UNIVERSIDADE DE ÉVORA

ESCOLA DE CIÊNCIAS E TECNOLOGIA

DEPARTAMENTO DE INFORMÁTICA

Estudo sobre Análise de Sentimentos em Textos

Mestrado em Engenharia Informática

Orientação: Henriques Fernando

Orientador: Irene Pimenta Rodrigues

Dissertação

Évora, 2013

Mestrado em Engenharia Informática

Dissertação

Estudo sobre Análise de Sentimentos em Textos

Henriques Fernando

Orientadora
Irene Pimenta Rodrigues

Sumário

O sentimento em opiniões terceiras sempre foi e continua a despertar interesse e extrema preocupação por parte dos gestores ou tomadores de decisões. Terceiros podem ser indivíduos, produtos, entidades empresariais, instituições de pesquisa e órgãos governamentais etc. dado que força de expressões e ideias controversas podem causar grandes celeumas. Desta forma o *feedback* emocional pode ser propulsor de mudanças, no sentido de proporcionar a busca contínua de melhorias por um lado ou determinar por outro o insucesso da entidade. Logo a análise de sentimentos é uma ferramenta indispensável no apoio ao processo de tomada de decisões. No processo de análise de sentimentos focamos na classificação de polaridade de opiniões em textos ou documentos em termos positiva ou negativa. Uma gama de Aplicações e recursos actualmente está alinhada à Língua Inglesa, o que mostra a existência de poucas experiências noutras Línguas. O objectivo é desenvolver um protótipo como ferramenta de auxílio a análise de opiniões em textos, auxiliado pelas técnicas de Aprendizado em Automática e Processamento em Linguagem Natural.

Na realização de experiências aplicamos a Classificação Supervisionado a dois Corpus nomeadamente *Review Movie e SentiCorpus-pt*, que contém textos com opiniões sobre diversos filmes e políticos Portugueses participantes de um debate eleitoral respectivamente. A metodologia aplicada baseia-se na classificação de padrões linguísticos tais como *PosTag, Chunking* e outras formas simples de negação. Para a melhorar a classificação determinamos a orientação semântica das palavras, desde os seus recursos léxicos através do *SentiWordnet Sense*, que é uma ferramenta em Inglês que faz a tradução de qualquer língua para o Inglês, antes de extracção de polaridade. A nossa abordagem é avaliar os dois corpus. Tarefa que é auxiliada pelo casamento de termos linguísticos com vista a melhorar a performance. O classificador utiliza para tal o modelo Balsa de Palavras como linha de base.

Palavras-chave: Análise de Sentimentos, Mineração de Textos, Classificação de Polaridade, Aprendizado Supervisionado, Características Linguísticas, Orientação Semântica, Aprendizado em Máquina.

Abstract

The feeling for third parties opinions has always been and it continues to arouse interest and extreme concern by managers or decision makers, third parties can be individuals, products, business organizations, research institutions and government institutions etc.. Knowing that expressions force and controversial ideas can cause major embarrassment. In this way, the emotional feedback can be change propeller, in order to provide continuous search for improvements on one side or to determine on the other hands the failure of the entity. So, the sentiment analysis is an indispensable tool to support the process of decision making. In the process of sentiment analysis we have focused on the classification of polarity of opinions in texts or documents in positive or negative terms . A range of applications and resources nowadays are aligned with English Language, which demonstrates the existence of few experiences in other languages . The aim is to develop a prototype as a supporting tool to analyse texts opinions, supported by learning techniques on Machine and Natural Language Processing.

When conducting experiments, we have applied the Supervised Classification into two Corpus namely *Review Movie and SentiCorpus-pt*, which contains texts with opinions of different films and Portuguese politicians participants in electoral debate respectively. The applied methodology is based on the classification of linguistic patterns such as Pos-tag, Chunking is other simple form of denial. To improve the classification and determine the semantic orientation of words, from the lexicon resources through SentiWordnet Sense, which is a tool in English that makes the translation from any language to English before identifying the polarity. Our approach is to evaluate the two corpus. Task that is supported by the marriage of linguistic terms in order to improve the performance. For such, the classifier uses the bags word model as the baseline.

Keywords: Sentiment Analysis, Opinion Mining, Polarity Classification, Supervised Learning, Linguistic Features, Semantic Orientation, Machine Learning.

Agradecimentos

A deus em primeiro lugar por me iluminar e dar forças para atingir este patamar;

A minha família por acreditar e aceitar o ser ausente e compreender tempo e hora as razões os benefícios que este novo degrau subido traz.

A minha orientadora Professora Drª Irene Rodrigues Pimenta, que de forma profissional e desinteressada abraçou a missão e acreditou desbravar as terras longínquos para tornar este sonho possível, num momento que o nosso país não possui instituições para tal feito, citar a sua disposição em ajudar mesmo em seu horário nobre.

A todos os Professores da Universidade de Évora que participaram directamente na nossa formação, ao acreditarem e deslocarem-se, para cumprir uma missão desinteressada, fica o meu reconhecimento, nas suas capacidades, destrezas na passagem de testemunho e conhecimentos.

A todos os meus colegas que ajudaram directamente ou indirectamente a chegar até aqui o meu reconhecimento.

IV

Lista de Acrónimos

APIs Interface de programas de aplicações

AM Aprendizado em Máquinas

AS Análise de sentimentos

BD Base de Dados

CRF Frequência condicional Randómico

e.g. Em geral

EC Extracção de Características

HTML Linguagem de Marcação de Hipertextos

INSS Instituto Nacional de Previdência Social

i.e Isto é

IA Inteligência Artificial

NLTK Quite de ferramentas de Linguagem Natural

MD Mineração de Dados

MT Mineração de Texto

NP Sigmas nominais

NGD Distância de Normalização de Google

PLN Processamento em Linguagens Naturais

PMI Informação pontual mútua

KDD Descoberta de conhecimento em Base de Dados

KNN Vizinhos mais próximos

RI Recuperação de informação

SVM Vectores de Suporte de Máquina

SO Orientação semântica

SC Selecção de características

SE Sistemas Especialistas

SNS Sítios de redes sociais

SAS Sistema de Análise Estatística

SPSS Statistical Package for the Social Sciences

TIs Tecnologia de Informação

t.c Tal como

LC Linguagens convencionais

TMT Técnicas de Mineração de Textos

Lista de Figuras

Lista de Tabelas

Conteúdo

Capítulo 1

Introdução

Neste Capítulo são apresentados os aspectos introdutórios e elementares do projecto e realiza-se o enquadramento da tese em referência ao problema tratado ou a solucionar, a sua delimitação, os objectivos as metodologias para a sua realização, assim como o horizonte organizacional ou estruturante da tese.

1.1 Considerações Iniciais

Existe uma dinâmica acelerada na produção e armazenamento de dados estruturados, semi-estruturados e não estruturados disponíveis sobretudo na forma digital e em diversas fontes e localizações remotas, algumas organizações possuem Trilhões de registos e diariamente são acedidos para pesquisas ou transacções electrónicas como é o caso do Google, Yahoo, YouTube, Amazon etc.o que contribui com aumento significativo do conteúdo, as informações que armazenam estão na ordem de milhões de Terabytes. Para termos ideia este volume cresce ou aumenta de forma exponencial a medida que o tempo evolui, resultados atribuídos a evolução acelerado e desenfreado

das Tecnologias de Informação e Comunicação (TICs). Afortunadamente a produção e armazenamento não representa nenhum problema no contexto actual, porque os meios de suporte existentes possuem capacidades para tal. Porém a capacidade de produção e armazenamento de conteúdo é extremamente superior a capacidade de interpretação e extracção de conhecimentos nos dados. Com vistas a dimuir o fosso na interpretação e extracção de conhecimentos, as pesquisas e experiências recentes concentram-se em traçar linhas mestres e estratégias viáveis para a realização desta tarefa, o que possibilitou a concepção e desenvolvimento de técnicas, métodos e ferramentas especializadas, cujos alicerces principais assentam na conjunção multidisciplinar de áreas tais como: Informática, Estatística, Linguística e Ciência Cognitivas etc. A principal actividade desde processo é a mineração e para a sua realização requer a combinação de várias técnicas e métodos Computacionais e Estatísticos e.g. Técnicas de Base de Dados; Inteligência Artificial (e.g. Aprendizagem automática, Abordagens Evolutivas, Simbólica e Conexionista), Métodos Estatística, Reconhecimento de Padrões, Recuperação de informação etc.

A busca de conhecimento em textos consiste em extrair informações relevantes, padrões e anomalias ou tendências em grandes volumes de textos simi-estruturados ou não estruturados disponíveis em Linguagem Natural sobretudo através de um processo automático, suas bases advém da mineração de dados, com uma diferença de que este último procura descobrir padrões úteis em fontes de dados estruturados. A realização da Mineração de Texto (MT) somente é possível com auxílio ou apoio de técnicas de Aprendizagem automática (AM) e Processamento em linguagem Natural (PLN), áreas de especialização de Inteligência Artificial (IA). Esta tarefa permite sobretudo recuperar informações, extrair dados, sumarizar documentos, descobrir padrões ocultos, associações e regras e fundamentalmente efectuar análise quantitativa e qualitativa que permite a tomada de decisões com base na interpretação dos resultados obtidos.

Os dados não estruturados sempre foram em termos de volumes superiores aos estruturados actualmente dado a sua facilidade de produção e armazenamento e a imperiosa necessidade do seu tratamento, várias técnicas e aplicações foram desenvolvidas e como entradas fontes incluem: *emails*, textos livres obtidos como

resultados de pesquisas, arquivos eletrônicas gerados por editores de textos, páginas da Internet onde existem diversidade de publicações, campos textuais em bases de dados, documentos eletrônicos digitalizados a partir de papéis, fóruns de discussões, médias sociais e etc. Pesquisas recentes apontam que 80 por cento do conteúdo *online* assim como também o equivalente em volume de dados guardado pelas corporações não é estruturado [79].

O foco do projecto é realizar um estudo das abordagens referentes a sistemas e técnicas actuais utilizadas na descoberta de conhecimentos e Análise de Sentimentos com intuito de implementar um protótipo de Avaliação de Opiniões em textos. Para a realização de experiências e obtenção dos resultados dois corpura foram utilizados nomeadamente o SentiCorpus-pt e Review Movie como entradas para a realização da tarefa com auxílio de algoritmos supervisonados. A nossa abordagem é avaliar os sentimentos contidos nos dois corpura com aplicação dos algoritmos Naive Bayes e Máxima Entropia, de formas a comparar a eficiência da avaliação. Outrossim, também é feita o casamento de termos linguísticos com vista a melhorar a performance do classificador em função da utilização de filtros como *Stopwords* e o método ou modelo Balsa de Palavras como linha de base e técnicas de Bigrama e utilização de melhores características para aumentar o desempenho e a precisão do modelo.

1.2 Definição do Problema

Qualquer organização actualmente lida com um grande volume de informações em diferentes formatos respeitante aos seus negócios ou opiniões de fontes externas que podem ser de parceiros ou concorrentes, estas informações podem ser constituídos por expressões que vão desde pequenos comentários, opiniões direitas e indirectas, aquelas expressas pelos actores formalizadores de opiniões, ou outros expostos em livros, textos, jornais ou outros meios supracitados, mantidos localmente ou remotamente. É importante frisar que existem várias indagações e incógnitas cujas respostas são dificeis de serem fornecidas. Vejamos a seguir alguns exemplos:

1. Que necessidades reais da organização as informações (i.e. de concorrentes, parceiros, ou da própria organização) guardadas ou geradas atedem?

2. Qual tem sido o proveito ou o benefício que essas informações têm proporcionado na tomada de decisões instituicionais?
3. As informações não estruturadas são convertidas num formato propício constituindo padrão, formas a facilitar a análise e descoberta de conhecimento?
4. É possível termos uma visão clara do que está a pensar o cliente e a concorrência em relação aos nossos negócios?
5. É possível saber se um determinado trabalho é integralmente do autor que diz a referência?

Estas perguntas e outras mais podem ser respondidas com a utilização da ferramenta de análise de sentimentos, que é um tema de realce nas pesquisas mais recentes e representa uma preocupação premente no contexto geral e nas novas formas de governação da Internet. A sua aplicação não é uma tarefa tão trivial, porque lida-se com emoções que são parte integral de como os humanos reagem a estímulos externos, certos acontecimentos ou se comunicam com o mundo externo, o que acontece através da fala, escrita, expressões faciais, gestos etc.

Uma frase tem uma força extrema em modificar um cenário, que pode passar de favorável para desfavorável se difundida pelas médias actuais e também pode ser pertinente a diversas entidades e determinar a emoção invocada por uma entidade. Outrossim a difusão de informações nos moldes actuais pode ser benéfica para uns e prejudicial para outros e.g. feita através de emails, mensagens por telemóveis, comentários na Web ou através de médias sociais (microblogs, twitter, facebook etc.), despertando interesse de vários participantes em frqcções de segundos e gerando diversos sentimentos e opiniões.

Os sistemas actuais de análise e técnicas existentes focam sobretudo nos sentimentos disponíveis em médias sociais e utilizam sob grande medida técnica de propósitos gerais de mineração de dados, outrossim, a maioria dos corpus disponíveis para este efeito estão em língua Inglesa, logo requer a necessidade de implementações em outras línguas dada a importância de soluções com esta natureza e a universalidade da sua aplicabilidade.

1.3 Motivação

Desde tempos remotos a informação foi e continua a ser um activo importantíssimo em quailquer organizações ou indivíduo. Os seus detentores sempre apresentam-se na vanguarda em coparação a aqueles que não a possuem, ou em contraste as que a detém mais que não a exploram o suficiente ou eficientemente para auxilia-los a tomar decisões que convergem a direccionar os negócios, anseios ou na realização efectiva do planeamento, para o efeito a análise para a descoberta de conhecimento possibilita a previsibilidade, além de fornecer uma grande probabilidade extrema de acertos nas decisões a tomar, interesse o conhencimento oculto sobre os seus negócios, objectos, personalidade etc.

As pesquisas actuais apontam uma corrida desenfreada na tentativa de encontrar soluções como saída para a extracção ou coleccionar informações existentes em diversas fontes (p.e. em sítios, fóruns, comentários em jornais online, microblogs, e redes sociais etc.) de maneira automática e outras ferramentas de análise com emprego de diversas técnicas, mecanismos e métodos de classificação ou agrupamento aplicados a múltiplas línguas com vistas à obtenção do conhecimento sobre certos assuntos de interesse p.e. Eleições, produtos, objectos, personalidades políticas etc. Porém o desafio é amplo, logo os estudos indicam que não existe, todavia uma linguagem padrão ou um dicionário para a classificação e análise de documentos semi-estruturados ou não estruturados numa perspectiva ampla em termos quantitativos, tangente a multilinguagens, devido a factores t.c. heterogeneidade, peculiaridades, aspectos culturais, estruturais e à semântica dos componentes intriseca a cada linguagem, assim como a ampla diversidade de sinónimos, lexemas associados a determinadas frases lógicas, a dinâmica e a forma da sua produção envolvendo gírias e erros associados, em função dos níveis de especialização e a diversidade dos seus produtores e a dinâmica da produção, além da ambiguidade nativa, que caracteriza as linguagens naturais, o que sem sombras a dúvidas torna a tarefa de descoberta conhecimento e análise mais complexa.

Em resumo as abordagens estudadas não descrevem ou fornecem uma linguagem padrão, método ou sistema para análise de documentos de texto e, por conseguinte às seguintes perguntas não são tão fáceis de serem respondidas:

1) É possível desmistificar a dependência sintáctica e semântica de cada linguagem de forma a criar um mecanismo de análise de texto não estruturado ou semi-estruturado padrão e multilíngue?
2) As técnicas e métodos actuais de classificação de opiniões basedos tanto nas características sintácticas como semânticas comportam-se de forma semelhante na análise de qualquer documento de texto e os resultados obtidos seriam semelhantes?
3) A precisão dos métodos actuais de análise de texto pode ser melhorada para proporcionar resultados com uma margem de erro ínfima?
4) Os actuais sistemas têm a capacidade de extrair documentos de forma automática e distinguir características relevantes e não relevantes em quaisquer corpora e fazer à análise de forma eficiente e eficaz em qualquer língua ou idioma?
5) Os resultados obtidos em qualquer língua se comparados com os obtidos noutras línguas e com a utilização dos mesmos algoritmos são equivalentes?

As respostas a estas indagações serão respondidas ao longo do desenvolvimento deste projecto. O campo análise de sentimentos está na infância e precisa ser explorada com a criação de técnicas e métodos que proporcionam vantagens técnicas e facilidades na elaboração e realização das tarefas.

1.4 Objectivos Gerais

A presente tese tem como objectivos centrais fazer uma abordagem sucinta as técnicas e sistemas actuais de análise de sentimentos e descoberta de conhecimento com vistas a construir um protótipo de análise de opiniões em textos, apoiado pelas técnicas de inteligência artificial, mineração de dados e textos, com vistas a marcar e classificar a subjectividade e a polaridade (positiva, negativa ou neutra) nos texto, de formas a obter respostas se expressam sentimentos negativos ou positivos, em relação ao um determinado assunto ou universo de discurso p.e. das notícias que falam do desporto

podemos retirar um subconjunto das frases que de facto falam destas como pilares para a classificação.

1.4.1 Objectivos Específicos

Para a realização deste projecto faz-se a abordagem e estudos concentrando-se em quatro pontos fundamentais:

1- Estudar e abordar as técnicas e métodos de mineração de dados e textos utilizadas para a análise de dados e opiniões ou sentimentos no contexto actual;
2- Estudar e fazer abordagem dos sistemas actuais de análise de sentimentos;
3- Implementar um protótipo para a análise de opiniões ou sentimentos disponíveis em médias electrónicas;
4- Produzir um material guia de referência para as futuras implementações e consultas a comunidade académica.

1.5 Metodologias de Investigação

A identificação e análise de sentimentos é uma tarefa árdua atendendo à natureza dos dados e justificada pela dinâmica da sua produção e volume caracteristico. Imaginemos o seguinte, actualmente grande parte dos serviços de governos e corporações que se presam são fornecidos de forma *online.* Alguns governos possuem bibliotecas digitais para guardar as informações veja a dimensão das informações que contém é difícil de se imaginar. Assim os métodos convencionais dificilmente dariam suporte eficiente a análise e descoberta de conhecimento, veja que a identificação de sentimentos não se resume apenas a análise das classes, frases, ou composição gramatical, mais também o contexto e o comportamento das entidades em função de aspectos linguísticos. Não é supressa que o método correcto para o efeito reveste-se no foco principal que será a exploração de suas propriedades fundamentais, através de Detecção e Análise de Subjectividade em virtude de que o estilo das frases pode ser objectivo ou subjectivo e a Análise de Polaridade pode ser positiva, negativa ou neutra.

O método utilizado para a realização da Análise de Sentimentos neste projecto é baseada nas técnicas Aprendizagem automática, e Processamento em Linguagem Natural e utilizou-se especificamente os algoritmos supervisionados Naive Bayes e Máxima Entropia. Outra técnica de realce utilizada para a tarefa de processamento e análise dos Corpus é o Bolsa de Palavras, que consiste em construir um vector de termos e associa-los a frequência de sua ocorrência no texto, para optimizar os resultados fez-se aplicação das técnicas filtragem e Bigrama e subsequentemente, a construção de modelos de treinamente e consequentemente a realização de teste com as melhores características para obtenção de melhores ganhos, cujas principais fases envolvidas na realização da tarefa são: Extracão de Inputs; Pré-Processamento; Extracção de Subjectividade; Extracção de Características; Classificação e Visualização.

