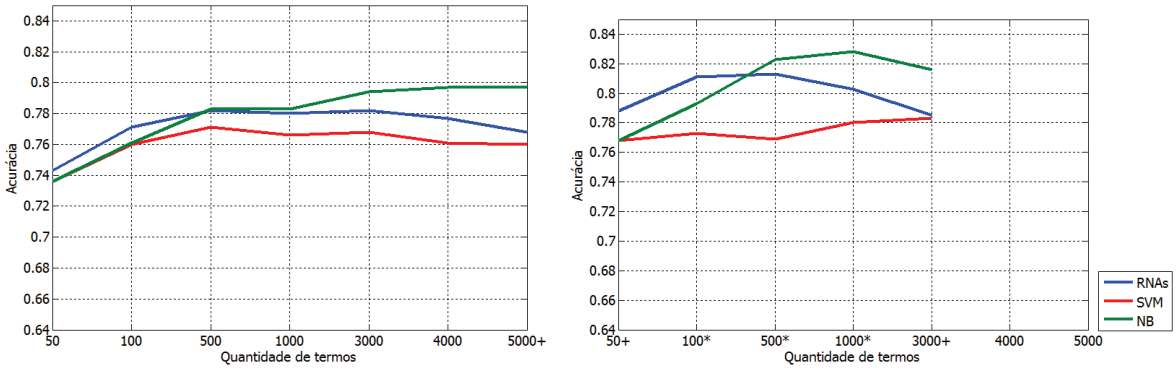
- Pode-se notar também, que as quedas bruscas da taxa de acurácia apresentadas pelo classificador neural, exibidas nas Figuras 23 e 24, estão associadas, além de experimentos com maiores quantidades de termos, com quedas da métrica *recall* da classe negativo e *precision* da classe positivo.

- Assim como as RNAs, o classificador NB também apresenta dificuldades na classificação de conjuntos com maior número de termos, apresentando classificações mais tendenciosas a classe majoritária. Já o classificador SVM contorna melhor o problema da agregação de termos mais ruidosos, apresentando uma melhor estabilidade com a variação da quantidade de termos, e em alguns casos, um pequeno aumento na taxa de acurácia para grandes quantidade de termos, mas que nem sempre são resultantes da melhora da taxa de *recall* da classe negativo.

4.5 Resultados com a Aplicação de *Undersampling*

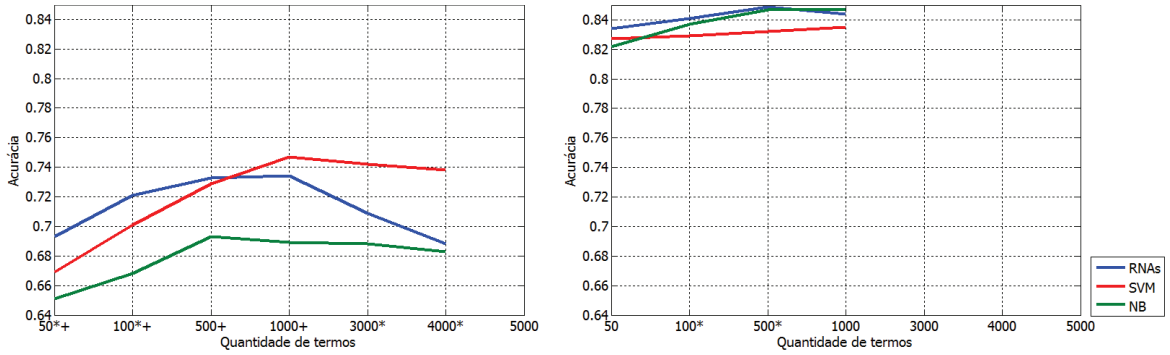
Nesta seção são apresentados os resultados dos experimentos realizados com as bases definidas na subseção 4.2.3 originadas da aplicação da técnica *undersampling* randômico. Os resultados representados na Figuras 30 e 31 se referem a taxa de acurácia em cada um das bases considerando a variação das quantidade de termos selecionados pelo algoritmo IG. Nesses experimentos, o teste de significância t também foi aplicado e as figuras também contém as indicações dos experimentos em que a diferença dos resultados das RNAs foi estatisticamente significativa.

Figura 30: Média de acurácia de classificação em função da quantidade de termos com a aplicação de *undersampling*. (a) base de opiniões referentes a filmes. (b) base de opiniões referentes a GPS.



(a) (b)

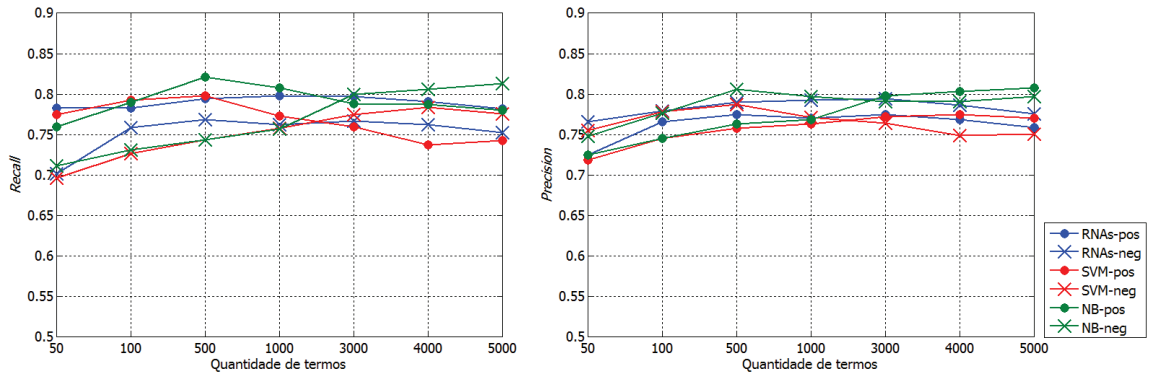
Figura 31: Média de acurácia de classificação em função da quantidade de termos com a aplicação de *undersampling*. (a) base de opiniões referentes a livros. (b) base de opiniões referentes a câmeras.



(a) (b)

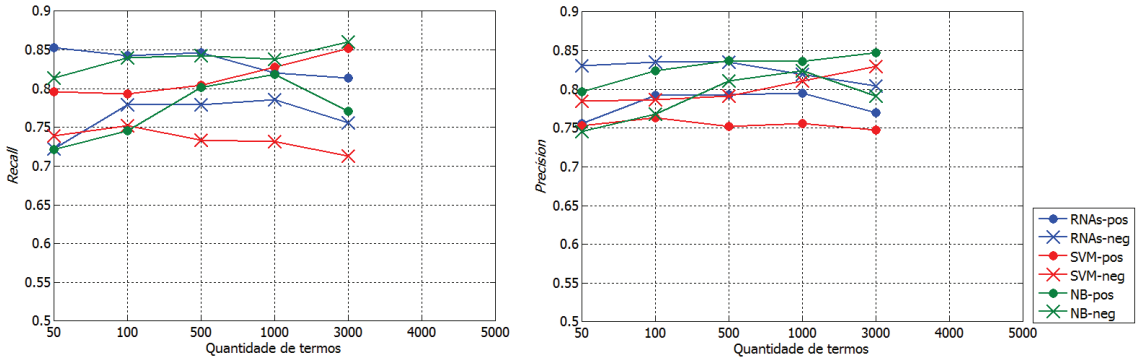
Já as figuras apresentadas a seguir, Figuras 32-35, apresentam os resultados da métricas *recall* e *precision* para cada uma das bases conforme a variação da quantidade de termos selecionados pelo algoritmo IG. Nestas figuras segue-se a mesma representação dos resultados dessas métricas dos experimentos com a base balanceada apresentados na seção 4.3 que adota linhas com marcações diferentes para cada uma das classes.

Figura 32: Média de *recall* (a) e *precision* (b) em função da quantidade de termos com a aplicação de *undersampling* na base de filmes.



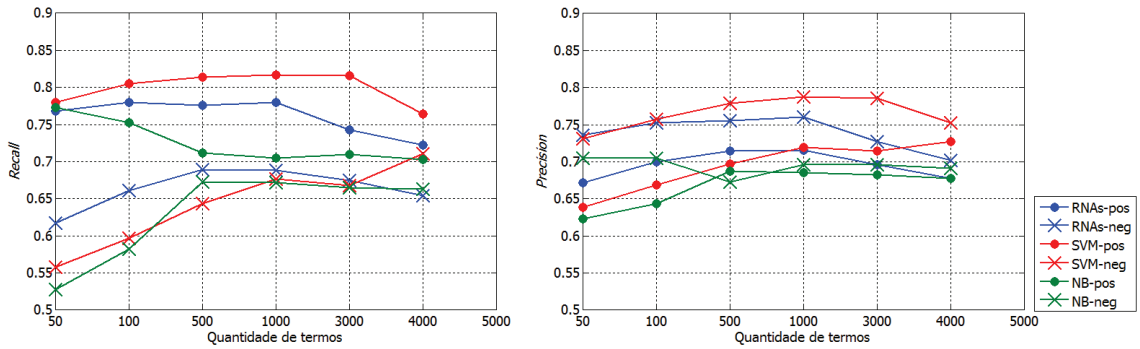
(a) (b)

Figura 33: Média de *recall* (a) e *precision* (b) em função da quantidade de termos com a aplicação de *undersampling* na base de GPS.



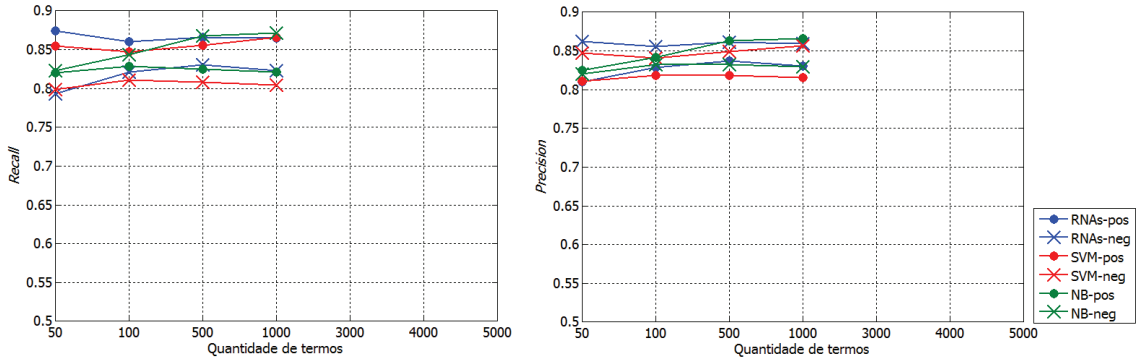
(a) (b)

Figura 34: Média de *recall* (a) e *precision* (b) em função da quantidade de termos com a aplicação de *undersampling* na base de livros.



(a) (b)

Figura 35: Média de *recall* (a) e *precision* (b) em função da quantidade de termos com a aplicação de *undersampling* na base de câmeras.



(a) (b)

Considerando esses resultados referentes aos experimentos realizados com a aplicação da técnica *undersampling*, as seguinte questões podem ser destacada:

- De maneira geral, as Figuras 30 e 31, mostram que com a aplicação do *undersampling* todos os classificadores retomam índices mais altos de acurácia comparados com os ín-

dices dos experimentos com bases desbalanceadas. Porém, esses índices ainda são mais baixos que os índices obtidos no contexto balanceado apresentados nas Figuras 12 e 13.

- Ainda considerando a acurácia, os resultados evidenciam que as RNAs voltam a ser competitivas com os demais classificadores, já que, em apenas 3 dos 22 experimentos algum outro classificador conseguiu superá-las significantemente, sendo que, em 5 experimentos elas conseguiram superar significantemente o NB e em 7 o SVM.

- Assim como nos experimentos no contexto desbalanceado, porém, de maneira menos expressiva, o aumento da quantidade de termos influencia negativamente no desempenho da RNAs, fato que também acontece no contexto balanceado mas tende a estabilizar a após uma certa quantidade. Porém, em função de que quantidades maiores de termos não puderam ser experimentadas em algumas bases, não pôde-se concluir que a estabilidade das RNAs apresentada no contexto balanceado se repete com a aplicação do *undersampling* nos dados.

- Apesar dos resultados inferiores aos do contexto balanceado, os resultados apresentados nesta seção apresentam menos variação em termos de valores absolutos em função da variação da quantidade de termos selecionados pelo algoritmo IG para os três classificadores.

- Quanto as métricas *recall* e *precision* os três classificadores apresentam comportamentos semelhantes com a variação da quantidade de termos, sendo que a classe positivo apresenta valores mais altos de *recall* e a classe negativo valores mais altos de *precision*.

4.6 Discussão de Resultados

Ainda que somente experimentos utilizando uma função de *Kernel* não-linear para o SVM tenham sido realizados, pode-se assumir que as classes do problema abordado neste trabalho não podem ser perfeitamente e linearmente separadas tendo como base os termos individuais como dimensões do espaço de representação. Os resultados deste trabalho evidenciam que o classificador SVM necessita de um grande número de *support vectors* para a classificação de documentos opiniativos como positivos ou negativos (ver Tabela 8). Tal fato acaba impactando em uma demanda de tempo de classificação muito maior que o das RNAs. Entretanto, apesar de haverem trabalhos na literatura, como Cristianini e Shawe-Taylor (2000), que afirmam que na prática o classificador SVM frequentemente seleciona um número pequeno de *support vectors*, os resultado reportados neste trabalho estão consistentes com aqueles apresentados na literatura de Mineração Textual, (COLAS et al., 2007), que relatam a construção de modelos de classificação SVM com uma grande quantidade de *support vectors*.

As RNAs ainda são raramente aplicadas ao problema de classificação de Mineração de Opiniões e Análise de Sentimentos e uma das razões para isso pode ser o excessivo custo compu-

tacional requerido para o seu treinamento em um contexto em que os dados são representados por uma alta dimensionalidade. Porém, os resultados apresentados nas seções anteriores indicam que um algoritmo simples de seleção de termos (IG) consegue refinar os dados para a composição da entrada de uma rede neural reduzindo seu custo computacional no processo de treinamento sem comprometer o desempenho de classificação correta.

Os resultados apresentados no contexto balanceado de bases de dados indicam que, com quantidades de termos acima de 1000 termos, as RNAs não somente apresentam um aumento no tempo de treinamento, mas também, nenhuma melhora significativa de desempenho quando considerada a taxa de acurácia, indicando que o algoritmo IG apresenta um desempenho satisfatório na filtragem de termos ruidosos. Em Mineração Textual, alguns trabalhos, como Gabrilovich e Markovitch (2004) e Taira e Haruno (1999), recomendam o treinamento do classificador SVM fazendo o uso de todos os atributos dos dados. Porém, os experimentos relatados revelam que o SVM também apresenta melhorias de desempenho com a seleção de termos realizada pelo IG, já que seu tempo de classificação pode ser significativamente reduzido quando menos termos são considerados na entrada do classificador.

Como pode ser visto nas Figuras 12 e 13, não é necessária a utilização de mais do que uma pequena fração do vocabulário original de cada base, menos de 10%, já que não há um aumento significativo na acurácia quando são considerados mais de 500 ou 1000 termos selecionados pelo algorítimo IG. Em síntese, embora a aplicação de alguma técnica para a redução de dimensionalidade dos dados seja crítica para tornar o treinamento das RNAs viável, essa aplicação não representa uma desvantagem para o SVM, sendo que este também é beneficiado com a redução de dimensionalidade (LI et al., 2009; DANG; ZHANG; CHEN, 2010; ABBASI et al., 2011), principalmente no contexto de Mineração de Opiniões e Análise de Sentimentos em grande escala discutido no trabalho de Bespalov et al. (2011).

A associação da aplicação do NB com o IG, principalmente em problemas de somente duas classes, poderia ocasionar uma classificação tendenciosa para uma das classes, já que ambos algoritmos se baseiam na probabilidade dos termos ocorrerem nas classes. Desta forma, se o conjunto de termos melhores classificados pelo IG conter muitos termos com altas probabilidades de ocorrerem em uma das classes, probabilidade essa utilizada pelo classificador NB (*likelihood*), a coocorrência de termos da outra classe poderia não ser suficientemente representativa para superar a dos poucos termos com alta probabilidade. Porém, este fato não foi observado em nenhum dos experimentos mesmo naqueles com apenas 50 termos.