1.6 Organização da Tese

A tese está estruturada ou organizada fundamentalmente em cinco capítulos que exploram de forma evolutivas e apresentam os aspectos conceptuais e práticos, descritos a seguir:

1. Capítulo 1 apresenta os aspectos introdutórios elementares, apresentando o enquadramento da tese em referência ao problema a ser resolvido a sua delimitação, os objectivos, as metodologias para a sua realização e a linha organizativa;
2. Capítulo 2 tem o foco principal voltado na apresentação dos estudos e soluções recentes a nível de aplicações e sistemas que fazem abordagem a Descoberta de Conhecimento e Análise de Sentimentos, elucidando fundamentalmente as técnicas, métodos, ferramentas e etapas seguidos na realização dos projectos;
3. Capítulo 3 apresenta os aspectos teóricos elementares para o processo de aquisição ou descoberta de conhecimento, mineração de dados e textos em geral, as fases para a realização das tarefas das distintas fases, assim como os mecanismos e métodos de extracção e pré-processamento de dados, mecanismo de indexação dos dados, recuperação de dados, processo de mineração e a

descrição dos algoritmos fundamentalmente utilizados para a realização da tarefa de mineração;

4. Capítulo 4 faz abordagem geral do processo de análise de sentimentos, descrevendo a sua importância e o seu contributo, outrossim, apresenta as fases, e a arquitectura geral de análise de sentimento e descreve taxativamente os níveis de incidência sobre as quais a análise pode se centralizar, assim como as áreas de aplicação e os principais desafios que a área apresenta;
5. Capítulo 5 apresenta o protótipo elaborado, objecto do presente projecto, os testes e os resultados alcançados com a utilização dos algoritmos supervisionados. São também descritos neste capítulo todos os recursos e instrumentos utilizados para a realização do trabalho, descrevendo, além das ferramentas e linguagens utilizados no contexto prático;
6. Capítulo 6 apresenta as considerações finais, os principais contributos, lições aprendidas, recomendações e trabalhos futuros.

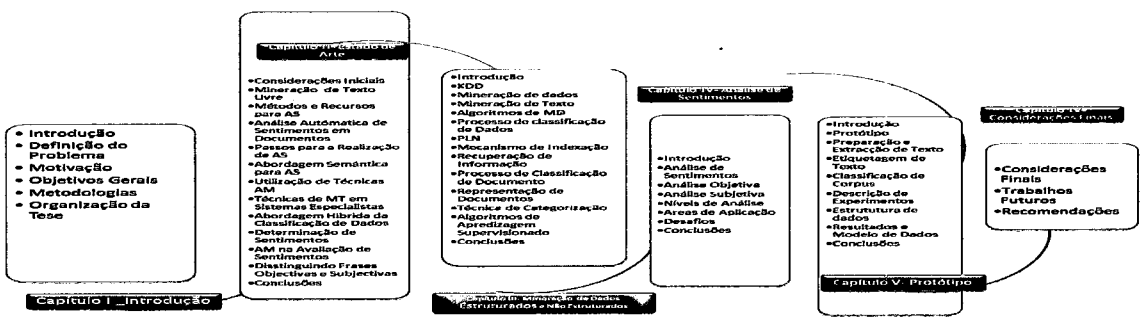

Figura 1.1: Estrutura da Tese

A figura 1.1 apresenta a estrutura detalha da tese, na qual estão espelhados todos os capítulos que a compõem, assim como as principais sessões de cada capítulo. O capítulo introdutório por exemplo apresenta as seguintes sessões: introdução que aborda os aspectos introdutórios; A definição de problema, onde estão catalogados os aspectos ao solucionar ao longo do desenvolvimento da tese; A motivação, na qual são ilucidados os aspectos que levaram a idealizar uma solução para o contexto, objecto do presente trabalho; os objectivos que indicam o norte a ser atingindo; As metodologias aplicadas para a consecução dos objectivos; o estado da arte que espelha os trabalhos desenvolvidos nesta área e a sua relação com o projecto desenvolvido e finalmente a última secção apresenta a estrutura organizacional da tese.

Capítulo 2

Estado da Arte

2.1 Considerações Iniciais

Este capítulo faz uma abordagem sucinta a soluções e estudos mais recentes elaboradas na área objecto de estudo e as áreas co-relecionadas, elucidando os principais mecanismo, técnicas, métodos e a visão de cada projecto, assim como o processo e etapas envolvidos na sua realização, a nível de aplicações, sistemas e tecnologias de base utilizadas. O foco principal considera sobretudo os estudos que abordam a problemática de descoberta de conhecimento e análise de sentimentos em textos.

2.2 Mineração de Textos Livres

Em [3] propõe-se uma estratégia consubstanciada na mineração de textos livres escritos em Português independentes de domínio. A análise baseia-se no estudo das propriedades estruturais linguísticos (i.e. morfologia e sintaxe), com vista a extrair informação útil em documentos textuais não estruturados que tem como principal característica a escrita informal, com possível presença de erros ortográficos, abreviações, gírias, símbolos, erros gramaticais e pontuação incorrecta ou a falta desta,

normalmente escritos por utilizadores com níveis de conhecimentos distintos [3]. Para o efeito faz-se a análise gramatical para a identificação das classes associadas às palavras (i.e. Substantivos, Verbos, Adjectivos, Artigos, Pronomes, Numeral, Advérbios, Preposições, Conjunção e Interjeição) contidos no texto e análise sintáctica que define a função que as palavras desempenham dentro da oração p.e. Sujeito, Adjunto Adverbial, Objecto Directo e Indirecto, Complemento Nominal, Aposto, Vocativo, Predicado etc. Para a realização do processo da análise utiliza uma metodologia composta por quatro etapas principais: 1) Configuração; 2) Pré-Processamento; 3) Processamento e 4) Pós-Processamento.

- Na fase de Configuração: diz respeito à formação inicial da base de dados (Léxico), que contém entre outras informações, termos e entidades nomeadas (i.e. palavras da classe de substantivos próprios e definem nomes de lugares, pessoas, organizações, acontecimentos, objectos, obras etc.).
- Na fase de Pré-Processamento: faz-se a verificação preliminar do conteúdo do documento que inicialmente é convertido em *tokens*. Um *token* consiste de um par ordenado (valor, classe). A classe indica a natureza da informação contida num valor.
- Na fase de processamento: faz-se a Mineração de Texto, na qual o documento é sumarizado e, com auxílio do Léxico, é realizado o processo de Análise Morfológica dos termos, complementado com a Análise Sintáctica dos termos do documento de formas a identificar as palavras e frases mais importantes de um documento ou conjunto de documentos e gera também um resumo ou sumário. Esse sumário pode dar uma visão geral do conjunto de documentos ou pode ainda salientar as partes mais importantes ou interessantes, em seguida faz-se a identificação dos períodos que dividem o documento sumarizado em orações. Cada período identificado é submetido aos processos de remoção de *Stopwords e Stemming*, na sequência é feita a classificação dos termos com a ajuda de heurísticas de categorização dos termos. Por fim, na fase de Pós-Processamento os dados adquiridos nas etapas anteriores são transformados, carregados e armazenados numa *Data Mart* que faz a gestão da informação textual extraída de vários portais de consumidores de acordo com as necessidades dos utilizadores e dá suporte à descoberta de novas informações.

No final deste trabalho, descreve-se a forma como a proposta pode ser estendida para a elaboração de uma estratégia que permite descobrir relações entre os termos dos textos submetidos à análise. Também se vislumbra os ganhos previstos com a construção

doravante de um Módulo de Polarização para enriquecer os resultados a serem obtidos com base numa nova perspectiva de análise e avaliação dos dados.

2.3 Métodos e Recursos para Análise de Sentimentos

Em [4] são apresentados alguns métodos e recursos para análise de sentimentos em vários Documentos escritos em diferentes línguas. A abordagem baseia-se amplamente na mineração de características e no paradigma de sumarização cuja teoria foi inicialmente concebida e descrita em [19] e concluída em [22], além de fazer a extensão das técnicas e métodos para a detecção e classificação automática de sentimentos expressos de forma directa, indirecta ou explícita em diferentes tipos de textos e idiomas, aplica o método da análise de sentimentos que se propõe no contexto de outras tarefas de PLN (processamento em Linguagens Naturais) e.g. busca de respostas e resumos automáticas. Utiliza a aprendizagem automática (AM) de forma automática como padrão para as pesquisas, sobretudo nos conjuntos de dados que têm centenas ou milhares de características, por exemplo, o processamento e categorização de textos extraídos na Internet. A selecção de atributos, variáveis ou características é uma técnica em que são escolhidas as amostras das características mais relevantes e representativas para um dado problema num universo de características de um conjunto de dados, além de propor um novo esquema de anotação e novos corpora, que visam padronizar a rotulagem de opiniões, de modo que diferentes tipos de comentários directos, indirectos, implícitos e as diferentes maneiras de expressar opiniões através de declarações subjectivas ou objectivas possam ser rotulados ou etiquetados correctamente, para outras línguas. A análise de sentimento é avaliada com base no contexto de um determinado produto. Utiliza a abordagem baseada na extracção de características e paradigma de sumarização de textos e documentos cujas bases teóricas foram descritos por [19,20,21] respectivamente, cria um método mineração de textos baseados nas características explícitas para recomendar produtos com base nas pontuações obtidas na quantificação de diferentes características na etapa de MT (mineração de texto), através da extensão do método apresentado em [22], cujo princípio de funcionamento é a comparação dos detalhes técnicos do produto em função de medição de termos relacionados (i.e. aplicação de distância de normalização) e a análise semântica latente.

O método de mineração e análise de sentimentos em produtos consiste em duas etapas distintas: pré-processamento e processamento principal. Cada etapa contém uma série de submodulos e utiliza ferramentas, linguagens e diferentes recursos.

- Na primeira etapa parte-se do princípio de que se o utilizador faz uma consulta do produto que deseja comprar, o motor de busca recupera uma série de documentos que contêm as referências do produto em questão em diferentes idiomas, em seguida realiza duas operações paralelas: a primeira faz a filtragem dos comentários utilizando o software identificador de idioma *Lextek* e obtém resultados em Inglês e Espanhol. A segunda operação determina a categoria do produto (p. e. câmara digital, laptops, impressoras, livros etc.) e determinada a categoria, procede-se para a extracção das características específicas do produto e atributos do produto através da utilização de WordNet e o *ConceptNet* e mapeamento correspondente em Espanhol utilizando o *EuroWordNet* (i.e. base de dados com wordnets para várias línguas Europeias), com as classes, características e atributos tem-se o que é considerado como núcleo de características e atributos independentes do produto cuja importância determina a sua frequência de ocorrência no universo de pesquisa;

- A segunda fase consiste no processamento principal, nesta fase duas entradas de clientes uma em Espanhol e outra em Inglês são executados em paralelo pelo sistema e como saída é gerada a sumarização. A análise começa com filtro de acordo com a linguagem e para cada linguagem utiliza uma ferramenta especializada para a resolução anafórica (t.c. *JavaRAP* para Inglês e Supar para Espanhol), também faz-se a separação das frases do texto de acordo com as características de interesse, em seguida, utiliza-se um detector ou identificador de entidade nomeadas para distinguir nomes de produtos, marcas ou lojas, para esta última tarefa utiliza-se como ferramenta o *LingPipe* que é um conjunto de bibliotecas Java para a realização de análise em Língua Natural. Completada a tarefa de identificação de frases e atributos, procede-se a seguir com a extracção com o propósito de processar apenas características referenciados pelo produto, e também faz-se o *parser* para obter a estrutura das frases e as dependências entre os componentes. Para esta finalidade são utilizados a ferramentas *Minipar* [19] para Inglês e *FreeLing* para Espanhol. E por fim de modo a atribuir a polaridade para cada um dos atributos empregam-se os métodos ou algoritmos de classificação *Support Vector Machines* - Optimização mínima sequencial (SVM) e aprendizagem automática [25], assim como a distância de Normalização de Google (NGD). SVM é um algoritmo de classificação originalmente desenvolvido por Vladimir Vapnik, cuja versão de

referência foi proposta por [24]. O algoritmo tenta estimar uma função $f: R^n \rightarrow \{\pm 1\}$ utilizando um conjunto de treinamento, onde cada elemento deste conjunto é um vector N-Dimensional $(xi, yi) \in R^n \times \{\pm 1\}$ de formas que esta função será capaz de classificar uma nova instância (x,y) ou seja f(x)=y. Já a abordagem baseada em aprendizagem automática utiliza classificadores para separar ou identificar frases de interesse. Técnicas de aprendizagem automática são utilizadas em reconhecimento automático de termos, que são projectadas para atender uma classe específica de entidades e utilizam dados de treinamento para aprender em função das características que são úteis e relevantes no reconhecimento e a classificação de termos [26]. NGD tem por finalidade medir e proporcionar como resultado em função de determinar quão próximo estarão os termos, no espaço dos documentos indexados pela *Google*. É uma amostra robusta que pode ser tomada como corpus[1] representativo da Língua actual. O algoritmo de cálculo indica que se deseja medir o NGD dos termos, p.e. se quermos consultar os termos Professor e Aluno devemos pesquisar na Google sobre a quantidade de documentos que contém o primeiro termo, e depois o segundo e finalmente ambos os termos, para tal é utilizado uma fórmula para o cálculo da distância, computa-se a distância entre os términos a avaliar, para a obtenção dos resultados.

No final do seu trabalho, vislumbra a possibilidade de abordar um módulo para extracção automática de taxonomia e extensão do sistema de anotações de revisões e resumo noutros idiomas para além de Inglês e Espanhol.

2.4 Análise Automática de sentimentos em Documentos

Em [27] é apresentada uma abordagem para análise automática de sentimentos em documentos, que cobre várias tarefas de análise de sentimentos para ilustrar novas mudanças e oportunidades, descreve como modelar diferentes tipos de relacionamentos em várias abordagens para problemas de análise de sentimentos ou modelos que podem ter implicações fora desta área. Faz a classificação de polaridade através da análise e classificação de informações textuais de filmes. Considera várias abordagens incluindo algumas que aplicam as técnicas de categorização subjectiva do texto, descreve também uma abordagem baseada em técnicas eficientes para encontrar o custo mínimo nos gráficos que incorporam relações em nível de frases. Também considera meta

[1] Corpus: conjunto de documentos, dados e informações sobre um determinado assunto de interesse.

algoritmos para explorar relacionamentos explícitos entre classes ou problemas de classificação multi-classe. Faz a classificação da polaridade em documentos aplicando a técnica ou método de aprendizagem automática, formalmente para a classificação utiliza o algoritmo *Naive Bayes e Máxima Entropia e Support Vector Machine.* Para a avaliação ou análise de polaridades utiliza uma base de dados de documentos na qual os dados são subdivididos em três áreas de igual tamanho e para a marcação de texto utiliza o formato *HTML (Hiper Text Mark-up Language)*[2] de formas a separar os itens léxicos e aplicação do método *N-Grama* para a caracterização e desambiguação das palavras utiliza o programa *Oliver Mason's Qtag program.*

Por fim como extensões ao trabalho propõe o desenvolvimento doravante de um conjunto de dados que cobrem o alcance, de modo a representar *consultas* com intuito de retornarem páginas Webs completas pesquisadas no motor de busca qualquer.

2.5 Passos para Realização de Análise de Sentimento Multiligue

Em [28] é apresentada uma abordagem de quatro passos para a realização da Análise de Sentimentos Multilíngue em médias sociais, cujas principais fase são: 1) Identificação da Língua; 2) *Part-Of-Speech (POS);* 3) Detecção de subjectividade e; 4) Detecção de Polaridade. Cada passo aborda um sub-problema e descrever o problema separadamente e apresenta uma solução para cada um destes. O processo é feito da seguinte maneira, a entrada é constituída por pequenos textos não estruturados, as quais não passam conhecimento prévio, nem mesmo o meio social do qual se originam. A cada passo mais informações são adicionadas e como saídas obtém-se a polaridade, o que implica que apenas textos subjectivos sejam processados pelo passo de detecção da polaridade, porque é sabido de que os textos objectivos por defeito não possuem qualquer polaridade. A lógica que está por trás da utilização da abordagem de quatro passos é que podemos mais especificamente construir modelos numa etapa posterior, quando o conhecimento é apresentado nos passos anteriores. A separação da subjectividade e detecção de polaridade é inspirada em [29].

---

[2] HTML: é uma linguagem que informa ao navegador que elementos estão na página, p.e. arquivos (imagens, sons) que eles contém e onde eles estão p.e., um certo trecho é identificado como o título principal do documento e outro trecho como um link.

Para a identificação da linguagem propõe o algoritmo *LIGA* que captura a gramática da linguagem na adição da ocorrência de características, através da utilização de formalismo gráfico, incorpora a gramática no modelo, entretanto a técnica de Processamento de Língua Natural nomeadamente *Part-Of-Speech (POS)* pode ser utilizada para auxiliar as expressões regulares na identificação dos termos relevantes contidos numa frase. O etiquetador POS consiste em rotular as palavras segundo a sua classe gramatical normalmente em: substantivos, adjectivos, advérbios, verbos e preposição, que são alguns exemplos de classes gramaticais [30]. Regras simples sem POS é um tipo de regra convencional em que o desenvolvimento do padrão é dependente do domínio. As regras apenas com POS utilizam padrões POS mais específicos visando extrair termos com baixa ocorrência de falsos positivos. As regras com POS e termos representativos utilizam-se de termos representativos (e.g. verbos ou palavras) para identificar se uma frase contém ou não um termo [30]. Para a tarefa de POS-TAG*ging* utiliza-se a solução *TreeTagger* desenvolvida pela Universidade de Stuttgart, Para a detecção de subjectividade aplica-se o *AdaBoosts* que utiliza o modelo de AM e para a detecção de polaridade propõe-se o algoritmo *RBEM* que utiliza regras heurísticas para criar um modelo emissivo nos padrões.

Cada estrutura pode ser estendida, modificada ou substituído para adequa-la a diferentes propósitos. É um pipiline na qula a saída de um passo constitui a entrada do próximo passo i.e. não existe uma dependência operacional entre os algoritmos, mais sim entre os passos, isto significa que alguns passos podem ser substituídos de alguma forma ao longo de entrada e permanecerem nalguma saída, por exemplo, o passo de detecção de subjectividade pode ser substituído pelo algoritmo *Naive Bayes* para realizar em duas tarefas a classificação de termos em subjectivo ou objectivo.

2.6 Abordagem Semântica para Análise de Sentimentos Automaticamente

Em [31] é apresentada uma abordagem semântica para análise de sentimentos em textos de forma automática, com base na classificação da linguagem tomado como referência para a avaliação efectiva das frases, utiliza-se como referência o trabalho apresentado em [33], que consiste em subdividir as tarefas em três subactividades interrelacionadas: 1) determinar se certa unidade da linguagem é subjectiva; 2) determinar a orientação da

polaridade em função da subjectividade de unidades da língua; e 3) determinar a força da orientação semântica. Desta maneira a análise é feita em função da unidade em que a pesquisa é focada p.e. Palavra, frase ou textos completos, o sistema de análise focada nesta abordagem determina a polaridade do texto na sua plenitude, cuja pesquisa inicial para a classificação da polaridade foi dominante em duas abordagens nomeadamente: o modelo AM popularizado descrito por [19] e a outra que foca em palavras e frases como portadores de orientação semântica ou SO (*semantic orientation*) [34], que não é mais do que a média da soma dos SOs das partes de um documento. Implementa ou rescreve a calculadora SO e adiciona a esta novas características com um dicionário e regras para detecção de palavras negativas, intensificação, modularidade, repetição e outros fenómenos que afectam a orientação semântica i.e. Algoritmo para a contagem de palavras, em função do seu peso no texto, e negação. Constrói um dicionário em Espanhol análogo ao do Inglês, que inclui adjectivos, substantivos, verbos, advérbios, e intensificadores, criando uma lista de 157 palavras e expressões baseadas na lista em Inglês. Com este dicionário três métodos diferentes foram testados para comparar a performance de corpus em Espanhol, no primeiro teste fez-se a tradução automática de Inglês para Espanhol preservando o valor da orientação semântica da palavra, para a tradução automática empregam-se dois métodos. O primeiro consiste num dicionário bilingue online onde são extraídos as primeiras definições da categoria sintáctica ignorando alguns casos de expressões com múltiplas palavras em ambas as línguas. O segundo método consiste na tradução automática e envolve simplesmente *pluggins* ou um dicionário Inglês a partir do tradutor *google* e o resultado do analisador sintáctico, novamente excluindo expressões com múltiplas palavras. Para o segundo método criou-se o dicionário com uma lista *spanishdict.com* e fixou-se manualmente as entradas erradas como é óbvio, ou seja, envolve a remoção de *stopwords.* Finalmente o terceiro método consiste em criar todos os dicionários desde o início.