Considerando a avaliação da convergência da RNAs para uma solução satisfatória, optou-se pelo treinamento de três modelos com a mesma configuração de estrutura, número de neurônios e suas disposições nas camadas, porém com diferentes pesos iniciais. Os resultados obtidos com um modelo neural são referentes ao modelo, dentre os três, que apresentou a maior taxa de acurácia. Tal fato ocasiona uma significante desvantagem em comparação ao treinamento do classificador SVM, já que seu método de otimização sempre converge para uma única solução. Apesar disso, há trabalhos que objetivam o estudo de técnicas que permitem a divisão

do processamento realizado pelas RNAs e a execução das partes em paralelo, como propõem Atakulreka e Sutivong (2007). Além disso, se o tempo para a realização da classificação for um fator mais interessante que o tempo de treinamento, um cenário comum na aplicação de soluções, a escolha de um modelo neural pode ser a melhor opção quando esses apresentarem o melhor desempenho de classificação correta, já que já que as RNAs apresentaram tempos de classificação bem abaixo daqueles obtidos com a aplicação do SVM

Apesar das vantagens das RNAs apontadas anteriormente, os resultados no contexto desbalanceado de base de dados indicam que o SVM, e principalmente o NB, são mais estáveis que as RNAs conforme há a inserção de termos ruidosos e o desbalanceamento fique maior. Visto que, pode-se assumir que as bases de livros, GPS e câmeras contêm termos mais ruidosos do que a base de filmes clássica da literatura, o comportamento da acurácia de classificação em função do desbalanceamento das bases evidencia que o desempenho da RNAs é significativamente inferior aos dos demais classificadores, principalmente quando mais termos são considerados para a representação dos documentos.

Considerando o classificador NB, embora seus resultados no contexto balanceado tenham sido significantemente inferiores aos das RNAs, os resultados dos experimentos no contexto desbalanceado revelam que esse classificador apresenta a surpreendente capacidade de estabilidade, considerando a variação da proporção de desbalanceamento da base, sendo superior inclusive ao SVM, em grande parte dos experimentos, que, assim como as RNAs, apresenta um algorítimo muito mais sofisticado e complexo do que o NB, apresentados na seção 2.6.

Os altos índices de acurácia observados na aplicação do NB no contexto desbalanceado são consequência de uma classificação menos tendenciosa à classe majoritária, equilibrando assim as métricas *recall* e *precision* entre às classes, mesmo não apresentando os maiores valores absolutos delas. Além disso, os índices mais altos obtidos pelo classificador foram em experimentos com quantidades menores de termos, revelando que a seleção de poucos termos gerada pelo algoritmo IG pode selecionar termos que apresentam uma probabilidade grande de ocorrência na classe negativa, facilitando a classificação do NB, característica que não consegue ser captada pelos outros classificadores. Em relação ao tempo de treinamento e classificação, em ambos o NB foi o classificador que apresentou os menores valores, consequência do baixo custo computacional requerido pelo seu método estatístico em ambas as fases, treinamento e classificação.

Com a aplicação da técnica de *undersampling*, mesmo que não tenha sido realizada mais de uma escolha aleatória das opiniões da classe negativo mantidas no treinamento dos classificadores, as RNAs se mostraram mais uma vez competitivas apresentando melhores resultados que os demais classificadores e repetindo algumas características observadas no contexto balanceado. Porém, os fatores tempo de treinamento, tempo de classificação e a redução da vantagem de desempenho apresentada, fazem com que o classificador neural não seja o mais indicado neste contexto, já que a vantagem de desempenho de classificação não justifica a escolha do modelo neural quando considerado o tempo de treinamento muito maior que os dos demais

classificadores

Embora a aplicação do *undersampling* tenha sido benéfica para o desempenho dos três classificadores, a aplicação desta técnica deve ser avaliada conforme as particularidades de desbalanceamento de cada base. As baixas quantidades de amostras utilizadas para o treinamento dos classificadores resultou em taxas de classificação consideravelmente mais baixas que aquelas obtidas com as bases balanceadas completas, demonstrando que esses conjuntos menores de amostras não foram capazes de representar, ou então, salientar, todos os padrões discriminativos das classes, já que as opiniões utilizadas para testes foram as mesmas nos experimentos do contexto balanceado e com *undersampling*. Além disso, outra questão a ser ressaltada quanto a baixa quantidade de amostras, é que, em alguns casos, as quantidades de termos em que os classificadores apresentaram melhores taxas de acurácia no contexto balanceado, não puderam ser experimentadas.

Apesar das questões negativas apontadas, em todos os experimentos realizados com bases em que a técnica de *undersampling* foi aplicada, as taxas de *recall* da classe minoritária do contexto desbalanceado, a negativo, foram superiores aos dos experimentos com as bases desbalanceadas. Porém, ressalta-se que, como visto nos resultados deste trabalho, a proporção de desbalanceamento das bases exercem grande influência no desempenho dos classificadores e, sendo assim, é válida uma avaliação prévia quanto a aplicação do *undersampling* para a verificação se, com a taxa de desbalanceamento apresentada pela base desejada, os classificadores realmente apresentarão um melhor desempenho.

5 CONCLUSÃO

Os classificadores *Support Vector Machines* (SVM) e *Naïve Bayes* (NB) vêm sendo amplamente aplicados na literatura de Mineração de Opiniões e Análise de Sentimentos, enquanto abordagens que utilizam o classificador Redes Neurais Artificiais (RNA) são pouco exploradas para o aprendizado de máquina na identificação da polaridade de sentimentos em dados textuais. Em outras áreas, comparações entre os classificadores SVM e RNAs são realizadas, mas não há um consenso claro de qual dos dois classificadores apresenta o melhor desempenho. Desta forma, os estudos apontam que o desempenho de ambos é fortemente relacionado com o problema e o contexto ao qual os dados pertencem. Considerando a literatura revista neste trabalho, uma investigação comparativa realizada sobre os mesmos dados e a representação destes, entre RNAs e classificadores clássicos da área é desconhecida. Desta forma, neste trabalho é apresentado um estudo empírico comparativo entre a aplicação de RNAs e os classificadores NB e SVM para o problema de classificação de polaridade de opiniões.

Os classificadores aplicados foram comparados objetivando principalmente o melhor desempenho de classificação correta. Com os experimentos realizados foram avaliados os modelos de classificação com a representação dos dados realizada através da modelagem *bag-of-words* considerando unigramas. Em relação à literatura, as principais contribuições resultantes da realização deste trabalho são:

- Em termos de taxa de acerto de classificação, as RNAs conseguiram superar com significância o classificador SVM em grande parte dos experimentos realizados com a base clássica da literatura de Pang e Lee (2004) considerando tanto o contexto balanceado como o desbalanceado de bases de dados.

- Considerando os experimentos que avaliaram os classificadores no contexto balanceado, em 13 dos 28 experimentos realizados as RNAs superaram significantemente o classificador SVM, enquanto o SVM superou o classificador neural em apenas 2 deles.

- Nestes mesmos experimentos, o classificador RNAs superou com uma diferença estatisticamente significativa em 19 experimentos o classificador NB sendo que, este último, também apresentou uma taxa de acurácia maior que as RNAs somente em 2 experimentos.

- Contudo, os experimentos também revelam que os classificadores SVM e, principalmente o NB, tendem a ser menos afetados com a presença de ruídos na representação dos dados do que as RNAs quando a quantidade de documentos em uma das classes diminui em relação a da outra.

- Como esperado, o tempo de treinamento do classificador neural é consideravelmente maior que os dos demais classificadores. Porém, considerando a tarefa de classificação as RNAs se mostram competitivas, já que apresentam valores de tempo consideravelmente

mais baixos que os do SVM e, dependendo do desbalanceamento da base, valores maiores de taxa de acerto do que o NB.

- A técnica *Information Gain*, que apresenta um algoritmo de baixo custo computacional para a seleção de termos, pôde ser utilizada para a redução da complexidade do problema sem afetar de forma significante nenhum dos três classificadores testados. Considerando o aumento do número de termos de entrada para os classificadores, os resultados obtidos indicam que há limiares de quantidade de termos que, acima deles, são pequenas as melhorias de taxa de classificação correta que podem ser atingidas. Além disso, eles ainda evidenciam que o IG (i) torna viável a aplicação do treinamento das RNAs considerando a abordagem de representação unigramas e (ii) contribui também para a redução da complexidade da tarefa de classificação do SVM, embora esta complexidade não seja necessariamente dependente do número de variáveis de entrada do SVM (JOACHIMS, 1998; SUYKENS; VANDEWALLE; MOOR, 2001).

- A aplicação da técnica *undersampling* para o tratamento do desbalanceamento de bases de dados é uma alternativa para a obtenção de modelos de classificação menos tendenciosos à classe majoritária, mesmo que a pequena quantidade de dados tende à interferir negativamente no aprendizado dos classificadores. Desta forma, tal aplicação deve ser avaliada, já que em bases que não apresentam uma grande diferença na quantidade de amostras pertencentes às diferentes classes, alguns classificadores podem apresentar taxas de acerto de classificação acima das do cenário em que o *undersampling* é aplicado.

- Além dessas contribuições, parte deste trabalho foi publicado na revista *Expert System with Application* com o título *Document-level sentiment classification: An empirical comparison between SVM and ANN* (MORAES; VALIATI; GAVIãO NETO, 2013). O trabalho aqui apresentado se difere da publicação por utilizar um algoritmo diferente do classificador NB, incluir uma comparação mais detalhada do desempenho dos classificadores no cenário com a menor proporção de desbalanceamento e também, realizar uma investigação do comportamento dos classificadores em um cenário cujo a técnica *undersampling* é aplicada às bases de dados desbalanceadas.

Em resumo, os resultados indicam que o classificador RNAs é melhor que os demais avaliados quando as bases de dados utilizadas para o treinamento apresentarem pouca ou nenhuma diferença na quantidade de amostras pertencentes às diferentes classes já que nestes contextos, as RNAs se mostraram, na maior parte dos experimentos, serem estatisticamente superiores quanto a taxa de classificação correta.

Considerando estas conclusões, a discussão dos resultados apresentadas na seção 4.6 e os estudos da literatura citados e apresentados neste trabalho, como possíveis extensões deste estudo oito aspectos são apontados:

- Um estudo comparativo da aplicação de RNAs utilizando-se outras abordagens de representação das opiniões, como *Part-of-Speech* ou *Joint Sentiment-Topic* utilizada por He, Lin e Alani (2011).

- O classificador ME tem mostrado resultados promissores quando aplicado à Mineração de Opiniões e Análise de Sentimentos, sendo assim, um objetivo futuro é a incorporação de resultados deste classificador a este estudo.

- Os resultados dos experimentos que avaliaram os classificadores quanto à variação da proporção de desbalanceamento das bases de dados sugerem que há uma relação entre a quantidade de dados e o fato das RNAs serem superiores ao SVM. Deste modo, torna-se interessante a investigação de um algoritmo de seleção de termos, como o IG, para a predição do melhor classificador a ser utilizado em determinado conjunto de dados.

- Considerando que não somente o desbalanceamento entre as classes, mas também que a quantidade de opiniões pode influenciar no treinamento dos classificadores, um estudo considerando diversas quantidades de termos, quantidades de documentos de treinamento e diferentes taxas de desbalanceamento pode revelar características particulares dos classificadores e questões a serem consideradas para a escolha de um deles ou para a aplicação de técnicas de tratamento do desbalanceamento de dados.

- Embora que a literatura aponte que técnicas de *undersampling* apresentam melhores resultados no tratamento do desbalanceamento das bases, estudos que realizam aplicação de técnicas mais elaboradas de *oversampling*, como a SMOTE (CHAWLA et al., 2002), não são encontrados na literatura. Sendo assim, a aplicação de técnicas mais robustas do que a *oversampling* randômica para a aplicação de *oversampling* se mostra promissora ao problema de Mineração de Opiniões e Análise de Sentimentos.

- Apesar do aumento de dimensionalidade do problema, as conclusões de Bespalov et al. (2011), de que modelagens de opiniões através de expressões formadas por mais de uma palavra (bigramas, trigramas, etc.) podem resultar em melhores resultados de classificação correta, um estudo comparativos entre os classificadores considerando este fato e a aplicação de técnicas de seleção de termos, na tentativa de reduzir os efeitos negativos do aumento da dimensionalidade, se mostra relevante.

- Considerando a existência de diversos tipos de Redes Neurais Artificiais (HAYKIN, 2001), outra sequencia a ser dada a este trabalho é a realização de uma investigação e aplicação de um modelo neural de classificação, além do MLP, mais apropriado ao tratamento de dados de grande dimensionalidade, ruidosos e esparsos.

- Outro fator não investigado neste trabalho, mas que o estudo proporcionado por ele mostra ser relevante na comparação entre os classificadores RNAs e SVM, é a utilização da

mesma função para a definição da função de ativação dos neurônios das RNAs e da função de *Kernel* utilizada no SVM. A comparação realizada na Tabela 2 mostra que estes dois componentes dos classificadores são equivalentes e, sendo assim, quando ambos são definidos considerando a mesma função, a diferença entre o desempenho dos dois classificadores pode reduzir. Neste sentido, tal investigação objetiva definir se o mérito de desempenho de classificação de cada um dos classificadores está na forma com que a informação é processada como um todo, estrutura de cada um deles, ou especificamente como a otimização de cada um deles é realizada.

REFERÊNCIAS

ABBASI, A. Intelligent Feature Selection for Opinion Classification. **IEEE Intelligent Systems**, [S.l.], v. 25, n. 4, p. 75–79, July-Aug 2010.

ABBASI, A.; FRANCE, S.; ZHANG, Z.; CHEN, H. Selecting Attributes for Sentiment Classification Using Feature Relation Networks. **IEEE Transactions on Knowledge and Data Engineering**, [S.l.], v. 23, n. 3, p. 447 –462, march 2011.

ALPAYDIN, E. **Introduction to Machine Learning**. 2nd. ed. [S.l.]: The MIT Press, 2010.

ATAKULREKA, A.; SUTIVONG, D. Avoiding local minima in feedforward neural networks by simultaneous learning. In: AUSTRALIAN JOINT CONFERENCE ON ADVANCES IN ARTIFICIAL INTELLIGENCE, 2007. **Proceedings...** [S.l.: s.n.], 2007. p. 100–109.

BERGER, A. L.; PIETRA, V. J. D.; PIETRA, S. A. D. A maximum entropy approach to natural language processing. **Comput. Linguist.**, Cambridge, MA, USA, v. 22, n. 1, p. 39–71, Mar. 1996.

BESPALOV, D.; BAI, B.; QI, Y.; SHOKOUFANDEH, A. Sentiment classification based on supervised latent n-gram analysis. In: CIKM'11, 2011. **Anais...** [S.l.: s.n.], 2011. p. 375–382.

BISHOP, C. M. **Pattern Recognition and Machine Learning**. 2. ed.. ed. [S.l.]: Springer, 2007.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. **J. Mach. Learn. Res.**, [S.l.], v. 3, p. 993–1022, Mar. 2003.

BLITZER, J.; DREDZE, M.; PEREIRA, F. Biographies, Bollywood, Boom-boxes and Blenders: domain adaptation for sentiment classification. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 2007. **Proceedings...** [S.l.: s.n.], 2007. p. 440–447.

BURGES, C. J. A Tutorial on Support Vector Machines for Pattern Recognition. **Data Mining and Knowledge Discovery**, [S.l.], v. 2, p. 121–167, 1998.

BURNS, N.; BI, Y.; WANG, H.; ANDERSON, T. Sentiment Analysis of Customer Reviews: balanced versus unbalanced datasets. In: **Knowledge-Based and Intelligent Information and Engineering Systems**. [S.l.]: Springer Berlin / Heidelberg, 2011. v. 6881, p. 161–170.

CHAKRABARTI, S. **Mining the Web**: discovering knowledge from hypertext data. [S.l.]: Morgan-Kauffman, 2002.

CHANG, C.-C.; LIN, C.-J. LIBSVM: a library for support vector machines. **ACM Transactions on Intelligent Systems and Technology**, [S.l.], v. 2, p. 27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

CHAOVALIT, P.; ZHOU, L. Movie Review Mining: a comparison between supervised and unsupervised classification approaches. **Hawaii International Conference on System Sciences**, Los Alamitos, CA, USA, v. 4, 2005.