2.7 Utilização de Técnicas de Aprendizagem Automática

Em [35] é feita uma abordagem que utiliza técnicas de AM no reconhecimento de entidades nomeadas em Português através da comparação dos resultados obtidos com a aplicação das técnicas de AM tanto supervisionados como não supervisionados nomeadamente: e.g *Naive Bayes, SVM* e a árvore de decisão (*Decision Table*) no

processo de classificação das entidades nomeadas nos documentos escritos na Língua Portuguesa. Utiliza também corpus disponíveis em [36], com anotações de POS, classes gramaticais de cada palavra e entidades nomeadas associadas à palavra. *Naive Bayes*: é por sinal o classificador mais utilizado considera que as entradas são independentes entre si, o que não ocorre na maioria dos problemas ou casos práticos, i.e. baseia-se numa premissa de ingenuidade, porém embora esse princípio seja mantido por defeito ao trabalhar com este algoritmo, ainda assim não compromete a qualidade dos resultados reportados pelo classificador. O SVM: é uma técnica utilizada no processo de reconhecimento de padrões e regressão linear, baseada na utilização de *kernels* (núcleo) e de regras não-lineares. A ideia central é o núcleo é tido como o produto interno entre o chamado vector de suporte e um vector retirado do espaço de entrada. Os vectores de suporte são um subconjunto dos dados de treinamento [36]. *Decision Table*: é uma das técnicas mais simples de aprendizado supervisionado é uma das mais simples de se entender atendendo o seu princípio de funcionamento. Algumas tabelas utilizam valores verdadeiros ou falso (*true ou false*) na representação de alternativas e condições (p.e. if-then-else). Outras utilizam alternativas numeradas (p.e. switch-case) e, ainda existem aquelas, que utilizam a lógica fuzzy (i.e. lógica difusa que apresenta vários níveis entre verdadeiro e falso) ou representações probabilísticas para as alternativas condicionais [37].

2.8 Técnicas de Mineração de Texto e Sistemas Especialistas

Em [38] é Abordada a aplicação de Técnicas de Mineração de Textos (TMT) e sistemas Especialistas (SE) na liquidação de processos trabalhistas. Utilizam-se as TMT para a tarefa de classificação, e definir em Linguagem Convencional (LC) e SE para definir os pedidos fundamentados em freses judiciais; Identificar o resultado de cada um destes pedidos (i.e. deferido ou indeferido); Extrair cada uma das incidências (reflexos) geradas pelos pedidos e Capturar eventuais parâmetros para o cálculo e valor de algum tipo de pedido. O SE é utilizado para para calcular o valor associado à acção julgado como compensação ao cliente. As principais fases consistem em: 1) Aplicar as técnicas relacionadas às tarefas de classificação ou categorização para identificar quais pedidos (p.e. hora extra, adicional de perigosidade, equiparação salarial, *ticket* de transporte, etc.) estão definidos na fundamentação do Exmo. Juiz, cada fundamentação pode ser

decomposta em uma "bolsa de palavras" extraida da frase, Com base nos métodos de MT, em que um grupo de "bolsa de palavras" passa por um algoritmo de aprendizado, utilizando técnicas de classificação, como SVM , Naive Bayes [39,40], novas "bolsas de palavras" poderão ser classificadas. Os pedidos seriam as "classes" e estariam relacionados às "bolsas de palavras"; 2) utilizar Linguagens Convencionais para identificar se cada pedido foi deferido ou indeferido. Se o pedido foi deferido, utilizar também LC ou MT para capturar os reflexos e outra vez LC para capturar outros parâmetros necessários ao cálculo; 3) Passar as informações para um Sistema Especialista, via uma interface, baseada nas regras obtidas anteriormente com os especialistas, e assim calculará o valor exacto da reclamação, também o que deverá ser pago à Previdência Social (INSS) e à Receita Federal (IRRF). O vínculo decorre da necessidade de se fazer cálculos de forma repetitiva, objectivando rapidez, evitando erros e a necessidade de utilizar uma grande quantidade de regras, isto remete para os Sistemas Especialistas (SE).

Propõe como extensões ao trabalho a criação e implementação do módulo central do sistema entre MT e SE, que contempla a Identificação se o pedido foi deferido ou indeferido, isto pode ser feito através de um programa de linguagem normal acoplado a um dicionário de dados (*thesaurus*), visto existirem várias palavras similares que tem o mesmo significado tal como: defere é igual a deferimento, dou seguimento, é devido, etc... E indefere é igual a indeferimento, nego seguimento, não é devido.

2.9 Abordagem Híbrida de classificação de dados

Em [48] é Proposta uma nova abordagem para tratar o problema de pequenos disjuntos, onde a classificação é feita através de regras, que fundamentalmente são utilizadas para identificar um pequeno número de exemplos. Para tratar o problema utiliza-se a técnica de árvore de decisão de maneiras a descobir regras de grandes disjuntos e algoritmos genéticos para evidenciar regras para pequenos disjuntos, i.e. forma normal disjuntiva. Também avalia o conhecimento descoberto em termos de qualidade consubstanciado no grau ou medidas de interesse dos utilizadores das regras descobertas. Para a descoberta de grandes disjuntos utilizam-se como técnicas os algoritmos de indução de regras, os quais ignoram os pequenos disjuntos, o que pode incorrer na redução de forma

significativa a precisão preditiva na classificação. O método híbrido proposto explora as características mais fortes das duas técnicas de descoberta de conhecimento, haja visto que uma é mais adequada para a descoberta de grandes disjuntos (i.e. técnica de árvore de decisão) e a outra para pequenos disjuntos (i.e. técnica de algoritmos genéticos). Nota-se uma ampla deficiência dos algoritmos de árvores de decisão em lidar com regras sempre que se atinge uma profundidade maior, porém este tratamento não representa nenhum problema difícil quando são utilizados algoritmos genéticos. Em geral esta é característica típica de todos os algoritmos evolucionários, pois trabalham com soluções alternativas ou candidatas em vez de envolver uma única solução como é feita na maioria das técnicas baseadas nas regras de indução. O sistema é composto por duas fases de treinamento. Na qual na primeira fase é executado o algoritmo de classificação C4.5 para a indução da árvore de decisão, a qual depois de podada é transformada em um conjunto de regras. Na segunda fase utiliza-se o algoritmo genético para descobrir regras exemplos que cubram os exemplos pertencentes aos pequenos disjuntos.

Para trabalhos futuros aponta a optimização dos algoritmos utilizados com adoptação de outros algoritmos de classificação da classe, ou ainda o valor dos parâmetros dos algoritmos genéticos independentes da base de dados a ser minerado. Por outra linha poderia implementar uma função de avaliação de ajustamento (*fitness)* que para além da predição levaria em consideração a questão de simplicidade, ou ainda através de algoritmos genéticos desenvolver uma função de avaliação de vários objectivos. E por fim outra linha poderia desprender-se em aprofundar a análise da relação entre a precisão preditiva e a simplicidade das regras descobertas e por fim perspectiva estender as experiências computacionais para incluir algoritmos baseados em instâncias mais complexos e o desenvolvimento do novo método para combinar técnicas de *emseble* ou relacionadas aos resultados de várias medidas data-*driven* de interesse de regras de uma com uma única medida global para estimar a subjectividade e o grau de interesse do utilizador nas regras descobertas de forma mais eficaz em relação as medidas individuais.

2.10 Determinação de Sentimentos

As pesquisas em volta de predição de sentimentos com a utilização do método de orientação semântica de adjectivos foram iniciadas por [32], um algoritmo não supervisionado foi utilizado em [84] para terminar a orientação semântica de termos individuais, o algoritmo inicialmente começa com 7 termos positivos e 7 negativos conhecidos, na sequência e com base nestes faz pesquisa com a utilização do motor de busca Altavista utilizando o operador NEAR para encontrar vários documentos que contém os termos. Diferentemente de *Pointwise Mutual Information* (PMI) *Score*, que oferece um grau para termo positivo e outro para negativo [85]. O *score* PMI de duas palavras W1 e W2 é dada pela possibilidade de ocorrência de duas palavras dividida pela possibilidade de ocorrência de cada uma em particular, conforme a fórmula a seguir:

$$PMI(w1,w2) = log\frac{P(w1,w2)}{P(w1),P(w2)} = log\frac{hits(w1,w2)}{hits(w1)hits(w2)}$$

A fórmula para a orientação semântica da palavra pode ser descrita da seguinte forma:

$$\boxed{SO - PMI(palavra) = PMI(palavra,p-query) - PMI(palavra-n-query)}$$

Onde : p-query representa os termos positivos e n-query os termos negativos. p.e. p-query = *bom ou agradável ou excelente ou afortunadamente ou correcto ou superior* e *n-query= mal ou desagradável ou negativo ou desafortunado ou incorrecto ou inferior.*

Os operadores OU e Near são oferecidos pelo motor de busca, embora o Near não é amplamente suportado, e por aproximação os valores utilizam números de *resultados* retornadas pela pesquisa e ignora o número de documentos no corpus (N), a fórmula segue abaixo:

$$SO - PMI(palavra) = log\frac{hits(palavra\ NEAR\ p-query)hits(n-query)}{hits(palavra\ Near\ n-query)hits(p-query)}$$

A orientação semântica de bigramas pode também ser determinada [85] em termos e frases e utilizada para determinar os sentimentos nas frases na sua plenitude. Para experiências foram tomadas 410 opiniões extraídas de opinions.com para avaliação de documento, na computação de determinadas frases em diferentes tipos de análise e

obteve-se como resultado 84 porcento de precisão na avaliação de automóveis e 66 porcento na avaliação de filmes.

2.11 Aprendizagem automáticas na Avaliação de Sentimentos

Um dos métodos utilizados para a classificação de documentos em positivos ou negativos consiste na utilização de algoritmos AM. Vários algoritmos são comparados em [60] e conclui-se que o algoritmo SVM representa melhor desempenho do que os demais, unigram, bigrama, POS-TAG. A posição de termos no documento é utilizada como característica. Entretanto a utilização de unigram apresenta entre todas técnicas os melhores resultados. Os algoritmos Bayesianas e Redes Neurais também são utilizados para a análise de documentos.

A classificação de sentimentos para análise de *feedback* de consumidor é apresentada em [28]. As variedades de características foram utilizadas em SVM na tentativa de dividir o conjunto de dados não apenas em negativos e positivos, mais associando um *ranking, pontuação* nomeadamente 1,2,3,ou 4, onde 1 significa não satisfazivel, 4 muito satisfazivel, o sistema proposto foi bom para distinguir as classes de 1 a 4 com 60 porcento aproximadamente de precisão estes resultados foram alcançados utilizando 2000 características seleccionados por tipo em função da razão.

2.12 Distinguindo Frases Objectivas e Subjectivas

O método para a exportação de subjectividade são apresentados em [89]. Na subjectividade a indicação inclui palavras baixa frequência, colecções, adjectivos e verbos identificados, utilizando a similaridade da sua distribuição. O processo inicialização é rico em aprendizado linguístico e extrai expressões subjectivas, com alta precisão do classificador de rótulos não anotados para criar um grande conjunto de

treinamento, que fornece a extracção de padrões com algoritmos linguísticos. Os padrões de aprendizagem então utilizam outras frases subjectivas. O método para distinguir frases subjectivas dos objectivos é apresentado em [58]. Este método é baseado na suposição de que as frases objectivas e negativas aparecem com mais frequência em princípio cada frase é associado a uma *pontuação* que indica se a frase é muito mais do tipo positivo ou negativo utilizando o classificador *Naive Bayes* para treinar um conjunto de dados anotada manualmente. O sistema ajuste a subjectividade baseando-se no fecho de outras frases subjectivas ou objectivas. Existem algumas melhorias neste método que não são significantes estatisticamente, e foi demonstrado que se pode reduzir o tamanho do documento, removendo frases objectivas mas descresse a qualidade da análise de sentimentos. Experiências similares são apresentados em [90]. O Nive é utilizado para descobrir frases com opiniões para treinar o rótulo do conjunto de dados.

O método combina muitos classificadores para a realização de algumas tarefas onde cada um destes focas em aspectos típicos, p.e. Nive foca em diferentes partes das características. O conjunto das características, POS-TAG, inclui unigram, bigrama, trigram, informações e polaridade, uma vez isto descoberto, i.e. se uma frase é objectiva ou subjectiva. Um simples classificador com características unigram foi utilizado para determinar o sentimento das frases.

2.13 Conclusões do Capítulo

Em tese grande parte dos trabalhos apresentados e actuais utilizam as técnicas de aprendizagem automática no contexto supervisionado, o que permite a classificação automática dos textos submetidos para treinamento e testes. As técnicas mais utilizadas em quase todas as abordagens são o *Naive Bayes* e Árvore de Decisão na realização de classificação, predição e para a obtenção de subjectividade e polaridade as técnicas POS-TAG e Bolsa de palavras. Alguns trabalhos utilizam técnicas de mineração e sistemas especialistas de forma a combinar os aspectos de classificação e aquisição de conhecimentos através da utilização de simulação e conceito de agentes inteligentes.

Capítulo 3

Mineração de Dados Estruturados e Não Estruturados

Realizar a mineração de Dados ou texto é uma tarefa que envolve a utilização de várias técnicas e métodos que auxiliam na sua prospecção, tratamento, transformação, sumarização e interpretação ou avaliação do conhecimento gerado com a finalidade ou propósito de apoio na tomada das decisões de formas a propiciar uma linha a seguir na gestão dos activos, pois além de predição do cenário para a evolução, fornece um estudo permenorizado que servirá de guia para futuras consultas e dimensionamento.

Este capítulo faz uma abordagem aos aspectos teóricos e discute o processo de mineração de dados e textos enquanto tarefa de descoberta de conhecimentos, cujo foco fundamental centra-se nos principais algoritmos utilizados na realização da classificação e agrupamento de texto.

3.1 Introdução

Os dados e informações são activos fundamentais para qualquer corporação, logo a sua gestão, tratamento e controlo de forma eficiente e eficaz, assim como a sua disponibilização para fins de análise e tomada de decisões são factores capitais e podem contribuir de forma diferencial para superar a concorrência. Observa-se com o advento

das TIs (Tecnologias de informação) um número elevado de dados e informações armazenadas nas grandes, pequenas ou médias empresas em diferentes suportes referentes aos seus processos, em geral em diferentes formas e formatos, o que dificulta o processo de extracção e descoberta de conhecimento porque os dados podem ter inconsistência decorrentes p.e. de erros de inserção, e outros factores que podem comprometer os resultados e desempenho do conhecimento a ser extraido, quando o tratamento é feito nos moldes tradicionais. As organizações e entidades produzem informações com extrema facilidade e velocidade expressiva o que porprociona que haja monumentos crecentes de textos e documentos de forma exponencial no atendimento das necessidades organizacionais. Nas organizações existem muitos dados supérfluos do ponto de vista de descoberta de conhecimentos. Os métodos e técnicas ou mecanismos de mineração actuais auxiliam o processo de descoberta de conhecimento através de mecanismos automáticos o que minimiza o processo que normalmente é árdua.

3.2 Descoberta de Conhecimento nas Base de Dados

Knowledge Discovery in Databases (KDD) é um processo de identificação de novos padrões válidos, úteis e compreensíveis num conjunto de dados armazenados numa Base de Dados (BD), através de empregue de técnicas e métodos específicos aplicadas em fases dependentes, geralmente este processo é realizado por especialistas com conhecimentos em diversas áreas [25]. O processo de descoberta de conhecimento da qual faz parte o processo de mineração de dados é abrangente e envolve três grandes etapas que agrupam as principais tarefas nomeadamente: pré-processamento, mineração e pós- processamento.

1. Pré-processamento: é a etapa em que ocorre a filtragem de dados de formas excluir dados que fundamentalmente não seriam tão relevantes para o contexto em questão e.g. eliminação de inconsistências nos dados em função dos erros decorrentes no registos e presumíveis exagero de registos nulos, podendo se necessário substitui-los por valores padronizados, através de operações de

processamento e a transformação dos dados que tem como objectivo converter-los para que tenham uma mesma semântica, assim como permitir um único formato de entrada para os algoritmos de mineração.

2. Mineração: consiste na aplicação dos principais algoritmos para a busca de conhecimento de formas a extrair padrões nos dados com vistas a se obter os resultados necessários e viáveis ao estudo que se realiza e requerem conjunção de conhecimento especializados de diversas áreas entre as mais importantes destacam-se: estatística, inteligência artificial e base de dados. Em síntese é o processo de exploração de grandes quantidades de dados para encontrar padrões de consistência p.e. regras de associações ou sequências temporárias para detectar relacionamentos sistemáticos entre variáveis, de formas a produzir novos subconjuntos de dados;
3. Pós - Processamento: observa-se a análise e interpretação do conhecimento descoberto, verificando por exemplo se os objectivos preconizados foram atingidos, podendo-se repetir o ciclo das etapas percorridas.

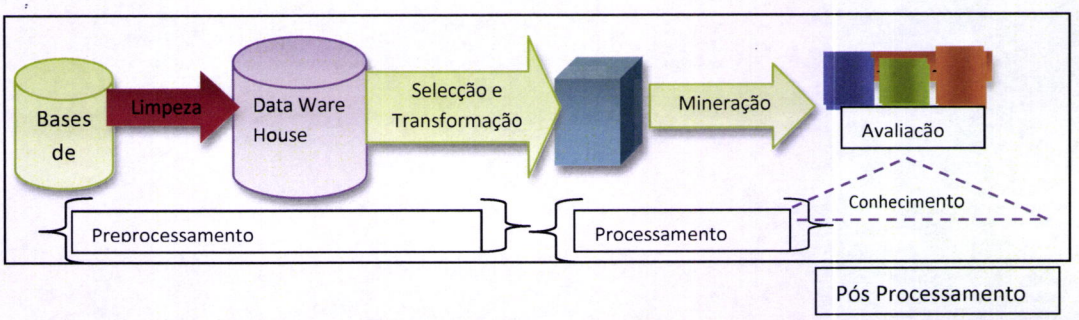

Figura 3.1: Processo de Descoberta de Conhecimento. Adaptada fonte: [54].

A análise dos resultados obtidos com a aplicação das técnicas de mineração são avaliados e interpretados com o objectivo de seleccionar o conhecimento considerado útil para os propósitos e as necessidades a que as tarefas pretendiam dar suporte ou seja em termos elementares para a sua utilização na tomada das decisões venha a contribuir para a elevação dos índices positivos da entidade.

1.6.1 Actividades de KDD

As principais actividades do processo de descoberta de conhecimento em base de dados são:

- Selecção de informações: que é realizada a partir de diferentes fontes de dados e.g. Base de dados, folhas de cálculos etc.;

- Pré- processamento: que visa limpar os dados p.e. retirar ruídos, eliminar registos nulos etc.;

- A transformação: que tem como propósito reduzir o tamanho de texto de formas a melhorar a eficiência da pesquisa;

- E por último a mineração que consiste na aplicação dos algoritmos aos dados cujos resultados são interpretados para a tomada de decisões.

Estas três primeiras fases fundamentalmente constituem o que é conhecido como processo de Data warehouse[3], cujos propósitos são criar condições para melhorar o tempo de resposta na realização de pesquisa de dados.

3.3 Mineração de Dados (MD)

A mineração de dados como já referenciado é a fase mais importantes do processo de descoberta de conhecimento, consiste fundamentalmente na aplicação dos inúmeros algoritmos especializados aos dados a serem analisados em função do propósito. É importante e já está a ser aplicada em larga escala em corporações como a *Google e Yahoo, YouTube*, *Faceboock* etc. Veja de que serviriam os dados captadas diariamente sem o processo de mineração, talvés a museus e até mesmo para incineração a cada cinco anos. Estudos recentes indicam que a *Walmart* conhecida como a maior empresa de *e-commerce*[4] do mundo é uma das organizações na vanguarda na aplicação de

[3] DataWareHausing (Armazém de Dados): local onde armazenamos os dados limpos, dados que os algoritmos vão consultar com intuito de aumentar a performance. Ao contrário se essas consultas fossem feita numa BD (Base de Dados) corportativo.

[4] E-commerce: consiste nas transacções comerciais efectuadas online entre pessoas ou entidades.

tecnologias de mineração de dados e consequentemente na aplicação dos resultados aos seus negócios.

O processo de MD reveste-se na transformação de grandes quantidades de dados em padrões e regras significativas, com vistas a proporcionar conhecimentos antes implícitos ou oculto, em função do contexto e propósitos da análise pode ser classificada em: direccionada ou não direccionada.

3.3.1 Mineração Direccionada

Diz-se que a mineração é direccionada quanto ela é feita em função de alguns atributos ou variáveis específicas, onde a avaliação é realizada de forma preditiva na obtenção do conhecimento preconizado, p.e. obter o valor fixo da conta de luz pago numa casa que se deseja alugar em função do valor pago nas outras residências do mesmo local.

3.3.2 Mineração Não Direccionada

A mineração é dita não direccionada quanto precisamos criar grupos de dados ou queremos encontrar padrões nos dados que dispomos, em geral este tipo mineração é tipicamente aplicado em dados censitários.