CHAWLA, N. V. Data Mining for Imbalanced Datasets: an overview. In: MAIMON, O.; ROKACH, L. (Ed.). **Data Mining and Knowledge Discovery Handbook**. [S.l.]: Springer US, 2010. p. 875–886.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. SMOTE: synthetic minority over-sampling technique. **Journal of Artificial Intelligence Research**, [S.l.], v. 16, p. 321–357, 2002.

CHEN, L.-S.; LIU, C.-H.; CHIU, H.-J. A neural network based approach for sentiment classification in the blogosphere. **Journal of Informetrics**, [S.l.], v. 5, n. 2, p. 313 – 322, 2011.

CHURCH, K. W.; HANKS, P. Word association norms, mutual information, and lexicography. **Comput. Linguist.**, Cambridge, MA, USA, v. 16, p. 22–29, March 1990.

COLAS, F.; PACLÍK, P.; KOK, J. N.; BRAZDIL, P. Does SVM really scale up to large bag of words feature spaces? In: INTELLIGENT DATA ANALYSIS, 7., 2007, Berlin, Heidelberg. **Proceedings...** Springer-Verlag, 2007. p. 296–307.

CRISTIANINI, N.; SHAWE-TAYLOR, J. **An Introduction to Support Vector Machines and Other Kernel-based Learning Methods**. 1. ed. [S.l.]: Cambridge University Press, 2000. 204 p.

DANG, Y.; ZHANG, Y.; CHEN, H. A Lexicon-Enhanced Method for Sentiment Classification: an experiment on online product reviews. **IEEE Intelligent Systems**, [S.l.], v. 25, p. 46–53, 2010.

DANTAS, E. R. G.; ALMEIDA, J. C.; LIMA, D. S.; AZEVEDO, R. R. O Uso da Descoberta de Conhecimento em Base de Dados para Apoiar a Tomada de Decisões. In: SEGET – SIMPóSIO DE EXCELêNCIA EM GESTãO E TECNOLOGIA, 2008. **Anais...** [S.l.: s.n.], 2008.

DASGUPTA, S.; NG, V. Mine the Easy, Classify the Hard: a semi-supervised approach to automatic sentiment classification. In: ACL-IJCNLP 2009: PROCEEDINGS OF THE MAIN CONFERENCE, 2009. **Anais...** [S.l.: s.n.], 2009. p. 701–709.

DEERWESTER, S.; DUMAIS, S. T.; FURNAS, G. W.; LANDAUER, T. K.; HARSHMAN, R. Indexing by latent semantic analysis. **JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE**, [S.l.], v. 41, p. 391–407, 1990.

DUA, S.; DU, X. **Data Mining and Machine Learning in Cybersecurity**. [S.l.]: Taylor & Francis, 2011.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P.; UTHURUSAMY, R. (Ed.). **Advances in Knowledge Discovery and Data Mining**. [S.l.]: AAAI/MIT Press, 1996.

FELDMAN, R.; SANGER, J. **The Text Mining Handbook**: advanced approaches in analyzing unstructured data. [S.l.]: Cambridge University Press, 2007.

FORMAN, G. An extensive empirical study of feature selection metrics for text classification. **Journal of Machine Learning Research**, [S.l.], v. 3, p. 1289–1305, 2003.

FREEMAN, J. A.; SKAPURA, D. M. **Neural networks**: algorithms, applications, and programming techniques. Redwood City, CA, USA: Addison Wesley Longman Publishing Co., Inc., 1991.

GABRILOVICH, E.; MARKOVITCH, S. Text Categorization with Many Redundant Features: using aggressive feature selection to make svms competitive with c4.5. In: THE TWENTY-FIRST INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 2004, Banff, Alberta, Canada. **Proceedings...** Morgan Kaufmann, 2004. p. 321–328.

GONCALVES, T.; SILVA, C.; QUARESMA, P.; VIEIRA, R. Analyzing Part-of-Speech for Portuguese Text Classification. In: COMPUTATIONAL LINGUISTICS AND INTELLIGENT TEXT PROCESSING (CICLING), 2006. **Proceedings...** [S.l.: s.n.], 2006.

HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. H. The WEKA Data Mining Software: an update. **SIGKDD Explorations**, [S.l.], v. 11, 2009.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining**: concepts and techniques, second edition (the morgan kaufmann series in data management systems). [S.l.]: Morgan Kaufmann, 2006. 800 p.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. H. **The elements of statistical learning**: data mining, inference, and prediction: with 200 full-color illustrations. [S.l.]: New York: Springer-Verlag, 2001. 533 p.

HAYKIN, S. **Redes neurais** : princípios e prática. 2. ed.. ed. Porto Alegre: Bookman, 2001.

HE, Y.; LIN, C.; ALANI, H. Automatically extracting polarity-bearing topics for cross-domain sentiment classification. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: HUMAN LANGUAGE TECHNOLOGIES, 49., 2011. **Anais...** [S.l.: s.n.], 2011.

HEARST, M. A. Text-based intelligent systems. In: JACOBS, P. S. (Ed.). **Text-based intelligent systems**. Hillsdale, NJ, USA: L. Erlbaum Associates Inc., 1992. p. 257–274.

HEARST, M. A. Untangling text data mining. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS ON COMPUTATIONAL LINGUISTICS, 37., 1999, Stroudsburg, PA, USA. **Proceedings...** Association for Computational Linguistics, 1999. p. 3–10. (ACL '99).

JOACHIMS, T. Text Categorization with Suport Vector Machines: learning with many relevant features. In: EUROPEAN CONFERENCE ON MACHINE LEARNING, 10., 1998, London, UK, UK. **Proceedings...** Springer-Verlag, 1998. p. 137–142. (ECML '98).

JUSZCZAK, P.; DUIN, R. P. W. Uncertainty sampling methods for one-class classifiers. In: IN PROCEEDINGS OF THE ICML'03 WORKSHOP ON LEARNING FROM IMBALANCED DATA SETS, 2003. SIGKDD EXPLORATIONS. VOLUME 6, ISSUE 1 - PAGE 5, 2003. **Anais...** [S.l.: s.n.], 2003.

KOTSIANTIS, S.; KANELLOPOULOS, D.; PINTELAS, P. Handling imbalanced datasets: a review. **GESTS International Transactions on Computer Science and Engineering**, [S.l.], 2006.

KUBAT, M.; MATWIN, S. Addressing the Curse of Imbalanced Training Sets: one-sided selection. In: IN PROCEEDINGS OF THE FOURTEENTH INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 1997. **Anais...** Morgan Kaufmann, 1997. p. 179–186.

LAL, T.; CHAPELLE, O.; WESTON, J.; ELISSEEFF, A. Embedded Methods. In: GUYON, I.; NIKRAVESH, M.; GUNN, S.; ZADEH, L. (Ed.). **Feature Extraction**. [S.l.]: Springer Berlin / Heidelberg, 2006. p. 137–165. (Studies in Fuzziness and Soft Computing, v. 207).

LARSON, R.; FARBER, B. **Estatística Aplicada**. 4. ed. [S.l.]: Pearson, 2010. 638 p.

LI, S.; HUANG, C.-R.; ZHOU, G.; LEE, S. Y. M. Employing personal/impersonal views in supervised and semi-supervised sentiment classification. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 48., 2010, Stroudsburg, PA, USA. **Proceedings...** Association for Computational Linguistics, 2010. p. 414–423. (ACL '10).

LI, S.; WANG, Z.; ZHOU, G.; LEE, S. Y. M. Semi-supervised Learning for Imbalanced Sentiment Classification. In: INT. JOINT CONF. ON ARTIFICIAL INTELLIGENCE, 2011. **Proceedings...** [S.l.: s.n.], 2011. p. 1826–1831.

LI, S.; XIA, R.; ZONG, C.; HUANG, C.-R. A framework of feature selection methods for text categorization. In: ANNUAL MEETING OF THE ACL, 47., 2009. **Proceedings...** [S.l.: s.n.], 2009. p. 692–700.

LIN, C.; HE, Y. Joint sentiment?topic model for sentiment analysis. In: THE 18TH ACM CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT (CIKM), 2009. **Anais...** [S.l.: s.n.], 2009.

LING, C.; SHENG, V. Cost-Sensitive Learning and the Class Imbalance Problem. In: _____. **Encyclopedia of Machine Learning**. [S.l.]: Springer, 2008.

LIU, B. **Web data mining** : exploring hyperlinks, contents, and usage data. 2nd ed.. ed. New York: Springer, 2011.

LIU, B. **Sentiment Analysis and Opinion Mining**. [S.l.]: Morgan and Claypool Publishers, 2012. 165 p.

LOVINS, J. B. Development of a Stemming Algorithm. **Mechanical Translation and Computational Linguistic**, [S.l.], p. 22–31, 1968.

MANNING, C. D.; RAGHAVAN, P.; SCHTZE, H. **Introduction to Information Retrieval**. New York, NY, USA: Cambridge University Press, 2008.

MANNING, C. D.; SCHüTZE, H. **Foundations of statistical natural language processing**. Cambridge, MA, USA: MIT Press, 1999.

MCCALLUM, A.; NIGAM, K. A comparison of event models for Naive Bayes text classification. In: IN AAAI-98 WORKSHOP ON LEARNING FOR TEXT CATEGORIZATION, 1998. **Anais...** AAAI Press, 1998. p. 41–48.

MLADENIC, D.; GROBELNIK, M. Feature selection for unbalanced class distribution and Naive Bayes. In: IN PROCEEDINGS OF THE 16TH INTERNATIONAL CONFERENCE ON MACHINE LEARNING (ICML, 1999. **Anais...** Morgan Kaufmann Publishers, 1999. p. 258–267.

MüLLER, M. F. A scaled conjugate gradient algorithm for fast supervised learning. **Neural Networks**, [S.l.], v. 6, n. 4, p. 525 – 533, 1993.

MORAES, R.; VALIATI, J. a. F.; GAVIãO NETO, W. P. Document-level sentiment classification: an empirical comparison between svm and ann. **Expert Syst. Appl.**, [S.l.], v. 40, n. 2, p. 621–633, Feb. 2013.

MOUNTASSIR, A.; BENBRAHIM, H.; BERRADA, I. An empirical study to address the problem of Unbalanced Data Sets in sentiment classification. In: SYSTEMS, MAN, AND CYBERNETICS (SMC), 2012 IEEE INTERNATIONAL CONFERENCE ON, 2012. **Anais...** [S.l.: s.n.], 2012. p. 3298–3303.

O'KEEFE, T.; KOPRINSKA, I. Feature Selection and Weighting Methods in Sentiment Analysis. In: AUSTRALASIAN DOCUMENT COMPUTING SYMPOSIUM, 2009. **Proceedings...** [S.l.: s.n.], 2009. p. 67–74.

PALTOGLOU, G.; THELWALL, M. A study of information retrieval weighting schemes for sentiment analysis. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 48., 2010, Stroudsburg, PA, USA. **Proceedings...** Association for Computational Linguistics, 2010. p. 1386–1395. (ACL '10).

PANG, B.; LEE, L. A Sentimental Education: sentiment analysis using subjectivity summarization based on minimum cuts. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 2004. **Proceedings...** [S.l.: s.n.], 2004. p. 271–278.

PANG, B.; LEE, L. Opinion Mining and Sentiment Analysis. **Foundations and Trends in Information Retrieval**, [S.l.], v. 2, n. 1-2, p. 1–135, 2008.

PANG, B.; LEE, L.; VAITHYANATHAN, S. Thumbs up? Sentiment Classification using Machine Learning Techniques. In: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, 2002., 2002. **Proceedings...** [S.l.: s.n.], 2002. p. 79–86.

PORTER, M. An algorithm for suffix stripping. **Program**, [S.l.], v. 14, n. 3, p. 130–137, July 1980.

PORTER, M. **Snowball**: a language for stemming algorithms. [S.l.: s.n.], 2001.

RIFKIN, R. M. **Everything Old Is New Again**: a fresh look at historical approaches in machine learning. 2002. Dissertação (Mestrado em Ciência da Computação) — Massachusetts Institute of Technology - Sloan School of Management, 2002.

ROMERO, E.; ALQUÉZAR, R. Comparing error minimized extreme learning machines and support vector sequential feed-forward neural networks. **Neural Netw.**, Oxford, UK, UK, v. 25, p. 122–129, 2012.

ROMERO, E.; TOPPO, D. Comparing Support Vector Machines and Feedforward Neural Networks With Similar Hidden-Layer Weights. **IEEE Transactions on Neural Networks**, [S.l.], v. 18, n. 3, p. 959–963, may 2007.

ROSENBLATT, F. **Principles of Neurodynamics**. [S.l.]: Spartan Books, 1962. 841–842 p. v. 11, n. 5.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning internal representations by error propagation. In: _____. **Parallel distributed processing**: explorations in the microstructure of cognition, vol. 1: foundations. Cambridge, MA, USA: MIT Press, 1986. p. 318–362.

RUSSELL, S. J.; NORVIG, P. **Artificial Intelligence**: a modern approach. 3rd. ed. [S.l.]: Prentice Hall, 2009.

SALTON, G.; ALLAN, J. Text Retrieval Using the Vector Processing Model. In: IN PROCEEDING OF THIRD SYMPOSIUM ON DOCUMENT ANALYSIS AND INFORMATION RETRIEVAL, 1994. **Anais...** [S.l.: s.n.], 1994.

SEMOLINI, R. **Support vector machines, inferencia transdutiva e o problema de classificação**. 2002. Dissertação (Mestrado em Ciência da Computação) — Universidade Federal do Rio de Janeiro, 2002.

SULLIVAN, D. **The need for text mining in business intelligence**. [S.l.: s.n.], 2000.

SUYKENS, J.; VANDEWALLE, J.; MOOR, B. D. Optimal Control by Least Squares Support Vector Machines. **Neural Networks**, [S.l.], v. 14, p. 23–35, 2001.

TAIRA, H.; HARUNO, M. Feature selection in SVM text categorization. In: AAAI CONFERENCE ON ARTIFICIAL INTELLIGENCE, 1999. **Proceedings...** [S.l.: s.n.], 1999. p. 480–486.

TAN, A. Text Mining: the state of the art and the challenges. In: IN PROCEEDINGS OF THE PAKDD 1999 WORKSHOP ON KNOWLEDGE DISOCOVERY FROM ADVANCED DATABASES, 1999. **Anais...** [S.l.: s.n.], 1999. p. 65–70.

TANG, H.; TAN, S.; CHENG, X. A survey on sentiment detection of reviews. **Expert Syst. Appl.**, Tarrytown, NY, USA, p. 10760–10773, September 2009.

TURNEY, P. D. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In: ANNUAL MEETING ON ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 40., 2002, Stroudsburg, PA, USA. **Proceedings...** Association for Computational Linguistics, 2002. p. 417–424. (ACL '02).

TURNEY, P. D.; LITTMAN, M. L. Measuring praise and criticism: inference of semantic orientation from association. **ACM Trans. Inf. Syst.**, New York, NY, USA, p. 315–346, October 2003.

VAN HULSE, J.; KHOSHGOFTAAR, T. M.; NAPOLITANO, A. Experimental perspectives on learning from imbalanced data. In: MACHINE LEARNING, 24., 2007, New York, NY, USA. **Proceedings...** ACM, 2007. p. 935–942. (ICML '07).

WEISS, G. M.; PROVOST, F. Learning when Training Data are Costly: the effect of class distribution on tree induction. **JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH**, [S.l.], v. 19, p. 315–354, 2003.

WEISS, S. M.; INDURKHYA, N.; ZHANG, T. **Text Mining. Predictive Methods for Analyzing Unstructured Information**. 1. ed. [S.l.]: Springer, Berlin, 2004. 248 p.

WERBOS, P. **Beyond Regression**: new tools for prediction and analysis in the behavioral sciences. 1974. Tese (Doutorado em Ciência da Computação) — Harvard University, Cambridge, MA, 1974.

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data Mining**: practical machine learning tools and techniques. 3. ed. Burlington, MA: Morgan Kaufmann, 2011.

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

YANG, Y.; PEDERSEN, J. O. A Comparative Study on Feature Selection in Text Categorization. In: FOURTEENTH INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 1997, San Francisco, CA, USA. **Proceedings...** Morgan Kaufmann Publishers Inc., 1997. p. 412–420. (ICML '97).

YU, J.; ZHA, Z.-J.; WANG, M.; CHUA, T.-S. Aspect Ranking: identifying important product aspects from online consumer reviews. In: ACL'11, 2011. **Anais...** [S.l.: s.n.], 2011. p. 1496–1505.