A figura 3.2 mostra as actividades da fase de MD nas quais e em função das necessidades de avaliação podem ser aplicados aos dados, as técnicas preditivas ou descritivas, as preditivas são aquelas que determinam a previsão ou a classificação de dados em conformidade com uma classe ou número predefinida ou determinado, a predição difere-se da classificação porque é feita em torno de números e não classe. As técnicas descritivas são associadas a métodos estatísticos onde se observa variáveis como moda, mediana, desvio padrão a media etc. de um determinado atributo, e a constituição de agrupamentos com base em propriedades.

3.4 Mineração de Texto

A mineração de texto é análogo a mineração de dados, com diferencial de que os dados utilizados para esta tarefa não são estruturados. Por tanto a sua finalidade é extracção de conhecimentos a partir de textos ou documentos disponíveis em várias fontes de dados.

A mineração tanto de dados como textos tem como principal papel filtrar, limpar os dados e converte-los em formatos apropriados aos algoritmos para extrair padrões adequados de formas a facilitar a interpretação. Utilizam na maioria das tarefas as mesmas técnicas e métodos, embora na fase de pré-processamento utilizem métodos e técnicas específicas. A MT difere dos mecanismos de busca porque nestes o utilizador tem plena consciência do resultado do objecto da pesquisa i.e. o resultado da busca, p.e.ver documentos que fazem abordagem a mineração de textos, a mineração de texto tem como principal foco descobrir conhecimentos ocultos nos textos, também não simula o conhecimento dos humanos como acontece com agentes inteligentes, pois utiliza algoritmos de mineração para a descoberta.

3.5 Algoritmos de Mineração de Dados

Para a realização do processo de mineração de dados são necessários algoritmos que realizam tarefas específicas t.c. predição, classificação, agrupamento, associação e sumarização. Para o processo as técnicas são aplicadas em separado ou de forma combinada na realização dos mais diversos propósitos, entretanto a escolha do algoritmo a utilizar depende substancialmente do propósito e da avaliação pretendida, os objectivos e metas a serem atingidos, assim como a natureza dos dados disponíveis.

3.5.1 Categorias de Algoritmos de Mineração de Dados

Os algoritmos de MD podem ser categorizados em função da tarefa a ser realizada em geral:

1- Classificação: consiste em construir um modelo com base em atributos conhecidos de uma determinada classe de objecto, em tese fornece uma nova classe para os novos dados obtidos através do processo em função de certas

variáveis ou parâmetros conhecidos p.e. podemos prever o sucesso na venda de um determinado produto a um grupo indivíduos ou classe social se conhecemos as suas preferências em função de compras de produtos semelhantes realizadas com base em critérios ou padrões de qualidade e outros parâmetros necessários para a predição dos resultados. Dois tipos de algoritmos de treinamento podem ser utilizadas para a classificação dependendo de amostras a utilizar. Se as amostras são conhecidas os algoritmos são denominados por algoritmos de classificação e de predição quando a classificação é feita com amostras não conhecidas e.g. árvores de decisão com regras, e métodos estáticos p.e. redes bayesianas. Imagine a classificação para melhor entendimento por exemplo que tenhamos vários atributos num modelo e dos quais é preciso seleccionar aqueles que darão maior ganho da informação. Fundamentalmente como se pode verificar para a realização da classificação utilizam-se os algoritmos de AM na categoria supervisionada (e.g. *Naives Bayes, KNN, redes MLP, SVM etc..*);

2- Estimativas: tem a ver com predição de resultados em função de certas variáveis contínuas tomadas como i.e. valores perdidos ou desconhecidos em contraste com a classificação que lida com valores discretos [1], como exemplo poderíamos destacar os estudos censitários de um determinado país ou região;

3- Associação: a associação consiste em inferir situações em que as preferências são moldadas de acordo com certos factores ou objectos. O exemplo clássico é "numa loja se informática seria preciso arrumar lado a lado portáteis e pasta de suporte naquelas situações na qual à venda não é inclusiva, de formas a permitir que quem compra um portátil tende faze-lo com a pasta, e para estimular as vendas conjuntas diversas acções e questões devem ser acauteladas", p.e. a proximidade, os preços etc.; p.e. se A,B então C. se alguém leva a e B a tendência é levar C.

4- Segmentação ou agrupamento (Clusters): técnica que consiste em agrupar registos de atributos de objectos em relação a sua similaridade, em contraste com a classificação onde o agrupamento é feita em classes pré-definidas. Por exemplo a título do exemplo anterior poderíamos dividir o grupo de compradores de um determinado produto em função das classes sociais. Em tese a tendência é encontrar grupos de documentos que não são conhecidos a priori, para a realização desta tarefa é necessário a utilização de algoritmos de AM na

vertente não supervisionado p.e. *K-means, Clustering Hierárquicos, redes SOM* etc;

5- Sumarização: a sumarização está consubstanciada fundamentalmente na realização de síntese que visa encontrar um subconjunto de dados a partir de um determinado conjunto, p.e. encontrar o desvio padrão para um dado universo de dados como forma de descobrir que dados não estão conformes a com padrão definido previamente.

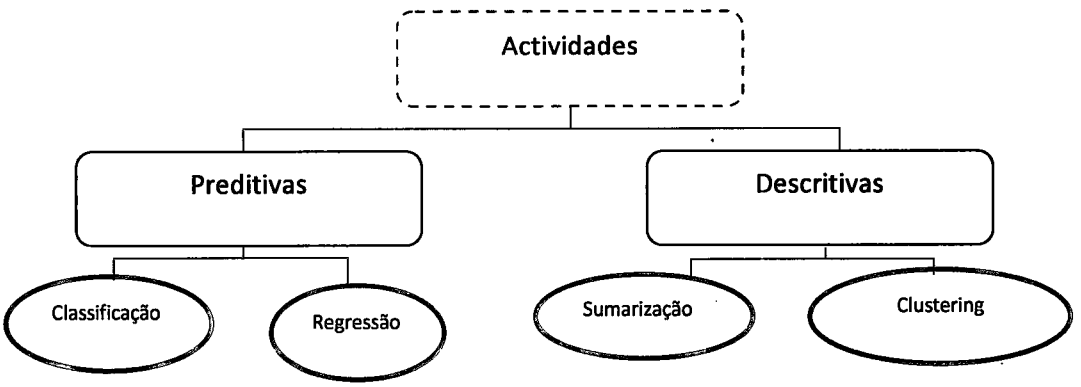

Figura 3.2: Tarefas de Mineração de Dados

3.6 Processo de Classificação de Dados

O processo de classificação em primeira instância gera-se um modelo como um produto intermediário, que incluem as regras para a classificação das tuplas da base de dados em torno de um conjunto de classes pré-determinadas, o modelo é gerado a partir de uma base de dados de treinamento, na sequência as regras geradas são testadas sobre uma outra base de dados, desta feita denominada de teste. Desta feita em função da comparação dos resultados obtidos com os dados de treinamento e os dados de teste a qualidade é mensurada de acordo com as quantidades de tuplas que o modelo consegue classificar de maneira satisfatória, as regras geradas caso sejam testadas na base de dados de treinamento, os resultados como é de esperar terão uma ampla probabilidade de correcção, haja vista que as regras foram extraídas a partir desta base de dados, és a razão pela qual para evidenciarmos a qualidade do modelo gerado é preciso que os testes sejam executados numa nova base de dados completamente diferente da utilizada para o treinamento, p.e. um candidato para um processo selectivo poderia ser seleccionado se a sua média fosse igual ou superior a quatorze, com este modelo de

treinamento, todos as avaliações posteriores com esta média ou acima convergiriam para a selecção, portanto o atributo de classificação neste exemplo é a média.

3.6.1 Processo de Mineração de Texto

Existem várias construções de soluções para gestão e recuperação dados através de mecanismos de buscas por registos estas tecnologias do ponto de vista de dados estruturados são funcionais, a maioria dos dados disponíveis são semi-estruturados ou não estruturados. E nesse caso surge a necessidade de mecanismos de buscas inteligentes ou avançados que permitam extrair dados relevantes e consistentes para os propósitos desejados. O processo de MT envolve cinco etapas fundamentais a serem realizadas sequencialmente para a consumação da análise pretendida, conforme a representação ou modelo figura 3.3.

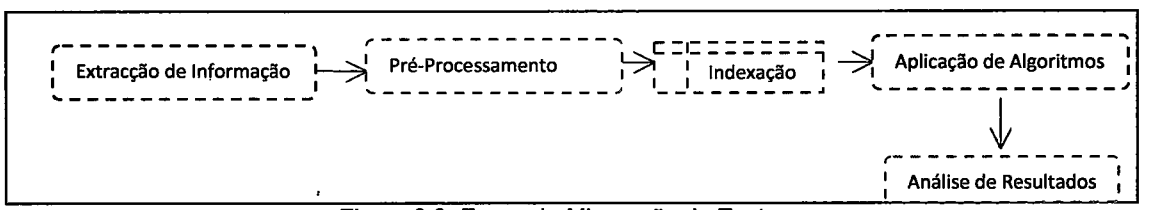

Figura 3.3: Fases de Mineração de Texto

3.6.1.1 Mecanismo de extracção de Texto

A tarefa consiste em identificar num documento de texto, os dados relevantes aqueles que contribuem normalmente para o ganho da informação. Geralmente para a extracção de informações utilizam-se as técnicas de engenharia de conhecimento, Processamento de LN e AM. A extracção pode ser feita manualmente ou de forma automática (Crawling)[5]. Existem três abordagens para a realização de extracção de informações [3;4]: baseada em regras; baseada em AM e a baseada em dicionário.

1. Extracção baseada nas regras: consiste na identificação de padrões através de utilização expressões regulares que são extremamente simples, mas as regras que utilizam são específicas a um determinado domínio, daí a dificuldade de adaptá-las a outros domínios;
2. Extracção baseada nas técnicas de AM: utiliza classificadores para identificar frases relevantes para o contexto em questão, as técnicas permitem reconhecimento automático de texto, para tal é necessário a utilização de um

[5] Crawling: é uma ferramenta típica de colecta de textos com capacidade de intensificação dos links mais relevantes

conjunto de dados de treinamento, com a desvantagem de que o melhor desempenho e precisão sejam alcançados a custo de um conjunto de dados de treinamento da classe equivalentes em termos quantitativos ao de outras classes, sob risco de comprometer os resultados obtidos na classificação;

3. Extracção baseado no dicionário: utilizada para o reconhecimento automático de texto, cujo princípio de funcionamento consiste em localizar a ocorrência de um determinado termo na lista contido no dicionário, cuja grande desvantagem encerra-se na probabilidade de que o termo pesquisado não estar contido no dicionário.

3.6.1.2 Pré-Processamento de Texto

Etapa que consiste adequar o texto a padrões de formatação e representação tarefa realizada com apoio de vários algoritmos ou através da combinação destes, haja visto que os textos extraídos contêm informações irrelevantes, natureza e formatos diferentes. Entre as técnicas aplicadas para o pré-processamento do texto podemos citar a tokenização que consiste em extrair *tokens* a partir de um determinado texto. Token pode ser uma palavra ou estar associado a mais de uma palavra, símbolos de pontuação ou caracteres especiais do nosso alfabeto, através da utilização da abordagem do processamento de linguagens naturais (PLN). Uma abordagem amplamente utilizada para este processo consiste em um dicionário para fins de interpretação dos termos em LN [25], normalmente este dicionário tem informações morfológicas, sintácticas e semânticas e as palavras associadas ao problema a ser tratado. O resultado desta fase normalmente consiste num vector de atributos que pertencem os documentos submetidos a fase [34]. Este vector pode ser representado ou construído através de duas abordagens fundamentais: modelo Bolsa de Palavras na qual as palavras de um documento de entrada são representadas como atributos do vector. O tamanho do vector é proporcional ao número de palavras contidos em cada documento, é associação a pesos binários a cada atributo que especifica a ausência ou presença de uma palavra no documento; TF-IDF (Term *Frequency- Inverse Document Frequency*) considerara a frequência da palavra no documento ou em todos os documentos, dado uma palavra w contido num dado documento d o peso TF-IDF é definido pela expressão:

$TF - IDF(w,d) = Tf(w,d).\log(N|Df(w))$, onde Tf (w,d) corresponde a frequência da palavra no documento, N é o número total de documentos e Df (w) é o número de documentos que contém a palavra w [25]. Outra forma de representação do documento numa bolsa de palavras é o modelo do espaço vectorial, onde os documentos e as respectivas consultas são representadas como atributos do vector de entrada os quais representam os termos contidos nos documentos, desta forma cada atributo é associado a um peso que é um valor numérico do termo de acordo com a frequência da sua ocorrência nos documentos por exemplo um termo que aparece no documento 5,7,13,2 e 10 vezes, seria representado no vector como: $\vec{d}j = (5,7,13,2,10)$, onde $\vec{D}j$ representa o espaço vectorial dos termos do documento. Em termos gerais a colecção pode ser representada de seguinte forma: $\vec{d}j = (w1,j,w2,j,..,wn,j)$ onde $\vec{d}j$ representa um documento de uma colecção e o vector contém os pesos de toda a colecção, logo w1,j refere-se que o termo tem peso 1 no documento de entrada j [37]. A pesquisa c pode ser representada analogamente ao vector do documento como segue: $\vec{c} = (1,1,1,0,0)$, em termos genéricos $\vec{c} = (w1,c,w2,c,..,wn,c)$. Em suma as principais tarefas desta fase são:

- Filtrar informações;
- Redução de dimensão do documento;
- Limpeza no documento;
- Tokenização: dividir o texto ou palavras contidas no texto em sua unidade menor;
- Remoção de stopword: consiste em remover palavras ou termos sobretudo empregues para dar realce ao texto mais que não agregam conhecimento útil a análise;
- Redução ao radical (stemming): consiste fundamentalmente na remoção de sufixos para converter as diferentes palavras no seu radical.

3.7 Processamento de Língua Natural

O PLN é uma subárea de inteligência artificial com vasta gama de aplicabilidade entre as quais destacam-se: a extracção de informação, a recuperação de informação, a tradução automática, a geração automática de texto, a geração de linguagens naturais e a sua interpretação, a simplificação de texto, a correcção ortográfica, o reconhecimento de voz e o reconhecimento de manuscrito. O PLN possui três tarefas fundamentais nomeadamente: a Análise léxica; a Normalização e a Análise Sintáctica.

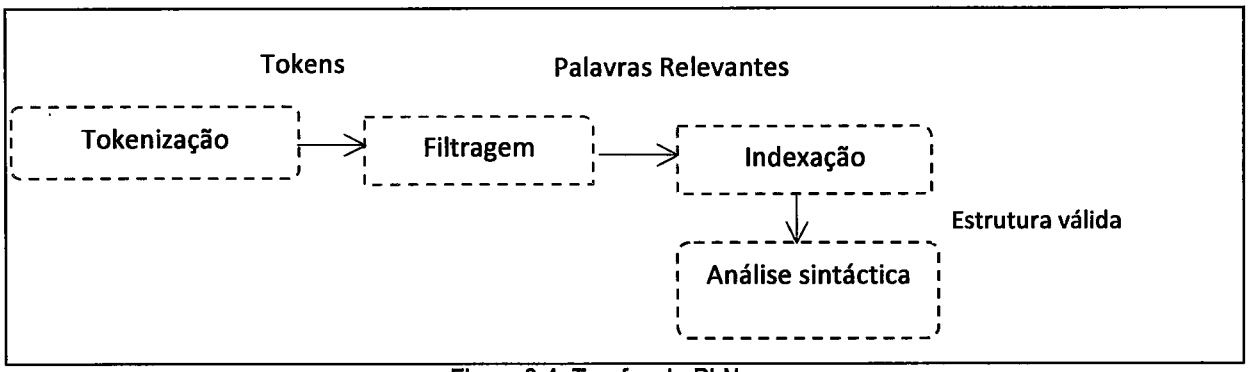

Figura 3.4: Tarefas de PLN

A figura 3.4 representa as tarefas que são realizadas no processo de processamento de PLN, onde em cada uma das etapas o texto é tratado e transformado em unidades chamados tokens, e posteriormente filtrado ou normalizado com vista a obter apenas unidades relevantes de maneira que permitam a eficiência na recuperação e busca da informação e por fim faz a análise sintáctica que visa fornecer estruturas válidas (chunks[6]) em função da gramática da linguagem alvo.

3.7.1 Tokenização ou análise léxica (Tonkenization)

É a tarefa que consiste na identificação e separação das unidades nas frases da linguagem e a sua associação a atributos ou traços gramaticais, assim como a segmentação do texto. As unidades são denominadas por tokens e comumente cada uma representa palavra ou símbolos de pontuação. A seguinte frase por exemplo a nível de segmentação e atomização ou tokenização seria subdividida nas seguintes unidades léxicas. A seguir descrevemos exemplos de segmentação e tokenização da seguinte frase: "O projecto começou a ser estruturado em 2012. O seu término está agendado para o primeiro trimestre de 2013."

- **Segmentação do texto**: [O projecto começou a ser estruturado em 2012.] [O seu termino está agendando para o primeiro trimestre de 2013.];
- **Tokenização do texto**: [O][projecto][começou][a][ser][estruturado][em][2012][.][O][seu][termino][está][agendando][para][o][primeiro] trimestre][de][2013][.].

3.7.2 Normalização do documento

[6] Chunks: representa a divisão do texto em estrutura sintacticamente relacionada

Consistem em fazer filtragem ou remover palavras irrelevantes para a acção de busca *(Stopword)*[7], i.e. são palavras que não tem grande influência ao serem removidos do texto, a busca tornana-se mais ágil por diminuir substancialmente a quantidade de palavras a serem pesquisadas, por não alterarem fundamentalmente o sentido da busca e muito menos os resultados esperados, são consideradas palavras irrelevantes, pertencem a este classe e.g. artigos, conectivos (i.e. conjunções, proposições etc.) e pronomes, assim como aquelas palavras que ocorrem repetidamente no texto. Ainda nesta fase as palavras são reduzidas para o seu radical gramatical *(lematização ou stemming)*[8], identifica-se a radical da palavra através da eliminação de sufixos e prefixos. A radicalização aplica-se tanto a verbos, adjectivos e assim os substantivos são identificados.

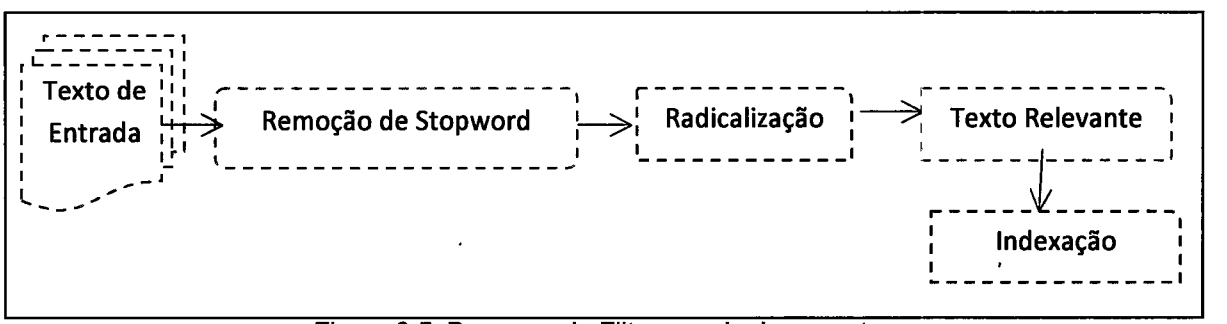

Figura 3.5: Processo de Filtragem de documento

A figura 3.5 mostra as etapas de filtragem de documentos e textos, onde numa primeira fase faz a remoção dos *stopwords*, na sequência a etapa de radicalização e consequente obtenção de texto relevante, livre de partes que cuja sua remoção torna as pesquisas mais eficiente, e por último a fase de indexação.

3.7.3 Análise sintáctica

Consiste na construção ou recuperação de estrutura sintáctica válida para as frases de entrada com base na gramática da língua em questão, porém a análise utiliza quase sempre uma gramática parcial pela mesmas razões invocadas na fase de filtragem, justificada nesta fase por contemplar todas as construções válidas para a aplicação,

[7] Stopword: palavras consideradas irrelevantes param a busca e portanto a sua indexação aumentaria o tamanho do índice, e tornaria a busca mais lenta.
[8] Stemming: consiste na normalização morfológica da palavra, onde a palavra é reduzida ao seu radical que é incluída ao índice.

evitando de tal maneira o volume de informações gramaticais que podem aumentar a complexidade na representação, assim como na análise propriamente dita. Como exemplo dada a seguinte frase a análise sintáctica pode ser da seguinte forma: "Nós estamos a fazer um projecto".

1. Nós: sujeito, pronome pessoal (plural);
2. Estamos: verbo;
3. A: determinante;
4. Fazer: verbo;
5. Projecto: substantivo.

3.8 Mecanismo de Indexação

O mecanismo de indexação é fundamentalmente quando estamos a falar em colecções de textos enormes, porque dificilmente uma pesquisa sequencial daria um retorno eficiente e satisfazivel nos moldes sem a indexação dos termos, por exemplo grande parte de pesquisa feita é *online* e parte dos documentos são semi-estáticos e não dinâmicos p.e. dicionários, enciclopédias, livros etc. sem índices associados aos termos fundamentalmente para acelerar o processo de busca, é uma actividade árdua e muitas vezes imprecisa. A indexação é o processo de geração de índices com base nos termos armazenados na base de dados textuais, associado índices a cada termo, como maneira de agilizar o processo de recuperação de texto ou dados.

3.9 Recuperação de Informação (RI)

Subárea de informática que utiliza algoritmos específicos para recuperar informações ou documentos contidos nalguma fonte de dados em função de entrada com palavras-chave, e desta feita os documentos que contém tais palavras são recuperados e servirão para a mineração e análise e.g. recuperam-se tabelas relacionais, documentos, ou textos publicados na Web [2]. O processo de recuperação inicia com a representação, e armazenamento, estende-se até a organização e a aquisição ou acesso a informação. Há uma sútil diferença entre recuperação de dados e informações, a recuperação de dados consiste no processo na qual se pesquisa numa colecção de dados utilizando palavras-chave ou frases e como resultado da busca é retornado um conjunto de dados associados a pesquisa, trabalha-se fundamentalmente com dados estruturados i.e. aqueles que

possuem sintaxe e semântica bem definida. Para dar exemplo vejamos a pesquisa para encontrar alunos que obedeçam os parâmetros abaixo: Buscar todos os alunos que: 1)- Foram licenciados em 2010 e; 2)- Que tenham alcançado médias superiores a 14, e deve agregar informações sobre as suas regiões de origem e tendências culturais e níveis sociais e 3)- Que tenham terminado em tempo regular. Como se pode verificar a frase está formulada na LN, logo para a recuperação de dados é preciso, transformá-la ou convertê-la numa linguagem de consultas (*queries*) de maneira a ser processada pelo sistema de busca. Esta mesma consulta poderia ser efectuada em sistemas de RI através do método na qual as informações relevantes e consistentes são extraídas do texto. O que torna possível o trabalho com formato de textos simi-estruturados ou não estruturados e muitas vezes com semântica ambígua. Nos sistemas de RI é passível trabalhar-se com o documento completo porém por efeitos de racionalização recursos atendendo os custos, durante a geração algumas transformações são efectuados p.e. a eliminação de *Stopwords,* redução das palavras a sua raiz gramatical (*Stemming),* identificação de substantivos e retirada de adjectivos, advérbios e verbos e para se tornar a pesquisa mais ágil, gera-se o índice num processo onde os dados são extraídos numa base de dados e através da utilização de operações de texto os quais resultam num documento e em seguida constrói-se o índice em função deste.

3.9.1 Classificação de Textos

A detecção de padrões é a parte central de PLN, os padrões observáveis na estrutura do texto e em função da frequência de ocorrência são utilizados para estabelecer e correlacionar as palavras com aspectos particulares de significados. É a tarefa ou acto que tem como objectivo associar um determinado documento à sua presumível categoria, funcionalmente inicia com a extracção de texto e na sequência faz-se a identificação das características do texto e por fim seu mapeamento para uma determinada classe pré-definida, i.e. com base nas características ou conteúdo o documento é atribuído a categoria correspondente, de formas a facilitar a tarefa de análise. A classificação pode associar o documento a mais de uma categoria ou ainda a nenhuma, pode ser manual ou automática. A manual é um processo que envolve especialistas para a sua realização, logo envolve altos custos, esforço, falta de eficiência etc. Já a automática apoia-se nas técnicas de AM, os resultados obtidos com as técnicas de categorização automática de texto tem-se mostrado bastante eficientes, embora a questão de precisão ainda seja crítica de uma técnica para outra. Sistemas de

classificação automática podem ser utilizados para realizar pré- filtragem como forma de auxiliar o especialista na sua actividade. Podemos classificar utilizadores, textos, sentenças (parágrafos, chunks de textos), predeterminar frases descritivas (<ADJ N>, <N N>, <ADV ADJ>, etc), palavras, Tweets/updates etc. Na Classificação de Palavras ou Frases Pequenas são construídos blocos de expressões sentimentais, frases pequenas podem ser importante como palavras: baixo preço, alta qualidade, necessitamos de uma abordagem para lidar com estes antes de seguir para outras tarefas de classificação.

3.10 Processo de Classificação de Documentos

Na década passada uma grande atenção estava voltada para o problema de classificação e muitos classificadores desenvolvidos baseiam-se na AM ou nos Métodos Estatísticos. Entre os quais temos a destacar o classificador Roccio que é extremamente influente na área de RI. Devido à importância e as vantagens que reveste a classificação automática como já frisado, deu-se origem a um problema na construção de classificadores de texto ao serem direccionados para múltiplos contextos, desta forma apenas um número pequeno teve êxitos, proporcionando ferramentas de classificação poderosas. As técnicas utilizadas na construção destas ferramentas podem ser divididas em algumas subclasses nomeadamente: Probabilística, Árvores de Decisão; Regras de Decisão; Modelo de Regressão; Redes Neurais; Support vector machine; e etc. Geralmente o sistema de categorização de texto constrói um modelo de representação vectorial. O vector representa o documento e contém os termos e os pesos atribuídos a cada termo do documento, a maneira como os pesos são atribuídos aos termos é um problema que está fora do âmbito déste trabalho.

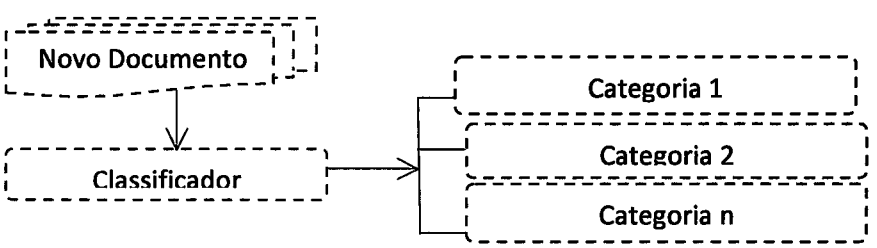

Figura 3.6: Classificação de documento

A figura 3.6 mostra o esquema de classificação do novo documento, onde o classificador mapeia o documento na sua categoria, em função do seu conteúdo ou características, p.e. documentos com conteúdo ou características do desporto, e política seriam agrupados em categorias desporto e política respectivamente.

3.11 Representação de Documentos

A representação de documentos é uma das técnicas de processamento que é utilizada para reduzir a complexidade do documento de formas a tornar a sua manipulação mais fácil, para o efeito o documento deve ser transformado completamente no formato vectorial. A representação é uma etapa crucial e muito importante na classificação do documento, este denota de forma compacta o mapeamento do documento em relação ao seu conteúdo. Um texto tipicamente é representado como um vector dos pesos dos seus termos (i.e. características das palavras) de um conjunto de termos (i.e. dicionário) onde cada termo ocorre pelo menos uma vez num determinado número mínimo de documento. Um dos principais problemas característicos de classificação de texto está ligado a dimensionalidade extremamente alta de dados de texto, o número de características potenciais, muitas vezes ultrapassa o número de documentos de treinamento. A definição de um texto consiste num conjunto de partes (termos) relacionadas que possuem vários padrões de ocorrências. A classificação é importantíssima em várias tarefas de gestão. Com o aumento de dados na Web tornam-se imprescindíveis os algoritmos de classificação porque podem melhorar eficientemente a análise com a manutenção da precisão e são amplamente desejados [11]. Pré-Processamento de documentos ou redução de dimensionalidade permite uma efectiva manipulação e representação de dados, as discussões nesta área produzirão ao longo dos últimos anos muitas propostas de modelos e técnica. A redução de dimensionalidade é extremamente importante na etapa de classificação, porque características redundantes e irrelevantes são eliminadas nesta etapa. As técnicas de pré-processamento podem ser classificadas em: técnicas Extracção de Características (EC) e Selecção de características (SC).

3.11.1 Técnicas de Extracção de Características

O processo de processamento é feito com base no quadro da estrutura de cada linguagem, desta forma eliminam-se muitos factores dependentes da linguagem, tokenização, remoção de *topwords* e *stemmings*. EC é o primeiro passo no processo de pré-processamento e tem a tarefa de condicionar o documento para melhorar a rapidez na sua manipulação, busca ou pesquisa. Os documentos de texto são representados com aumento de várias características, muitas vezes irrelevantes ou ruido. EC é exclusivo a várias palavras-chave, base preferível no processo estatístico para criar um vector com baixa dimensão. A técnica tem atraído bastante atenção porque reduz efectivamente a dimensão, o que torna a tarefa de aprendizagem mais eficiente e economiza mais o espaço de armazenamento. Formalmente os passos para a extracção de características são:

1. *Tokenization*: o documento é tratado como uma cadeia e particionado na lista de tokens;
2. *Removing stopwords*: Stopwords, são termos que ocorrem com muita frequência e são insignificantes para a estrutura das frases, p.e. artigos e conjunções t.c. "um(a)", "e"... etc , logo podem ser removidos;
3. *Stemming word*: a aplicação de algoritmos desta natureza converte as diferentes palavras nas suas similares formas canónicas. Este passo converte os tokens para as suas raízes e.g. conectiva para conectivo e computação para computação.

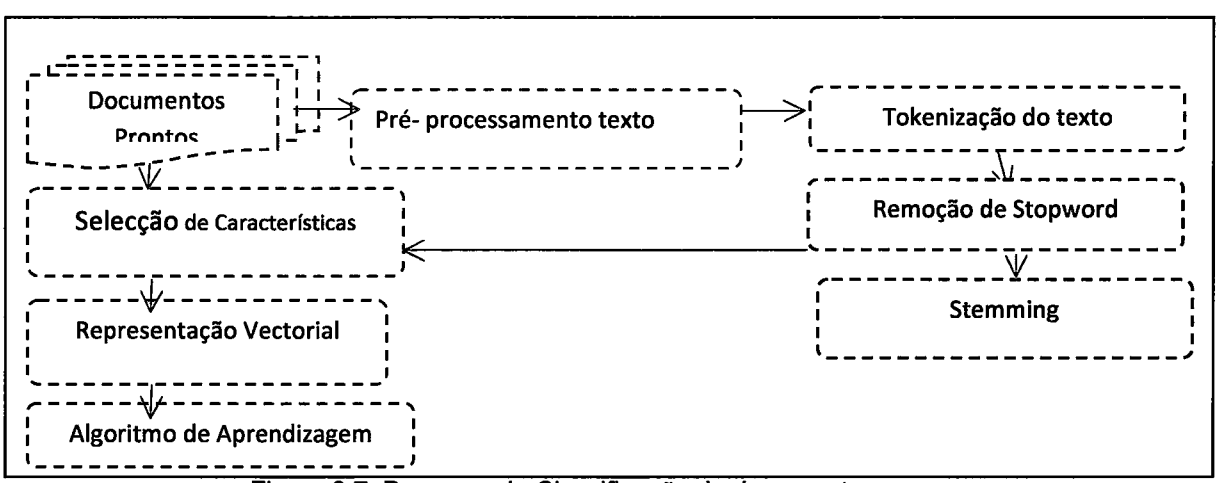

Figura 3.7: Processo de Classificação de documento

A figura 3.7 apresenta a classificação de documento, onde documentos extraídos são submetidos a filtragem, e a representação vectorial e por fim a etapa de classificação com a execução do processo de treinamento de acordo com o algoritmo desejado para a tarefa a realizar.

3.11.2 Técnicas de Selecção de Características

Depois de extracção o próximo passo de pré-processamento é selecção de características, cujo processo prende-se na construção de um espaço vectorial que visa melhor a escalabilidade, eficiência e precisão de um classificador de texto. Em geral um bom método de selecção de características deve considerar o domínio e as características do algoritmo.

3.12 Técnicas de Categorização

Existem várias técnicas de análise que podem ser utilizadas na análise de sentimento de forma individual ou combinadas, as quais podem ser agrupadas em duas categorias fundamentais [18]: 1)- Técnicas de Aprendizagem Simbólicas e; Técnicas de Aprendizagem automática.

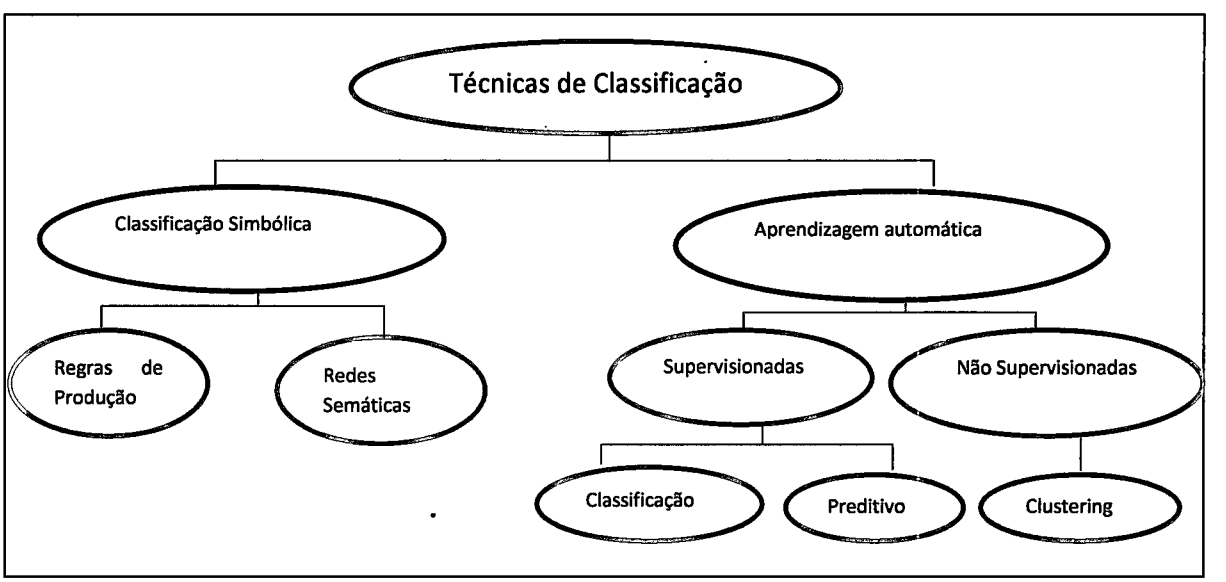

Figura 3.8: Técnicas de categorização: Adaptado de Monard e Baranauskas [53].

A figura 3.8 apresenta as diferentes categorias de técnicas e subcategorias que podem ser utilizadas para a classificação de texto ou documentos. Para este projecto foram adoptadas as técnicas de AM supervisionadas para mais detalhes ver sessão 5.2, a abordagem classificação simbólica esta fora do contexto deste projecto.

3.12.1 Técnicas Simbólicas

Tem o seu fundamento centrado na análise do léxico visando determinar a polaridade dos sentimentos e o estabelecimento de relações entre os componentes que compõem a estrutura com assuntos e frases. São tipicamente representados em expressões lógicas, regras de produção ou redes semânticas, como exemplo da aplicação podemos citar agentes inteligentes e árvores de decisão.

3.12.2 Aprendizagem automática

Aprendizagem é definida como resultado obtido de qualquer mudança no comportamento de um sistema referente à melhoria no seu desempenho quando uma tarefa é replicada ou repetida, ou quando outras tarefas são executadas com a mesma amostra [32].

AM é uma subárea de IA responsável pelo desenvolvimento de teorias computacionais focadas na criação de conhecimento artificial, na qual os sistemas desenvolvidos com base na abordagem possuem uma característica principal de tomarem decisões em função dos conhecimentos prévios acumulados através de interacção com o ambiente. Em tese a filosofia ou princípio do seu funcionamento é baseado em inferir o comportamento associado aos dados ou informação através de modelos treinados, que são criados com dados já processados com a finalidade de serem utilizados para a identificação de padrões similares nos dados a avaliar. Podemos dizer que um programa aprendeu se dado uma experiência (E) feita num dado contexto ou universo de tarefas (T), e a medida de desempenho obtido (P) se somente se o seu desempenho com as tarefas medido através do desempenho, melhora com a experiência[85]. Exemplo num contexto de pagamento de propinas pode-se detectar se os novos alunos são bons ou inadimplentes.

- Tarefa: classificar os potências novos alunos como bons ou inadimplentes;
- Medida de desempenho: percentagem de instâncias i.e. alunos classificados correctamente;
- Experiência de treinamento: uma base de dados na qual os alunos já conhecidos são previamente classificados como bons ou inadimplentes.

A AM pode ser categorizada ou classificada em duas categorias fundamentais: Supervisionado e não Supervisionado.

3.9.2.1 Aprendizagem supervisionada

A classe a qual pertence o conjunto de dados ou documento analisado é previamente conhecida. Para este contexto na análise não é considerada as opiniões neutras, como na parte dos estudos que a antecedem. A principal tarefa na classificação de sentimento é a escolha de um conjunto adequado de atributos, em geral utilizam-se nestes tipos de abordagem os algoritmos *Naive Bayesian e Support Vector Machines,* Máxima Entropia.

3.9.2.2 Aprendizagem não supervisionada

Os algoritmos utilizados para o efeito assumem sempre que não se conhece a classe a qual pertence o conjunto de dados analisados em contraste com os utilizados na aprendizagem supervisionado. Também pode ser denominado por segmentação ou agrupamento (clustering), como existe incerteza entre os resultados esperados normalmente utilizam-se como algoritmos para a realização da tarefa as redes de bayesianas, k-medias, fozzy, c-medias e mapas auto-organizáveis (SOM).

3.13 Algoritmos de Aprendizado Supervisionados

Existem muitos algoritmos de aprendizagem supervisionado utilizados na classificação de textos entre os mais importantes podemos destacar: Máxima Entropia, *Naive Bayes*, Support vector machine (SVM), K-Nearest Neighbors (K-NN - vizinhos mais próximos), Árvores de Decisão etc.

3.13.1 Máxima Entropia

É um método da classe dos algoritmos de aprendizagem supervisionados baseados em corpus, estes modelos busca a óptima distribuição de probabilidade assumindo a máxima ignorância dos dados possíveis, i.e. não se assume nenhum conhecimento que não estiver reflectido no conjunto de treinamento. A sua principal vantagem reside na sua capacidade de representação de fontes de informação de contexto heterogéneo. Trata-se de uma técnica alternativa incorporada as aplicações PLN [41], mostrou que a performance nos resultados obtidos com a utilização do algoritmo em alguns casos apresentam uma boa precisão noutros casos não se pode dizer o mesmo. Para a análise a probabilidade ou estimativas se baseia numa fórmula exponencial. Como exemplo

imagine quatro formas de classificação do texto para o maior entendimento da máxima entropia, se tivermos quatro classes diferentes e dizermos que apenas em média 40% dos documentos com a palavra Professor estão na classe Faculdade. Intuitivamente quando temos um documento com a palavra professor contido neste podemos dizer que tem 40% de probabilidade de pertencer a classe em questão e 20% de probabilidade para cada uma das três classes. Caso o documento não tem a palavra professor deverá ser feita uma distribuição uniforme com 25% de probabilidade para cada uma delas. Este é o princípio do funcionamento do modelo Máxima Entropia, o modelo é muito simples na prática para ser utilizado, mais quando temos várias restrições requer técnicas mais rigorosas para obtermos a solução óptima [41]. Máxima Entropia baseia-se no cálculo de distribuição de probabilidade p (X, Y) com entropia máxima que satisfaça as restrições definidas no corpus utilizado para treinamento, o processamento consiste em escolher a máxima entropia entre as possíveis distribuições de probabilidades dentro do corpus, em função desta escolha é possível determinar a distribuição de probabilidade conjunta de determinadas características contidas no texto que representa sentimentos. Em vista da qual podemos determinar se o sentimento expresso num texto é positivo ou negativo. Imaginemos um conjunto de treinamento $S = \{x1, y1, x2, y2, x3, y3, .., xn, yn\}$, com n exemplos de treinamento o modelo pode ser determinado através da seguinte fórmula S=$\sum_{x,y} p(x,y) = 1$, onde p representa a distribuição de probabilidade conjunta. Entretanto como várias distribuições podem satisfazer a condição representada pela equação dividindo o corpus em unidades menores tais como frases, palavras, facilmente cada unidade pode ser descrito por uma função dimensional de valores reais do vector de características [14], para análise de sentimentos podemos ter os pares unidade e Tag e isto é dado pela função:

$$f(x,k) = \begin{cases} 1, & \textit{se } x \textit{ é positivo e } k \textit{ é um adjectivo} \\ 0, & \textit{caso contrário} \end{cases}$$

Onde x representa uma unidade qualquer e k representa um adjectivo.