ZEIMPEKIS, D.; GALLOPOULOS, E. **TMG**: a matlab toolbox for generating term-document matrices from text collections. 2005.

ZURADA, J. **Introduction to artificial neural systems**. St. Paul, MN, USA: West Publishing Co., 1992.

APÊNDICE A EXEMPLOS DE OPINIÕES DE FILMES

- this film is extraordinarily horrendous and i'm not going to waste any more words on it
- claire danes , giovanni ribisi , and omar epps make a likable trio of protagonists , but they're just about the only palatable element of the mod squad , a lame-brained big-screen version of the 70s tv show . the story has all the originality of a block of wood (well , it would if you could decipher it) , the characters are all blank slates , and scott silver's perfunctory action sequences are as cliched as they come . by sheer force of talent , the three actors wring marginal enjoyment from the proceedings whenever they're on screen , but the mod squad is just a second-rate action picture with a first-rate cast
- kolya is one of the richest films i've seen in some time . zdenek sverak plays a confirmed old bachelor (who's likely to remain so) , who finds his life as a czech cellist increasingly impacted by the five-year old boy that he's taking care of . though it ends rather abruptly– and i'm whining , 'cause i wanted to spend more time with these characters– the acting , writing , and production values are as high as , if not higher than , comparable american dramas . this father-and-son delight– sverak also wrote the script , while his son , jan , directed– won a golden globe for best foreign language film and , a couple days after i saw it , walked away an oscar in czech and russian , with english subtitles
- this sometimes-tedious and often-moving documentary charts the life and times of anne frank , the young diarist and most-famous victim of adolph hitler . writer/director/producer jon blair has collected a staggering amount of historical material on both anne and the frank family . we meet miep gies , one of the family's protectors who is still alive . she recounts how she found the diary in the days after the germans captured the franks . we watch otto frank , anne's father and surviving family member , in interview footage filmed before his death . blair successfully combines these clips , footage , and other historical records to recount exactly what happened during that terrible period of european history . as narrated by kenneth branagh and with diary excepts read by glenn close , anne frank remembered retells more than just anne's story . we meet and learn about the * many * friends , family members , and acquaintances whose lives were touched by this young woman and her writings . winner of the last year's academy award for best documentary.

APÊNDICE B EXEMPLOS DE OPINIÕES DE GPS

- the tom tom appeared to charge but it wouldn't turn on. frustrating! but to amazon's credit, returning it was pain-free.
- while this gps system has a great screen and does well on giving you instructions as you are driving, i have found the maps to be out of date in numerous instances. i live in the charlotte nc area, and a loop which was built around charlotte at least two years ago isn't even acklowledged by the system. it kept telling me to turn where roads didn't even exist any longer and couldn't direct me to a shopping center which has been in existence for at least 5 years. more updating of maps would be a plus for those of us who are directionally challenged.
- this is the third tomtom gps i have owned and the best by far! has all the features and the lifetime map updates are fantastic
- i bought a garmin for myself 2 yrs ago which i have had no problems with..so it was easy to trust garmin when my parents needed a nav system. i recently used the new garmin (my parents garmin) on a trip to southern california. it performed very well. i especially liked the real time traffic delay notices which were overall right on point. the lane assist feature is a great addition and very useful. this new garmin calculates routes much faster then my old one. got it for $120 on amazon. well worth the price

APÊNDICE C EXEMPLOS DE OPINIÕES DE LIVROS

- i wasn't liking the book in general, it was going to slow for me at some parts and it was just way too depressing for me, seeing katniss's world through her prospective while everything was crashing down around her. too many people were dying, too many tears were shed. it's not something i want to read during my stressful life. when i finished the book, i was madder than i ever have been. i hated how the book ended, i hated the story. the first two books were absolutley amazing, but i have to say suzanne collins murdered the entire series with that book.

- this book is overrated. if you have common sense and a good set of values and can think for yourself, you don't need this book. on the other hand, if you are a follower who needs a role model to spoon feed you, this book is as good as any other self help book.

- this series is so fantastic that i can't seem to stop thinking about it. start with the first book (the hunger games) and you'll be hooked! i am over 30 and enjoyed this book immensely.

- this book is one of dozens of books i have read in the past two years regarding self-improving, balancing and organizing life, improving and understanding finances. i would say that this book is a much more general and umbrella-like book. it was one of the most inspiring books i read because it really motivated me and helped me realize so many things in my life. it will make you aware of everything in your life from your time-management, priorities and relationships with people in your life. i highly recommend anyone, especially if you are young, to read this book because it will help you look at the overall picture of your life and reanalyze your priorities. it certainly has affected my life for the better. i created my own mission statement for myself and i continously read it from time to time to help keep me balanced. finances, health, relationships, job and personal growth–this book will make you realize that they are all Part of one big picture that makes up your life. the 7 habits help you become more efficient in all aspects of your life and live a more productive and enjoyable life.

APÊNDICE D EXEMPLOS DE OPINIÕES DE CÂMERAS

- it was almost impossible to open and close the battery/memory card section. i have returned it hoping for a refund.
- i purchased this camera after what i thought was fairly extensive research to find a good replacement for my wife's aging canon powershot sd600. i especially wanted something equally portable with a more powerful zoom, and the panasonic dmc-fh20 provides both. apparently i did not do enough research, however, at least with respect to the quality of indoor or other low-light pictures, with or without flash. which, in a word, is terrible. shots in bright sunlight are great – sharp, with good color. but everything i have tried indoors has produced grainy, noisy images with unnatural color, no matter what image size or iso i select. if i had wanted a pocket-sized camera solely for outdoor, full sunlight photography, this would have been a good choice. otherwise, it is a serious disappointment.
- i love this camera, very easy to use...the pictures have come out great. the handy size makes it convenient to carry while out and about.
- for what i've used the camera for so far, i'm very pleased. the one thing i've had a problem with is they don't give enough instructions on how to set up the different keys so i kept dumping out good pictures that i wanted to keep. i "think"i have that part figured out now. i did think i'd be able to take pictures when it was at sunset because of the extra low light feature but they didn't come out as well as i had hoped and i even tried using the flash with it. but overall, i would recommend this camera to others.

APÊNDICE E STOPWORDS UTILIZADAS NO TRABALHO

Tabela 9: *Stopwords* utilizadas na etapa de pré-processamento.

a	about	above	according	across	actually	after
again	against	all	almost	along	already	also
although	always	am	among	an	and	another
any	anything	are	aren	aren't	as	at
away	b	back	back	be	because	been
before	behind	being	below	besides	better	between
beyond	both	but	by	c	can	cannot
can't	certain	class	could	couldn't	d	did
didn't	do	does	doesn't	doing	don't	down
during	e	each	else	enough	even	ever
f	few	for	from	further	g	get
going	got	great	h	had	hadn't	has
hasn't	have	haven't	having	he	he'd	he'll
her	here	here's	hers	herself	he's	high
him	himself	his	how	however	how's	i
i'd	if	i'll	i'm	in	instead	into
is	isn't	it	its	it's	itself	i've
j	just	k	l	later	least	less
less	let	let's	little	m	many	may
maybe	me	might	more	most	much	must
mustn't	my	myself	n	neither	never	new
no	non	nor	not	nothing	o	of
off	often	old	on	once	one	only
or	other	ought	our	ours	ourselves	out
over	own	p	perhaps	put	q	r
rather	really	s	same	set	several	shan't
she	she'd	she'll	she's	should	shouldn't	since
snot	snt	so	some	something	sometimes	soon
still	such	t	than	that	that's	the
their	theirs	them	themselves	then	there	therefore
there's	these	they	they'd	they'll	they're	they've
thing	this	those	though	three	through	till
to	together	too	toward	towards	two	u
under	until	up	upon	us	v	very
very	w	was	wasn't	we	we'd	we'll
were	we're	weren't	we've	what	what's	when
when's	where	where's	whether	which	while	who
whole	whom	who's	whose	why	why's	will
with	within	without	won't	would	wouldn't	x
y	yet	you	you'd	you'll	your	you're
yours	yourself	yourselves	you've	z		

APÊNDICE F TABELA DE RESULTADOS DA BASE DE FILMES NO CONTEXTO BALANCEADO

Tabela 10: Média dos resultados considerando o método *cross-validation* para a base de opiniões referentes a filmes no contexto balanceado.