3.13.2 Naive Bayes

É um método de aprendizagem supervisionada baseada nos métodos de classificação estatística, a classificação é tida como simples ou ingénua por não considerar as

dependências possíveis entre os dados. É um dos classificadores probabilístico mais simples, baseia-se na aplicação do Teorema de Bayes, os atributos tem ampla independência em relação a variáveis de uma classe i.e. as variáveis do domínio são fortemente independentes entre si, desta maneira torna a ordem das características irrelevantes porque nenhuma característica está relacionada a outra e consequentemente a presença de uma característica não afectará outras na tarefa de mineração. As estatísticas Bayesianas podem ser utilizadas para avaliar que modelo ou família de modelos é mais adequado para descrever certos dados. Assume um modelo probabilístico subjacente o que permite capturar a incerteza sobre o modelo em função de parâmetros desconhecidos através da associação a distribuição de probabilidade para obter os resultados, desta forma pode solucionar problemas de diagnóstico e predição. O modelo fornece uma perspectiva útil para o entendimento e avaliação de muitos algoritmos de aprendizagem, calcula a probabilidade explícita em função de uma determinada hipótese sobre a forma de distribuição de probabilidade. Está entre os algoritmos com mais sucesso na classificação de textos. É adequado sobretudo quando a dimensão das entradas é alta. Dependendo da precisão do modelo de probabilidades o classificador pode ser treinado de forma ou maneira mais efectiva, por requer um pequeno número de dados de treinamento. Fórmula para calcular a Probabilidade da categoria dado o documento $P(C|D) = \frac{P(C)}{P(D)} P(D|C)$ que representa a probabilidade de ocorrência de um documento dado que ocorre a categoria i.e. dada um documento, qual é a probabilidade de pertencer a uma determinada categoria. Desta maneira podemos observar que:

- P(C) é a probabilidade da ocorrência da categoria, e representa o resultado do quociente do número de documentos na categoria dividido pelo número de documentos na base de treinamento;
- P(D) representa a probabilidade do documento ocorrer, e resulta da multiplicação das probabilidades de cada termo que compõe o documento ocorrer na base de treinamento;
- P(C|D) representa a probabilidade da ocorrência do documento dado que a categoria ocorreu, e é obtido através do resultado do produto das probabilidades condicionais que compõem o documento.

Por exemplo dado um conjunto de documentos de treinamento este é considerado como uma lista ordenada de palavras e pode ser representado por $Wd_{i,k}$, para denotar a posição k da palavra no documento d_i, onde cada documento tem a forma de vocabulário $V =< w1, w2, ...w|y| >$. O vocabulário é o conjunto de todas as palavras consideradas na classificação. Também considera-se um conjunto de classes pré-definidas padrões $C = (c1, c2, ... c|c|)>$. Nesta óptica para melhorar a computação necessitamos a probabilidade a priori $\mathbf{Pr}[cj|di]$. Donde C_j é a classe e d_i é o documento. Portanto é necessário fazer o treinamento do classificador com os documentos de treinamento, para que se possa fazer subsequentemente a classificação de novos padrões com base nos resultados anteriores.

3.13.3 Support vector machine (SVM)

É uma técnica de aprendizagem automática amplamente utilizada. A classificação de texto pode ser efectuada manualmente com base nos critérios multiníveis, porém a automatização desses procedimentos facilita o processo da classificação cujo princípio é efectuar a classificação automática de documentos tarefa que é realizada efectivamente com o suporte ou auxílio dos algoritmos de classificação supervisionado conhecidos como SVM, o processo consiste em encontrar variáveis que sejam úteis e relevantes na descriminação de textos existentes associados a diferentes classes. Os sistemas que fazem a classificação automática são treinados com documentos previamente classificados e etiquetados de acordo com alguns tipos de critérios por exemplo assunto, matéria, origem, entidades nomeados etc. em conformidade com uma classe, entretanto o objectivo primordial da classificação é decidir em qual classe um novo texto ou documentò será incluído ou associado de acordo com um esquema de classificação previamente estabelecido. Imaginemos que agrupemos todos os documentos anteriores relacionados aos comentários sobre desporto na classe deporto e todos os documentos ligados a política na classe política, e sumariamente tenhas que incluir novos documentos ligados ao Desporto ou Política, esse algoritmo viabilizaria esta classificação dos documentos e associaria-os aos seus respectivos grupos com a utilização deste método. Considerado por muitos como o melhor classificador de texto fruto do desempenho que oferece, razão pela qual é a mais utilizada neste contexto. Em síntese pode ser concebido como um processo de aprendizagem estatístico durante a qual o algoritmo de classificação captura as características que distinguem cada

categoria de documento das demais ou seja aquelas características peculiares que um determinado documento deve possuir para pertencer a uma dada classe em particular, estas características são determinadas em função da graduação e escala, p.e. documentos que pertencem uma determinada característica terão uma grande probabilidade de pertencerem a mesma classe de maneira que quanto maior for o número de características maior será convergência a um resultado que consiste num coeficiente associados a cada uma das classes já conhecidas. Este coeficiente representa o grau de confiança ou certeza de que um determinado documento pertence a uma certa classe associada ao coeficiente resultante.

3.10.3.1 Representação Vectorial

No modelo vectorial tenta-se apanhar a relação ou vínculo de cada documento D_i, de uma determinada colecção de n documentos com um conjunto de m características desta colecção, assim formalmente um documento pode ser considerado como um vector que expressa a relação do documento com cada uma dessas características [42].

$$D_i \rightarrow \vec{d}=(c_{1k}, c_{2K...}, C_{nm})$$

Na representação vectorial do documento podemos verificar que o vector identifica em que grau o documento D_i satisfaz cada uma das m características ou seja o vector C_{ik} é um valor numérico que expressa em que grau o documento D_i, possui a característica. As noções das características concretizam-se na ocorrência de determinadas palavras ou termos dos documentos, embora não haja impedimentos de que outros aspectos sejam considerados, o que fundamentalmente é utilizado sobretudo no reconhecimento de objectos em que as características tenham aspectos perceptíveis p.e. cores, forma etc. [43].

3.13.4 K-Nearest Neighbors (K-NN - vizinhos mais próximos)

K-NN é um classificador de aprendizagem baseado no teste de grau de similaridade, entre os documentos e k dados de treinamento, armazena uma certa quantidade de dados de treinamento de maneira que para classificar uma nova instância se faça a busca nos dados armazenados casos sejam similares então são atribui a mesma classe. O conjunto de dados de treinamento são mapeados num espaço multidimensional de características, o espaço de características é particionado em regiões baseando-se nas categorias do conjunto de dados de treinamento, um ponto importante no espaço de características é

associar a uma categoria particular a um determinado dado se é a categoria muito próximo aos dados de treinamento, tipicamente utiliza-se a distância euclidiana na computação da distância entre os vectores. O elemento chave neste método é a disponibilidade da medida de similaridade para identificar os vizinhos de um documento particular [62]. A fase de treinamento consiste somente no armazenamento de vectores de características e categorias do conjunto de treinamento. Na fase de classificação a distância do novo vector representando um documento de saída para todos os vectores armazenados são computados os k exemplos mais próximos são seleccionados, a categoria do documento é predito baseando-se no próximo ponto que foi associado ou atribuído a categoria particular. Dado pela seguinte fórmula $argmaxi \sum_{j=1}^{k} Sim(Dj|D)x\ (\delta(C(Dj),i)$. Esta fórmula representa a equação padrão

para o cálculo de similaridade entre o documento de texto, em função da vizinhança e atribuição de teste a classe que contém muitos vizinhos.

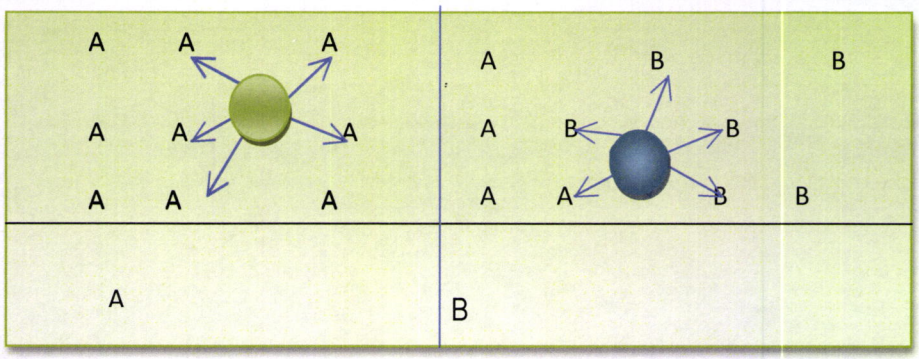

Figura 3.9: K- vizinhos mais próximos

A figura 3.9 apresenta os vizinhos mais próximos com k=5, onde podemos verificar os vizinhos associados a cada categoria das duas representadas. Deste forma cada nova classificação é feita com base na classificação da maioria dos seus vizinhos mais próximos. Para facilitar a atribuição de categoria de formas a evitar coincidências o K sempre será um número ímpar.

Este método é eficiente não é parametrizado e é fácil de implementar se comparado com o Rocchio's, muitas características locais do documento são considerados, contudo se o tempo de classificação for longo e dificultar encontrar o valor óptimo de k. Os dois algoritmos possuem algumas inconveniências que são identificados.

3.13.5 Árvore de Decisão

Reconstrói a categoria manual do documento de treinamento, construindo consultas falsos ou verdadeiras bem definidas na forma de estruturas de árvores, na estrutura de árvores as folhas representam a categoria correspondente ao documento e ramos correspondem a conjunção de características associados as características, a árvore de decisão bem organizada pode facilitar a classificação do documento colocado o nó na raiz do documento e deixa-lo correr através da estrutura de consulta que representa o objectivo de consulta do documento. A árvore de decisão é uma ferramenta de suporte a decisão que tem muitas vantagens, a principal é a simplicidade em que pode ser entendida e interpretada mesmo por aqueles utilizadores não especializados, além disso a explicação de um determinado resultado pode ser replicado utilizando simplesmente algoritmos matemáticos e prove uma visão consolidada da lógica de classificação que é uma informação útil para a tarefa, isto pode mostrar experimentalmente que a tarefa de classificação de texto evolve frequentemente várias características relevantes [79].

Com o método temos que ter em conta que pequenos detalhes podem comprometer a performance. Contudo quando temos um pequeno número de atributos na estrutura, a performance, a simplicidade e a facilidade de entendimento da árvore de decisão com base no modelo de conteúdo constituem aspectos vantajosos. O maior risco de implementação de árvores de decisão é se os dados de treinamento encaixarem-se com a ocorrência de uma árvore alternativa que categoriza por completo os dados de treinamento embora a categorização dos documentos será melhor organizado. Este algoritmo de classificação da árvore de decisão é adequado para a tarefa de classificação, contudo a limitação reside na negligência em torno da performance da classificação do documento, contudo estruturas enormes e complexos da árvore é feita através de um conjunto de dados com muitas entradas.

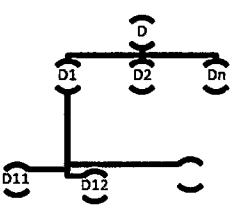

Figura 3.10: Árvore de Decisão

A figura 3.10 apresenta a Árvore de Decisão, cujo os nós representam os documentos e os ramos as características dos documementos.

3.13.6 Regras de Decisão

A regra de decisão é um método de classificação que utiliza as regras de inferência para fazer a classificação de documentos de formas a determinar a sua categoria [64]. O algoritmo constrói um conjunto de regras que descrevem o perfil para cada categoria, as regras são tipicamente construídos no formato de IF <condição> então conclusão, onde a parte de condição é preenchida pelas características da categoria e a parte de conclusão é representado com o nome da categoria ou outra regra para o teste o conjunto de regras para uma categoria particular é então construído pela combinação de cada regra, separando as regras das suas categorias com a lógica de operação, tipicamente utiliza-se os operadores *and ou or*. Durante a etapa de classificação não necessariamente cada regra no conjunto de regras necessita ser satisfeito, em caso de manipulação do conjunto de dados com número amplo de características para cada categoria a implementação de heurísticas é recomendado para reduzir o número de regras sem afectar a performance da classificação. A principal vantagem da implementação deste método de classificação é a tarefa de construção de um dicionário local para cada categoria individual durante a fase de extracção de características [64]. Os dicionários locais têm capacidade de distinguir o significado de uma palavra de diferentes categorias. Contudo a desvantagem do método é a impossibilidade de associar o documento a uma e exclusivamente uma categoria em função de regras de diferentes conjuntos.

3.13.7 Regras de Associação de Mineração

É uma tarefa canónica de mineração que tem como objectivo descobrir relacionamentos entre itens num conjunto de dados, são amplamente abordadas na literatura e sua eficiência na descoberta, fez com que tivesse ampla aceitação na comunidade de mineração de dados. Os melhores algoritmos são baseados em algoritmos a priori, que procuram pela frequência de ocorrência do conjunto de termos no documento, exemplo a regra do tipo $T \rightarrow C$, onde T representa o conjunto de termos do documento e C é a classe ou categoria associado. O classificador pode ser gerado para cada categoria ou para todas as categorias.

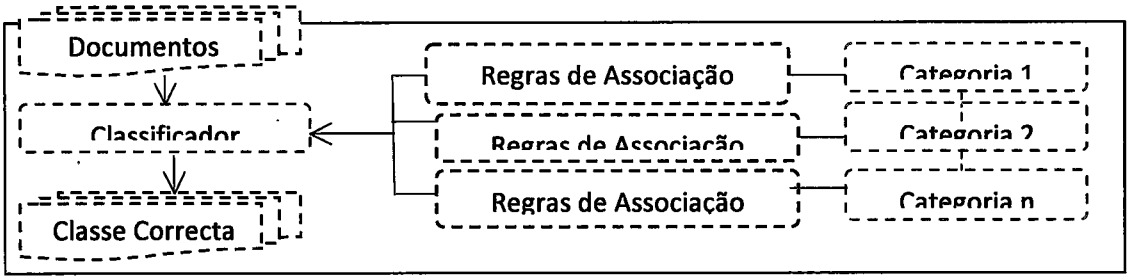

Figura 3.11: Classificador por Categorias

A figura 3.11 apresenta o classificador por categoria, cada categoria é considerado como uma colecção de textos de maneira isolada e aplicada a regras de associação afectos e neste caso simplifica o modelo de treinamento do documento para $Di = \{C, t1, t2 \ldots tn\}$ onde C é a categoria considerada e Ti o conjunto de termos associados a categoria. As regras geradas para todas as categorias separadamente são combinadas para se obter o classificador.

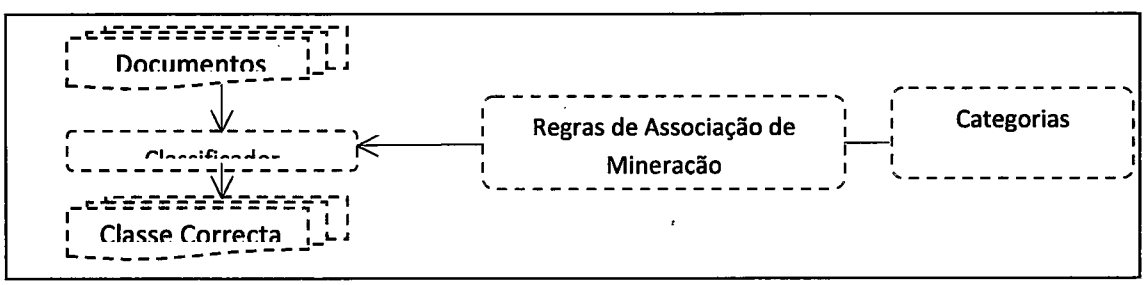

Figura 3.12: Classificador para todas as Categorias

A figura 3.12 apresenta a segunda abordagem de classificação com regras de associação, na qual todas as categorias formam uma simples colecção de dados de treinamento e as regras geradas são de facto o classificador.

3.13.8 Algoritmo de Rocchio

É um algoritmo de aprendizagem indutivo tradicional na recuperação de informação, desenvolvido originalmente para permitir um feedback em consultas de bases de texto. O algoritmo utiliza como método o espaço vectorial para roteamento ou filtrar a informação recuperada, constrói um vector protótipo para cada classe de documento utilizando um conjunto de documento de treinamento i.e. obtem-se a média sobre o vector de todos os vectores do documento de treinamento que pertence a classe C_i e calcular-se a similaridade entre o documento do teste e o vector protótipo que atribui o documento de teste a classe com a similaridade máxima. Quando dado a categoria que o documento pertence a esta categoria é dado um peso positivo e o vector do restante documento é atribuído o peso negativo. Ponderando o positivismo e o negativismo o

vector protótipo é obtido. É um algoritmo fácil de implementar e eficiente na computação e possui um mecanismo relevante de feedback, porém possui uma baixa precisão na classificação. Cujo cálculo é apresentado na fórmula seguinte $C_i = \alpha \times centroid_{ci} - \beta \times centroid_{\overline{ci}}$ (1). A combinação linear completa é também simples de classificar e as constantes α e β são empíricos, os algoritmos que utilizam operações no modelo de espaço vectorial fornecem um relevante feedback.

3.13.9 Redes Neurais

Corresponde a uma rede onde os termos representam as unidades de entrada e as saídas como as categorias associados aos pesos nas conexões que correspondem ao grau de relacionamento de interdependências, como exemplos o perceptrão [47].

3.14 Conclusões do Capítulo

Minerar dados ou textos está a tornar-se imprescindíveis para qualquer organização ou entidade que pretende fazer avaliação de algo, e independentemente da dimensão ou ramo de negócios, ajuda de maneira óbvia desde a gestão, controlo da informação, fornece à organização um mecanismo de maneira mais eficiente e eficácia além de facilitar de forma clara o trabalho dos tomadores de decisão e especialista em geral.

Vejamos que em grande medida os classificadores constroem os modelos de referência para conduzir os testes, cujos resultados gerados permitirão a análise da precisão dos com a classificação ou agrupamento. Os algoritmos a serem aplicados devem obdecer aos propósitos do contexto de referência, dos quais em alguns casos é preciso a combinação destes de forma comparativa para ressaltar a eficácia. Porque cada algoritmos possue vantagens e desvantagens a serem ponderados na hora da sua aptação.

Capítulo 4

Análise de Sentimentos

Este capítulo faz uma abordagem geral do processo de análise de sentimentos, descrevendo a sua importância, as fases e tarefas para a realização da avaliação, assim como, apresenta os principais problemas de pesquisa relacionados à análise de sentimentos e algumas das técnicas usadas para resolvê-los, apresenta por fim algumas os principias desafios e áreas de aplicação na actualidade.

4.1 Introdução

Grande parte de documentos e textos disponíveis possuem anomalias associadas a diversos factores os quais geralmente dificultam o processo AS, p.e. documentos tem quase sempre palavras irrelevantes ou redundantes para a actividade que se pretende realizar. Inconsistências podem ser decorrentes de erros de inserções, falta de preparo técnico, utilização de sarcamos, gírias etc. atentendo a forças de expressões e a dinâmica que caracteriza a interacção ou a produção dos conteúdos, por serem produzidas em contextos dinâmicos e por grupos heterogéneos ou de afinidades por exemplos os que são produzidos em fóruns de discussões numa comunicação assíncrono ou síncrono nas actuais médias sociais para expor opiniões ou sentimentos em tempo hábil em relação a

um determinado assunto, apoiados por meios de produção e uma dinâmica de interacção muito exigente em termos da eficiência com que devem expor ou difundir os seus sentimentos para que sejam avaliados e repercutidas a priori. Por outro lado e sob perspectivas de dados, existem diversas bases de dados textuais espalhados nas organizações ou entidades em suporte digital que tratam ou disponibilizam diversos corpus que constituem uma colectânea de dados ou opiniões semi-estruturados ou não estruturadas sobre os processos organizacionais ou entidades, objectos, acções ou medidas tomadas para sanar ou resolver algum problema que afecta directamente os clientes ou entidades, tais opiniões são auferidas por diversas entidades ou classes da sociedade que expressam os seus pontos de vistas e opiniões em relação a determinados assuntos.

Os dados quando não estruturados dificultam em grande medida a análise e a utilização para a finalidade a que se destinam, porque fundamentalmente são expressas e escritas sem uma linha padronizada e desta maneira são quase sempre ambíguas, dadas à natureza informal que os caracteriza, e em grande medida ilegíveis na sua plenitude aos beneficiários. A qualidade dos dados ou opiniões produzidos é de extrema importância para a análise e posterior suporte a tomada de decisões, verifica-se a falta de algum mecanismo de filtragem computacional, por exemplo, imaginemos o aumento de imposto sobre um determinado objecto, como filtrar em textos disponíveis, com determinada exactidão o posicionamento das pessoas sobre a contenda em relação a safisfactibilidade, insafisfactibilidade ou neutralidade. Haja vista, que especialistas e pessoas não especialistas muitas vezes adolescentes fazem o seu contributo quase sempre sem bases de apoio ou fundamentos técnicos e utilizam linguagens ou gírias que surgem a medida, para agilizar o processo de interacção. O que sem sombras a dúvidas dificulta seriamente o analista na tentativa de detectar a polaridade e as estatísticas em termos quantitativos e qualitativos por exemplo, mesmo com a utilização de técnicas e algoritmos concebidos para a referida tarefa t.c. como algoritmos de mineração de dados e texto.

A mineração de dados visa filtrar e limpar dados com vistas a converte-los em formatos apropriados aos algoritmos de *Data Mining* (mineração de dados) e desta maneira

extrair padrões para predizer situações ou ocorrências que sirvam para o direccionamento ou posicionamento efectivo do negócio ou da entidade, em tese para auxiliar os responsáveis, entidades singulares ou administradores na tomada de decisão. Desde os tempos remotos a tomadas de decisões requer ferramentas, técnicas e métodos de apoio ou suporte, actualmente o sucesso dos negócios ou ascensão de qualquer entidade está ligado à macro visão na qual se antecipam ou são preditas os desejos, necessidades dos clientes ou parceiros, o que constitui a chave para antecipar-se a concorrência, sobrevivência ou alcance do mais alto nível de excelência no mercado. Actualmente além da aquisição e disponibilização de dados, informações ou opiniões, é imprescindível a padronização e estruturação permitindo desta forma facilitar a sua análise com vista a proporcionar o conhecimento que convergem nas vantagens competitivas. Para a extracção e tratamento destes é necessário à utilização de métodos e técnicas que permitam à sua selecção a transformação e a sua mineração.

Especificamente a análise de sentimentos é análogo a ferramenta que auxilia o computador a apreender a avaliar ou entender bem os sentimentos expressos ou escritos em textos ou vários corpus[9] disponíveis geralmente na rede da Internet, de formas semelhante à feita pelos seres humanos. A mineração de opiniões ou detecção de subjectividade apoia-se nas linguagens naturais para o processamento de opiniões assim como as técnicas de aprendizagem automáticas para procurar estatísticas ou padrões nos textos que revelam atitudes. Actualmente a área está a ganhar notabilidade e importância extrema, porque é de aplicação imediata em vários ambientes de negócios, com a finalidade de sumarização de *feedback* para análise de produtos, e obter recomendações de parceiros ou auxiliar eleições em companhias ou corporações. A quantidade de informação tende a aumentar exponencialmente e a sua disponibilização nos médios tradicionais representa actualmente uma parcela muito pequena, em termos de tecnologias, a tendência é a mesma, por exemplo, corporações importantes como a Nokia que por muito tempo lideraram o mercado dos telemóveis, recentemente teve que demitir milhares de trabalhadores ao fechar filiais localizados na Asia, por ter perdido concorrência com corporações que lideram o mercado de *Ipads e Smartfones*, nomeadamente a *Sumsung e a Apple*. Logo é importante actualizar sempre os produtos, serviços e melhorar os métodos para a sua divulgação, baseando sempre no estudo do

[9] Corpus: colectânea de textos electrónicos compilado com base em critérios específicos e que servem para viabilizar a análise.

contexto e ambiente em função das exigências dos consumidores e adequar-se as novas formas de divulgação e relacionamento com os clientes, onde a Internet emerge como facilitador da interacção e comunicação, com o surgimento de fóruns de discussões e *blogs* passou-se a ter um espaço para intercâmbio e difusão de opiniões de maneira eloquente e eficiente e posteriormente com a necessidade cada vez mais premente desta partilha e interacção surgiram os serviços denominados por redes sociais como formas de interacção mais para a troca de informações sobre diversos assuntos, e consequentemente um espaço utilizado por milhões de pessoas diariamente, os quais fundamentalmente produzem um volume muito grande de informações diversas e de vários âmbitos e fóruns, importantes para os consumidores e interessados em geral. A exploração e tratamento desta informação constitui o segredo para a projecção de algumas entidades e individualidades, o que tem despertado bastante interesse actualmente as entidades, corporações e individualidades em geral, claramente na promoção de produtos, serviços, opiniões e diversos aprendizado sobre o que os outros pesam em relação aos seus serviços por exemplo, o que dispensou fundamentalmente a forma tradicional que normalmente a técnica empregue para o efeito eram as sondagens, que fundamentalmente despendiam muitos recursos sejam humanas, materiais e financeiras e muitas vezes sem levar em conta a volatilidade quase sempre nativa aos clientes se levarmos em conta o factor tempo. Tudo isso tem sido possível e prático devido ao grau de interactividade estabelecida pelas aplicações Web 2.0, que tem substancialmente mudado a essência ou a prioridade da Internet antes tido como uma fonte sobretudo de informação para passar a ser actualmente como uma fonte de opinião [8], a antevermos que qualquer que seja a dimensão de informação associado a produto, entidade objectos, lugares, individualidades etc. disponíveis *online* em lojas virtuais, sítios, medias sociais etc. podem ser comentados ou classificados de alguma maneira ou forma por milhares de pessoas as vezes simultâneamente e ser difundido em pouco tempo para milhares de entidades ou pessoas, o que de facto monstra o poder e a importância de análise de sentimentos. Especificamente para darmos exemplo, alguns estudos feitos no domínio de turismo (i.e. comentários ligados a restaurantes e hotéis) mostram que estas influência entre 73 a 87% das pessoas na hora de recorrer a serviços de um determinado estabelecimento de género [19]. De tal maneira podemos ver que a exploração e avaliação de opiniões beneficiam tanto a organização como os clientes.

4.2 Interação Dinâmica

Os *Microblogs* estão a tornar-se as ferramentas mais populares de comunicação entre os utilizadores da Internet, milhares de mensagens são disponibilizados diariamente em Web sites que forcem serviços de *microblogs* tais como *twitters, faceboocks, tumblr* etc. os autores das mensagens escrevem sobre diversas questões e assuntos por exemplo sobre as suas vida,trabalhos, projectos académicos, serviços, política, religião, cultura etc. combinando opiniões numa gama de tópicos e discussões relacionados a assuntos específicos, impulsionados pela formato livre das mensagens e a facilidade de acessibilidade das plataformas de *microblogs*, os utilizadores da Internet tendem a substituir as ferramentas de comunicação tradicional (p.e. blogs ou lista de mail) por serviços de *microblogs*, contam com milhares de utilizadores, estão a transformar-se em fontes de valor para sentimentos e opiniões. Os dados podem ser utilizados para marketing, estudos sociais etc.

A obtenção dos dados através destes serviços ou por outros moldes para o propósito de análise é uma tarefa que exige mecanismos de extracção os quais podem ser automáticas, semiautomáticas ou manuais, o seu tratamento fundamentalmente requer um trabalho árduo por exigir empregue de mecanismos, técnicas e ferramentas diversas, que formalmente filtram, classificam e qualificam a informação desejada para os propósitos a que se destina ou ao conhecimento desejado para a tomada de decisões.

4.2.1 Redes Sociais (SNS - Social Networking Site)

As redes sociais são espaços partilhados por pessoas e entidades quaisquer com a finalidade de intercâmbio de opiniões e relacionamentos (i.e. entre indivíduos, grupos, ou corporações) que em geral tem as mesmas afinidades utilizando-se para o efeito a Internet e consequentemente as plataformas que oferecem as facilidades para o efeito. A ideia básica por trás desses serviços é permitir relacionamento e interacção entre os participantes vulgarmente conhecidos por blogueiros que no geral tem afinidades culturais, familiares, amizades, colegas etc. a associação nestes espaços próprios, facilitada fundamentalmente porque eles disponibilizam sobretudo dados pessoais, por exemplo dados de identificação, do trabalho ou da entidade em que estudam ou estudou, cujo intuito por exemplo é rever os colegas e desta feita visar a reaproximação, partilhar ou trocarem ou ainda fortificar o grau de relacionamento.

2.4.2.1 SixDegrees

É o primeiro SNS, surgiu em 1997, e desde então evoluíram em várias plataformas com propósitos distintos, por exemplo prever a opinião dos blogueiros sobre factos ou produtos [45] ou identificação de comunidades de interesses (Java, [46]). Um exemplo típico das inúmeras vantagens que a sua adoptação encerra é sobretudo associado a sondagens de opiniões públicas, normalmente feitas com recurso a chamadas telefónicas e entrevistas os quais podem ser complementadas ou substituídas por uma minuciosa análise das mensagens colocadas pelos utilizadores nas redes sociais, que estão disponíveis sem custos, recorrendo às suas *APIs (Application Programming Interface).*

4.3 Tarefa de Análise de Sentimentos (AS)

Um sentimento está ligado ou relacionado plausivelmente com a recepção de impressões, sensações, emoções, conhecimento etc. por um determinado indivíduo ou entidade de maneira genérica que recebe estímulos. Por exemplo, a leitura de um determinado texto vinculado ao aumento do imposto de renda é um estímulo que pode despertar ou desencadear a diferentes leitores sentimentos diversos, dependendo da interpretação que possa ser feita por cada um deles em função dos interesses ou visão que ostenta da matéria. Por tanto a Análise de Sentimentos é uma área que explora as opiniões, emoções ou sentimentos de maneira global sobre um determinado corpus, sem a necessidade da utilização das técnicas tradicionais tais como sondagens ou entrevistas. Em tese explora textos que expressam opiniões (i.e. textos subjectivos) em contraste com textos objectivos que contém única e exclusivamente factos e por conseguinte não são viáveis para a análise. Na sua realização são empregues várias técnicas as vezes combinação para extracção e filtração de opiniões, assim poderia parecer desvantajoso, mas isto é feito quase a custo zero se comprado com as formas tradicionais de obtenção dos resultados e conhecimento, outro aspecto na análise é a determinação da polaridade, como forma de quantificar e qualificar as opiniões geralmente, esta quantificação da polaridade tem como principal contributo avaliar o grau de Positividade, Negatividade ou a Neutralidade das opiniões, emoções ou sentimentos expressos por uma amostra de indivíduos de um determinado universo em relação a um determinado assunto (e.g. ligado a política, produtos, serviços, pessoas etc.), para a sua realização é preciso a captação da orientação semântica dos termos que constituem o texto. A análise de

sentimentos ou mineração de opiniões é definido como uma tarefa que consiste em detectar e análise de polaridade no auxilio do processo de tomada de decisão que é afectada pelas opiniões formadas por líderes e pessoas comuns, p.e. quando uma pessoa quer comprar um produto *online* iniciará pela procura de comentários e opiniões escritas por outras pessoas sobre várias ofertas. A análise de sentimentos é uma das áreas de pesquisa mais promissoras em ciências de computação. Centenas de empresas estão a adoptar iniciativas de desenvolvimento de soluções de análise de sentimentos, também os principais pacotes estatísticos como SAS (Statistical Analysis System)[10] e SPSS (*Statistical Package for the Social Sciences*)[11] já incluem módulos de análise dedicados a AS . Os trechos de texto disponíveis em médias sociais são uma mina de ouro para as empresas e pessoas que deseja monitorizar a sua reputação e obter informações em tempo útil sobre os seus produtos e acções. Para analisar sentimentos ligados a uma determinada entidade devem ser considerados vários aspectos e factores, logo o sistema deve oferecer uma pontuação para a análise geral bem como analisar os sentimentos em relação a aspectos individuas da entidade. p.e. se tivéssemos que avaliar os sentimentos relacionados a um hotel, os aspectos individuais seriam, os sentimentos expressos em função da localização, ar condicionado, camas, acesso a Internet, qualidade serviços prestados e atendimento etc.

4.3.1 Objectivos da Análise de Sentimentos

Existem vários propósitos para a realização da análise de sentimentos disponíveis em um determinado corpus ou texto ou ligados a um determinado contexto, cujos metas consistem fundamentalmente em [19]:

1. Identificação de opinião: a partir de um corpus ou texto deve-se proceder a separação de factos das opiniões ou emoções;
2. Avaliação de polaridade: de posse de um determinado corpus ou corpora cujas opiniões já tenha sido separados dos factos e uma palavra-chave (e.g. que identifica se trata-se de uma pessoa pública ou uma corporação), classifique-as

[10] O SAS é um sistema integrado de aplicações para a análise de dados, que consiste de: Recuperação de dados, Gerenciamento de arquivos, Análise estatística, Acesso a Banco de Dados, Geração de gráficos, Geração de relatórios.

[11] Pacote este de apoio a tomada de decisão que inclui: aplicação analítica, Data Mining, Text Mining e estatística que transformam os dados em informações importantes que proporcionam reduzir custos e aumentar a lucratividade. Um dos usos importantes deste software é para realizar pesquisa de mercado.

como positivas ou negativas, ou indique o grau de negatividade ou positividade de cada uma delas;

3. Classificação de pontos de vista ou perspectivas: dado um conjunto de documentos contendo perspectivas ou pontos de vista sobre um mesmo tema ou conjunto de temas, classifique-os de acordo com essas perspectivas ou pontos de vista;
4. Reconhecimento de humor: dado um conjunto de textos com carácter emotivo ou sentimental, como *posts* de *blogs* pessoais, identifique que tipos de humor permeiam os textos ou classifique-os de acordo com as diferentes emoções encontradas.

A análise de documentos feita estritamente por humanos, a fim de compreender posicionamentos e opiniões, é capaz de atingir resultados superiores. Entretanto, ela nem sempre é possível especialmente quando há muitos documentos a serem analisados. A análise de sentimento pode responder diversas perguntas sobre uma gama de assuntos de interesse e relacionadas a diversas áreas, por exemplo numa organização produtora de bens ou serviços as seguintes perguntas poderiam ser satisfeitas:

1. O que os clientes pensam a respeito dos seus produtos em termos de serviços e a respeito da companhia em si;
2. Quão positivo ou negativo acham os clientes em relação aos seus produtos;
3. Que tipos de pessoas preferem ou gostam dos seus produtos.

Na área ou contexto político pode ser interessante saber ou conhecer se as pessoas em termos quantitativos, suportam ou não o programa de um determinado político ou força política ou se estão de acordo ou não com a linha de actuação de um determinado governo em relação aos aspectos sociais entre outros. Organizações sociais podem questionar opiniões de pessoas através de debates correntes que podem ser realizados através de plataformas *microblogs* com a realização de análise subjectiva ou objectiva, desta forma o peso da decisão de sobre o tipo de análise a realizar incide fundamentalmente sobre o contexto da sua aplicação. É comum classificar frases em duas classes principais no que diz respeito à subjectividade: frases objectivas que contêm informação factual e frases que contenham opiniões subjectivas explícitas, crenças e pontos de vista sobre as entidades específicas.

4.4 Análise de Objectividade

A objectividade está ligada é uma qualidade daquilo que é objectivo, externo a consciência, resultado de observação imparcial e independente de preferências individuais [12].

A análise objectiva pode ser aplicada a qualquer universo e para a sua realização utilizam-se algoritmos específicos que definem os testes a serem realizadas de forma precisa. O conhecimento objectivo é importantíssimo porque a ciência depende sobretudo dos experiências e análise do género, quando por exemplo outros pesquisadores validem a hipóteses.

4.5 Análise de Subjectividade

A subjectividade é um conceito ligado formalmente a forma de reagir a eventos ou ocorrências. É propriedade dos argumentos baseando-se no ponto de vista do sujeito e é influênciado pelos seus interesses particulares [12].

A análise subjectiva não pode ser replicada pois ela faz sentido apenas para um indivíduo ou uma amostra de um dado universo, envolve para tal que se conheça aspectos da entidade ou objectos sujeitos ao tratamento p.e. conhecimentos, comportamento e outras propriedades essenciais. O conhecimento subjectivo pode se tornar objectivo se for comprovado cientificamente, em outras palavras quando o modelo em análise envolver um determinado universo ou contexto.

No contexto geral os textos que expressão opiniões de certos autores são exemplos claros de subjectividade, já aqueles que expressam ou limitam os dados concretos e factos constituem a objectividade. A subjectividade e objectividade podem ser especificadas através de duas abordagens estatísticas nomeadamente: a abordagem Bayesiana e a Estatística Clássica.

- A Bayesiana: expande a inferência estatística objectiva e subjectiva onde a análise estatística deve ser encarada sob a perspectiva condicional (i.e. os dados observáveis devem ser conhecidos), considera-se que as quantidades

desconhecidas são variáveis aleatórias, portanto os parâmetros que interessam seguem uma distribuição a priori, que não depende de quantidades desconhecidas, entretanto a medida de probabilidade conjunta que governa as variáveis observáveis e não observáveis é conhecida;

- Na estatística clássica, as quantidades de interesse podem ser aleatórias ou fixas (índices de probabilidades), a medida que governa os dados observáveis e não observáveis não é conhecida. Tal medida de probabilidade pertence a uma família de probabilidades que será estimada utilizando os dados observados;
- Inferência estatística refere-se à obtenção de conclusões sobre quantidades não observadas θ a partir de dados observados y.

A diferença entre as duas abordagens consiste fundamentalmente no seguinte a inferência bayesiana aborda o problema alvo através especificação da probabilidade com subjectividade como função de uma medida plausível de uma proposição condicional na perspectiva da visão e conhecimento do observador. Logo a incerteza em relação à θ pode assumïr diferentes níveis ou graus, que geralmente se representam através de modelos probabilísticos para θ. Entretanto, tanto as quantidades observáveis, quanto os parâmetros do modelo estatístico são considerados quantidades aleatórias. Em contraste com a abordagem clássica onde os parâmetros do modelo estátistico são considerados quantidades fixas e desconhecidas [44]. Em tese a análise do género numa primeira instância visa determinar se uma frase é subjectiva ou objectiva, em contraste com a análise de sentimento que consiste em avaliar a polaridade de uma frase subjectiva, para responder se a frase expressa um sentimento positivo ou negativo. Para tal para a avaliação das frases utilizam-se técnicas de PLN, sobretudo a abordagem de aprendizagem automática, em que os documentos são objectos do um processo de classificação.

Um sistema de análise de sentimentos pode ser subdividido em cinco etapas fundamentais, em conformidade com esquema apresentada na figura 4.1.

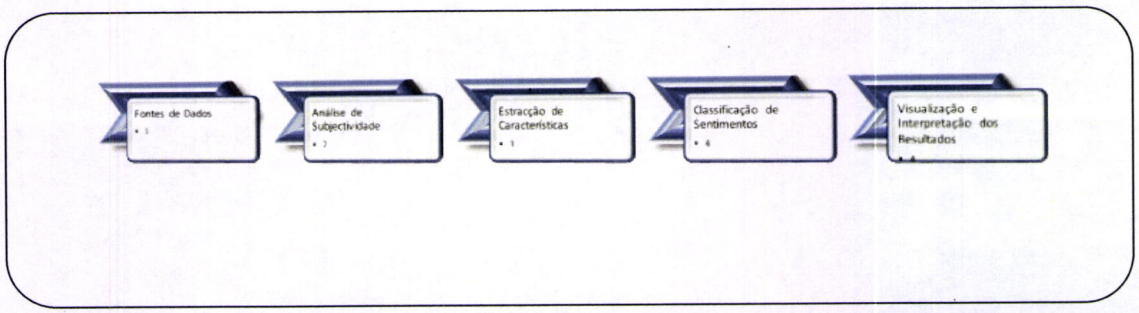

Figura 4.1: Sistema de Análise de Sentimento

A figura 4.1 mostra as fases que são executadas de forma sequencial para a realização da actividade de análise de sentimentos. A primeira fase consiste na recuperação de dados através das técnicas de RI, a fase de detecção de sujectividade consiste em encontra as frases, palavras ou padrões relevantes que podem ser utilizados para extrair a subjectividade, haja vista, que os textos extraídos podem ser subjectivos ou objectivos, a fase de extracção de características visa extrair as características relevantes que serão fundamentais e constituem opiniões e que serão obviamente utilizados para a realização do processo da análise e geralmente, a etapa da classificação de sentimentos extrai a polaridade das frases ou palavras, entretanto esta tarefa requer a utilização de técnicas de AM, onde o texto passa pelo processo de etiquetagem ou marcação. Os métodos utilizados na fase anterior são combinados na fase de análise de sentimentos. Em tese utilizam-se métodos de PLN, estatística ou aprendizagem automática para detecção ou extracção, identificar ou caracterizar conteúdo sentimental de um único texto. A última fase consiste na apresentação dos resultados da análise através de relatórios cuja apresentação pode ser na forma de tabelas, gráficos ou resumos em linguagem natural.