Classificador	Número de termos						
	50	100	500	1000	3000	4000	5000
Acurácia							
RNAs	80%	82,5%	86%	86%	**86,5%**	85,6%	85,8%
SVM	78,8%	82,6%	84,1%	85,2%	83,7%	83,7%	84,1%
NB	79,3%	81,7%	82,9%	83,5%	82,7%	83,5%	83,0%
Tempo de treinamento							
RNAs	2,3s	3,2s	6,7s	12,9s	40,3s	51,5s	65,5s
SVM	0,27s	0,63s	1,24s	2,2s	4,3s	4,75s	5,6s
NB	**0,018s**	0,02s	0,05s	0,07s	0,15s	0,18s	0,21s
Recall (POS : NEG)							
RNAs	0,81 : 0,79	0,84 : 0,81	0,85 : **0,87**	0,85 : 0,86	0,86 : **0,87**	0,85 : 0,86	0,86 : 0,85
SVM	0,82 : 0,76	0,83 : 0,82	0,85 : 0,83	**0,87** : 0,84	0,83 : 0,84	0,83 : 0,85	0,83 : 0,85
NB	0,8 : 0,78	0,83 : 0,8	0,85 : 0,81	0,85 : 0,82	0,83 : 0,83	0,83 : 0,84	0,82 : 0,84
Precision (POS : NEG)							
RNAs	0,79 : 0,81	0,82 : 0,83	**0,87** : 0,85	0,86 : **0,86**	**0,87** : **0,86**	0,86 : 0,85	0,85 : **0,86**
SVM	0,77 : 0,81	0,82 : 0,83	0,83 : 0,85	0,84 : **0,86**	0,84 : 0,83	0,84 : 0,83	0,85 : 0,83
NB	0,78 : 0,8	0,81 : 0,83	0,82 : 0,84	0,82 : 0,84	0,83 : 0,83	0,84 : 0,83	0,84 : 0,82

APÊNDICE G TABELA DE RESULTADOS DA BASE DE GPS NO CONTEXTO BALANCEADO

Tabela 11: Média dos resultados considerando o método *cross-validation* para a base de opiniões referentes a GPS no contexto balanceado.

Classificador	Número de termos						
	50	100	500	1000	3000	4000	5000
Acurácia							
RNAs	80,1%	83,6%	86,5%	**87,3%**	85,7%	85,2%	85,2%
SVM	79,8%	83,2%	84,7%	84,5%	84,3%	83,9%	83,7%
NB	78,1%	82,4%	85,6%	85,0%	85,3%	85,5%	85,3%
Tempo de treinamento							
RNAs	3,1s	4,6s	11,1s	14,5s	45,9s	64,5s	75,4s
SVM	0,2s	0,5s	0,6s	0,8s	1,1s	1,2s	1,3s
NB	**0,02s**	**0,02s**	0,04s	0,06s	0,12s	0,16s	0,19s
Recall (POS : NEG)							
RNAs	0,85 : 0,75	0,85 : 0,82	**0,89** : 0,84	0,87 : **0,87**	0,87 : 0,84	0,87 : 0,83	0,88 : 0,82
SVM	0,81 : 0,78	0,84 : 0,83	0,86 : 0,83	0,87 : 0,82	**0,89** : 0,8	**0,89** : 0,79	**0,89** : 0,78
NB	0,74 : 0,82	0,79 : 0,85	0,84 : **0,87**	0,84 : 0,86	0,86 : 0,84	0,86 : 0,85	0,85 : 0,86
Precision (POS : NEG)							
RNAs	0,77 : 0,83	0,83 : 0,85	0,85 : **0,88**	**0,87** : 0,87	0,85 : 0,87	0,84 : 0,87	0,83 : **0,88**
SVM	0,79 : 0,81	0,83 : 0,84	0,84 : 0,86	0,83 : 0,86	0,82 : **0,88**	0,81 : **0,88**	0,81 : **0,88**
NB	0,8 : 0,76	0,85 : 0,80	0,86 : 0,85	0,86 : 0,84	0,85 : 0,86	0,85 : 0,85	0,86 : 0,85

APÊNDICE H TABELA DE RESULTADOS DA BASE DE LIVROS NO CONTEXTO BALANCEADO

Tabela 12: Média dos resultados considerando o método *cross-validation* para a base de opiniões referentes a livros no contexto balanceado.

Classificador	Número de termos						
	50	100	500	1000	3000	4000	5000
Acurácia							
RNAs	75,6%	78,9%	80,5%	**81,8%**	80,8%	79,6%	79,2%
SVM	73,9%	78,4%	79,8%	80,9%	81,4%	81,7%	81,4%
NB	72,7%	75,4%	75,6%	76,2%	76,3%	75,6%	75,4%
Tempo de treinamento							
RNAs	3,7s	5,5s	12,1s	21,3s	42s	63,9s	69,4s
SVM	0,22s	0,39s	0,54s	0,78s	1,2s	1,4s	1,5s
NB	**0,01s**	0,02s	0,04s	0,05s	0,12s	0,15s	0,2s
Recall (POS : NEG)							
RNAs	0,84 : 0,67	0,84 : 0,74	0,83 : 0,78	0,84 : **0,8**	0,83 : 0,78	0,85 : 0,75	0,81 : 0,78
SVM	0,84 : 0,64	0,83 : 0,73	0,84 : 0,75	0,87 : 0,75	**0,88** : 0,75	0,87 : 0,76	0,87 : 0,76
NB	0,75 : 0,7	0,74 : 0,76	0,73 : 0,78	0,73 : 0,79	0,72 : 0,81	0,71 : 0,8	0,71 : 0,8
Precision (POS : NEG)							
RNAs	0,72 : 0,81	0,76 : 0,82	0,79 : 0,82	**0,81** : 0,83	0,79 : 0,82	0,77 : 0,83	0,79 : 0,8
SVM	0,7 : 0,8	0,76 : 0,81	0,77 : 0,83	0,78 : 0,85	0,78 : **0,86**	0,79 : **0,86**	0,78 : 0,85
NB	0,72 : 0,74	0,76 : 0,76	0,77 : 0,74	0,78 : 0,75	0,79 : 0,74	0,78 : 0,74	0,78 : 0,73

APÊNDICE I TABELA DE RESULTADOS DA BASE DE CÂMERAS NO CONTEXTO BALANCEADO

Tabela 13: Média dos resultados considerando o método *cross-validation* para a base de opiniões referentes a câmeras no contexto balanceado.

Classificador	Número de termos						
	50	100	500	1000	3000	4000	5000
Acurácia							
RNAs	84,9%	86,5%	89,9%	**90,3%**	89,8%	89,6%	88,8%
SVM	85,1%	88,0%	88,8%	89,6%	89,8%	89,7%	89,9%
NB	84,1%	85,8%	87,4%	88,2%	88,3%	88,4%	88,1%
Tempo de treinamento							
RNAs	4,5s	5,6s	11s	18,8s	45,2s	54,6s	77,2s
SVM	0,2s	0,3s	0,4s	0,6s	0,9s	1s	1,1s
NB	**0,02s**	**0,02s**	0,03s	0,06s	0,12s	0,15s	0,2s
Recall (POS : NEG)							
RNAs	0,87 : 0,82	0,89 : 0,84	0,9 : **0,89**	0,91 : **0,89**	0,91 : 0,88	**0,92** : 0,87	0,9 : 0,87
SVM	0,88 : 0,82	0,89 : 0,87	0,9 : 0,88	0,91 : 0,88	**0,92** : 0,88	**0,92** : 0,87	**0,92** : 0,88
NB	0,84 : 0,84	0,85 : 0,87	0,86 : 0,88	0,87 : **0,89**	0,88 : 0,88	0,87 : **0,89**	0,87 : **0,89**
Precision (POS : NEG)							
RNAs	0,83 : 0,87	0,85 : 0,89	**0,9** : 0,9	**0,9** : 0,91	0,89 : 0,91	0,88 : **0,92**	0,8 : 0,9
SVM	0,83 : 0,87	0,87 : 0,89	0,88 : 0,9	0,89 : 0,91	0,88 : 0,91	0,88 : 0,91	0,88 : 0,91
NB	0,84 : 0,84	0,86 : 0,85	0,88 : 0,86	0,88 : 0,87	0,88 : 0,88	0,89 : 0,87	0,89 : 0,87