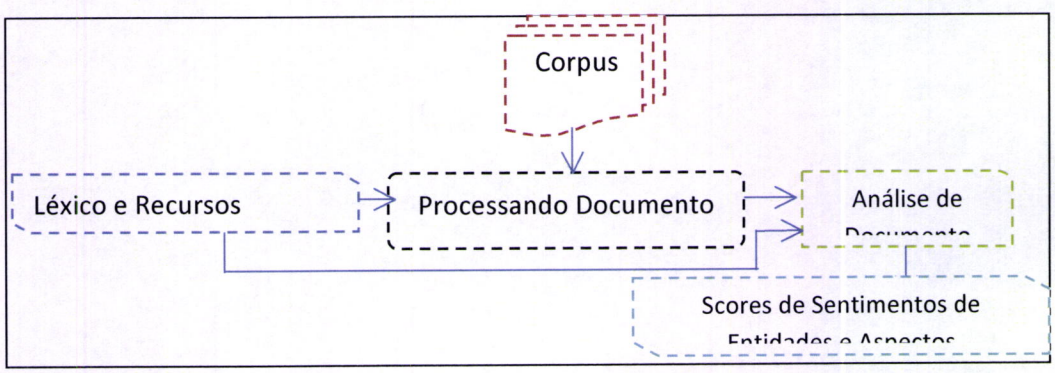

Figura 4.2: Arquitectura Geral de Sistema de Análise de Sentimentos

A figura 4.2 apresenta a estrutura geral de um sistema de análise de sentimentos, na qual se pode verificar a submissão do corpus ao processamento para a realização da tarefa de classificação e gera as pontuações de classificação, que serão úteis para a análise que se pretende.

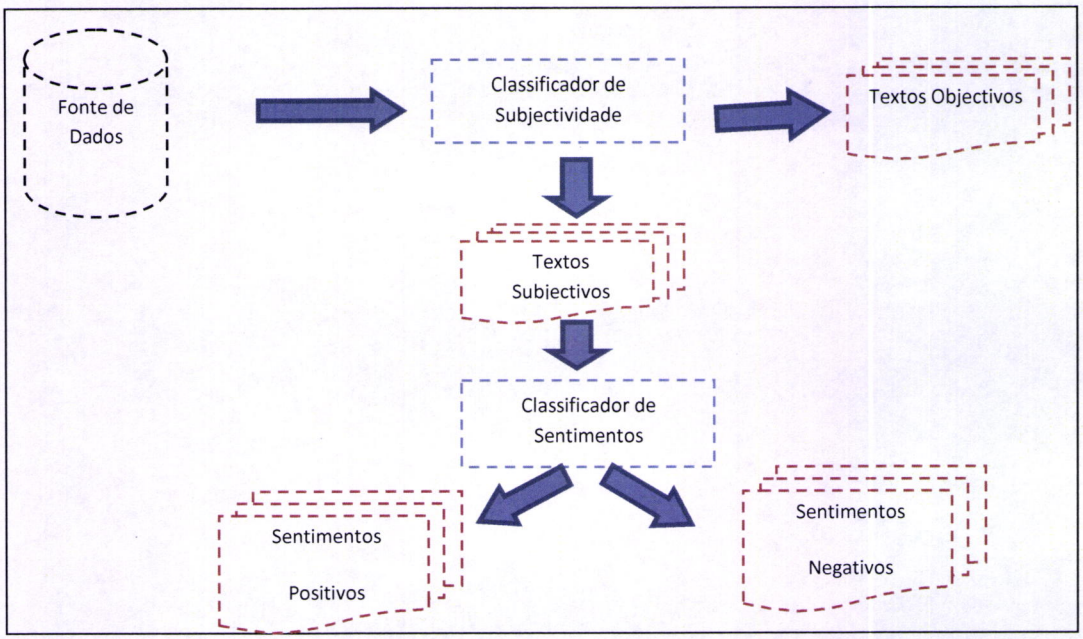

Figura 4.3: Processo de análise de sentimento

A figura 4.3 representa o processo de análise de sentimento, na qual textos submetidos a classificação passam inicialmente por um classificador para detecção e identificação de subjectividade, passo na qual apenas aqueles classificados como subjectivos serão utilizados para a realização da classificação de sentimentos. Após a realização deste processo como resultado final os documentos são agrupados em duas categorias fundamentais, aqueles cuja sua polaridade resulta em positiva para a classe positiva caso contrário para a classe negativa. Porém o processo da análise é complexo e pode ser feito de forma detalhada ou específica em conformidade com os objectivos preconizados e podem ser utilizadas técnicas descritivas ou preditivas, p.e. na preditiva a classificação é a tarefa de escolher as classes corretas para uma determinada entrada e na descritiva as classes não são conhecidas a priori. Nas tarefas básicas de classificação, cada entrada é considerada separadamente das outras entradas, e o conjunto de etiquetas é previamente definida. Alguns exemplos de tarefas de classificação são: Decidir se um email é spam ou não, neste caso as Classes são Spam e Não Spam; Decidir se uma determinada ocorrência da palavra Banco é usado para se referir a uma Margem do Rio, uma Instituição Financeira, ou acto Banco para sentar, ou o acto de depositar algo numa

Instituição Financeira; Decidir qual é o tema de um artigo de jornal é, de uma lista fixa de áreas temáticas, como "Esportes", "Tecnologia" e "Política". As classes seriam Esportes, Tecnologia e Política. Portanto a base da tarefa de classificação é o número de variáveis de interesse. Por exemplo, na classificação multi-classe, cada instância pode ser atribuído várias etiquetas, na classificação aberta, o conjunto de etiquetas não é definida previamente, e na sequência da classificação, uma lista de entradas são classificadas em conjunto.

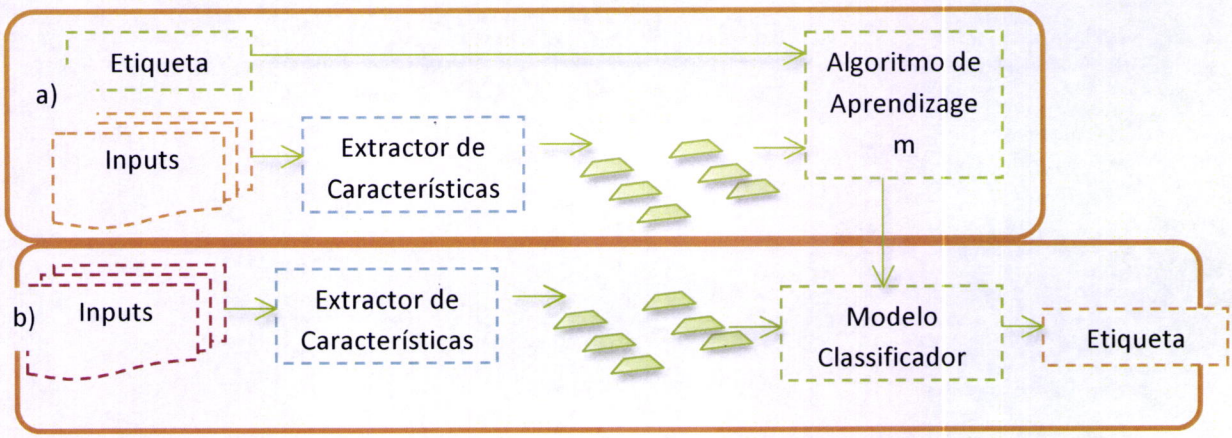

Figura 4.4: Classificação supervisionado

Conforme a figura 4.4 Durante o processo de treinamento as características extraídas são utilizadas para converter cada valor de entrada para um conjunto de características. Esse conjunto de características captura a informação básica das entradas que devem ser utilizadas para a classificação. Na sequência as características extraídas e etiquetas são inseridas no algoritmo para gerar o modelo. Durante a predição algumas características extraídas são utilizadas para converter entradas ocultas para conjunto de características, este conjunto de características são então inseridos no modelo que gera as etiquetas de predição.

4.6 Níveis de Análise

A análise de sentimento pode ser realizada de acordo os seguintes níveis: Relacionada ao documento; Relacionada à frase; Com base em Aspectos; Para Análise Comparativa e Para aquisição do léxico sentimental.

4.6.1 AS a Nível de Documento

Esta é a forma mais simples de realização de análise de sentimentos e presume-se que o documento contém opiniões sobre um objecto principal. Existem duas abordagens principais para a análise a este nível: aprendizado supervisionado e aprendizado não-supervisionado.

A abordagem supervisionada assume que há um conjunto finito de classes em que o documento deverá ser classificados e dados de treinamento estão disponíveis para cada classe. O caso mais simples é quando há duas classes: positivo e negativo. Extensões simples também podem adicionar uma classe neutra ou há alguma escala discreta numérica no qual o documento deve ser colocado (p.e. sistema de cinco estrelas utilizado pela Amazon). Considerando os dados de treinamento, o sistema treina um modelo de classificação usando um dos algoritmos de classificação comuns, tais como SVM, Naive Bayes, regressão logística, ou KNN. Esta classificação então é utilizada para marcar os novos documentos em suas classes. Quando um valor numérico em algum intervalo finito deve ser atribuído ao documento a regressão então pode ser usada para prever o valor a ser atribuído ao documento. Várias pesquisas têm demonstrado que uma boa precisão é alcançada mesmo quando cada documento é representado como uma simples bolsa de palavras. Em tese observa-se o sentimento expresso pelo documento de forma global.

4.5.1.1 Modelo Bolsa de Palavras

O modelo de bolsa-de-palavras é uma representação simplificada utilizada no processamento de LN e RI. Neste modelo, um texto (como uma frase ou um documento) é representado como uma colecção não-ordenada de palavras, gramática, ignorando até mesmo a ordem das palavras [13], é comumente utilizado em métodos de classificação de documentos, onde os termos e as frequências de cada palavra são utilizadas para gerar uma matriz. Em tese não é mais do que a representação numérica o documento, como se pode observar no exemplo 3.1. Para gerar um BoW é preciso realizar as seguintes tarefas: leitura e conversão (tokenização), extracção e limpeza dos termos (remoção de spowords, verificação de sinónimos e radicalização), contagem dos termos (quantidades de vezes que cada termo ocorre, criação de lista de termos e

frequência) e cálculo de frequência (combinar a frequência de termos e a frequência inversa do documento).

Exemplo de implementação de modelo BoW

Katia gosta de assistir filmes. Pedro gosta também.

Katia também gosta de assistir jogos de futebol.

Com base desses textos o dicionário pode ser construído

{"Katia": 1, "gosta": 2, "de": 3, "assistir": 4, "filmes": 5, "também": 6, "futebol": 7, "jogo": 8, "Pedro": 9, "também": 10}

Como tem 10 palavras distintas e utilizando a indexação no dicionário, cada documento representado por 10 entradas no vector:

[1, 1, 1, 1, 1, 0, 0, 0, 1, 1]

[1, 1, 1, 1, 0, 1, 1, 1, 0, 0]

Exemplo 4.1- Matriz

Donde cada entrada do vector referência um contador correspondente a entrada no dicionário. Cada linha representa um documento (Di) e a coluna um termo (Tj) e o Valor de DiTj é o valor de frequência do termo no documento. A representação vectorial não preserva a ordem das palavras na frase. A referência clara deste modelo no contexto linguístico pode ser encontrada em [14]. Representações mais avançadas utilizam TFIDF[12], POS léxicos sentimento, e as estruturas de análise.

Abordagens individuais de AS a nível de documento são baseadas na determinação da orientação semântica (SO) de frases específicas no documento. Se a média SO destas frases estiver acima de um limiar pré-definido o documento é classificado como positivo e caso contrário é considerado negativo. Existem duas abordagens principais para a selecção das frases: um conjunto de padrões predefinidos de padrões POS pode ser usado para seleccionar frases ou o léxico de palavras sentimentais e frases que podem ser utilizadas.

[12] É uma medida estática que representa a relevância de uma palavra de um documento em relação a colecção de documentos. A frequência de um termo (TF) representa é a quantidade de vezes que o termo aparece no documento.

O método clássico para determinar o SO de uma dada palavra ou frase é calcular a diferença entre o PMI (*Pointwise Mutual Information*) da frase com duas palavras que expressam sentimentos. PMI (P, W) mede a dependência estatística entre o P e o W frase e palavra respectivamente com base na sua co-ocorrência. Em [15] são utilizadas como exemplo as palavras 'excelente' e 'pobre'. As medidas, se P está mais perto de significado para a palavra ("excelente") implica positiva ou negativa para a palavra ("péssima").

4.6.2 AS a Nível de Frases

Um único documento pode conter múltiplas opiniões ainda que sejam sobre a mesma, entidade. Quando queremos ter uma visão mais detalhada de diferentes opiniões expressas no documento sobre as entidades. Assumimos que conhecemos a identidade da entidade associada, outrossim, de que existe uma única opinião em cada frase. Esta hipótese pode ser construída através de divisão da frase em frases na qual cada frase contém apenas uma única opinião. Antes de analisar a polaridade das frases que deve determinar se as frases são subjectivas ou objectivas. Apenas frases subjectivas serão posteriormente analisadas. Algumas abordagens também analisam frases objectivas, porém são mais difíceis. A maioria dos métodos utilizam abordagens supervisionadas para classificar as frases nas classes. Uma abordagem *bootstrap* foi sugerida em [16], a fim de reduzir a quantidade de trabalho manual necessário quando se prepara um grande corpus de treinamento. Uma abordagem única baseado nos cortes mínimos foi proposta em [17]. A premissa principal de sua abordagem é que as sentenças vizinhas devem ter a mesma classificação de subjectividade. Depois de termos mapeadas as frase subjectivas apenas podemos classificar essas frases em classes positivas ou negativas. Como mencionado anteriormente, a maioria das abordagens para análise a este nível são baseadas em aprendizagem supervisionada ou não supervisionada [97]. A última abordagem é de natureza semelhante em [36].

Pesquisas recentes mostraram que é aconselhável para tratar diferentes tipos de frases por diferentes estratégias. Sentenças que necessitam de estratégias únicas incluem frases condicionais, frases e frases sarcásticas. Sarcasmo é extremamente difícil de detectar, e aparece fundamentalmente nos contextos políticos. Uma solução para a identificação de

frases sarcásticas é descrito em [35]. Observa-se em suma os sentimentos expostos nas frases em termos da sua polaridade.

4.6.3 AS Baseadas em Aspectos

É o problema de pesquisa que se concentra no reconhecimento de todas as expressões de sentimento dentro de um determinado documento e os aspectos a que se referem. As duas abordagens anteriores funcionam bem quando se considera um documento na totalidade ou em cada frase individual refere-se a uma única entidade. No entanto, em muitos casos, as pessoas falam de entidades que têm muitos aspectos (atributos) e eles têm uma opinião diferente sobre cada um dos aspectos. Isto acontece muitas vezes em opiniões sobre produtos ou em fóruns de discussão dedicados a categorias específicas de produtos (p.e. carros, câmaras, *smartphones* e até mesmo medicamentos farmacêuticos). Assim alguns aspectos serão analisados de forma positiva, enquanto outros avaliados como sentimentos negativos no mesmo texto.

A abordagem clássica, que é usada por muitas empresas comerciais, para a identificação de todos os aspectos de um corpus de produtos, consiste em extrair todos os sintagmas nominais (NP) e manter apenas as PN, cuja frequência está acima de algum valor determinado experimentalmente [12]. A abordagem visa reduzir o ruído no NPs. A ideia principal é medir cada NP candidato do PMI com frases que estão fortemente relacionados com a categoria do produto (como telefones, impressoras ou câmaras). Apenas aqueles que têm um NPs e PMI acima de um limiar aprendidas são retidas. Por exemplo, para a categoria de impressora frases candidatas, por exemplo, seria "impressora vem com" ou "impressora tem." Assim nesta abordagem o problema de pesquisa se concentra no reconhecimento de todas as expressões de sentimento dentro de um determinado documento e os aspectos a que se referem. Outra abordagem para a identificação de aspectos é a utilização de um analisador de dependência de frases que utiliza expressões sentimentais conhecidos para encontrar aspectos adicionais (mesmo aqueles raros) [39]. Podemos ver o problema da identificação de aspecto como um problema de extracção de informações e, em seguida, usar um corpus marcado para treinar um classificador sequencial como um campo condicional Randómico (CRF) [18] para encontrar os aspectos [14]. O exposto até agora discute a identificação de aspectos explícitos, i.e, aspectos que são mencionados explicitamente nas frases. No entanto, há

muitos aspectos que não são mencionados explicitamente nas frases e podem ser inferidas a partir das expressões de sentimento que as mencionam implicitamente. Esses aspectos são chamados aspectos implícitos. Exemplos de tais aspectos são de peso, que pode ser inferida a partir do fragmento "este telefone é muito pesado", ou tamanho, que pode ser inferida a partir de "a câmara é bastante compacta." Uma maneira de extrair tais aspectos implícitos é sugerido em [10]. Onde uma das duas fases de co-ocorrência associação a abordagem a regras de mineração é usada para combinar aspectos implícitos (expressões de sentimento) com aspectos explícitos. Com estes dois conjuntos, podemos usar um algoritmo [2] simples que determina a polaridade de cada expressão de sentimento baseado num léxico sentimental, sentimento shifters (t.c. palavras de negação), e um tratamento especial de conjunções adversativas, como 'mas' por exemplo. A polaridade final de cada aspecto é determinada por uma média ponderada de todas as polaridades de expressões sentimentais inversamente ponderados pela distância entre a face e a expressão do sentimento.

4.6.4 AS Comparativa

Em muitos casos, os utilizadores não emitem uma opinião directa sobre um produto, mas sim fornecer opiniões comparáveis, p.e. " o modelo x de automóvel parece muito melhor do que o modelo y", "Eu testei o Tundra, ele não controla melhor do que o Land Cruizer, não chega mesmo nem perto." O objectivo do sistema de análise de sentimento, neste caso, é o de identificar as frases que contenham opiniões comparativas, e delas extrair a entidade preferida em cada opinião. Um dos trabalhos pioneiros sobre a análise comparativa de sentimento poder ser vista em [15]. Este trabalho mostra que a utilização de um número relativamente pequeno de palavras pode cobrir 98% de todas as opiniões comparativas. Estas palavras são:

- Advérbios e adjectivos de comparação tais como: "mais", "menos" (por exemplo, "mais leve");
- Adjectivos superlativos e advérbios, tais como: "mais", "pelo menos" (por exemplo, 'o melhor');
- Frases adicionais, tais como "favor", "exceder", "superperformance", "preferir", "de", "superior", "inferior", "número um", "contra".

Uma vez que esta palavra leva a uma abrangência muito alto, mas com baixa precisão, um classificador de Bayes pode ser usado para filtrar sentenças que não contêm opiniões comparativas. Os padrões sequenciais foram descobertos pela dominação de classe dos algoritmos sequênciais (RSE). Um simples algoritmo para identificar as entidades preferenciais com base no tipo de comparação utilizado e a presença de negação é descrito em [3].

4.6.5 AS para Aquisição Léxico

Como vimos na discussão anterior, o léxico sentimental é o recurso mais importante para a maioria dos algoritmos de análise de sentimento. Existem três abordagens para adquirir o léxico sentimental: abordagens manuais em que as pessoas codificam o léxico manualmente, baseada em dicionário em que um conjunto de palavras da frase é expandido por recursos utilizando como WordNet [8] e abordagem baseada em corpus no qual um conjunto de palavras da frase é expandido por meio de um grande corpus dos documentos a partir de um único domínio. Claramente, o trabalho manual em geral, não é viável porque cada domínio requer seu próprio léxico e um tal esforço é árduo, e honroso.

4.6.6 Abordagem baseada em dicionário

Começa com um pequeno conjunto de palavras da frase adequada para o domínio em questão. Este conjunto de palavras é, então, expandida usando sinónimos e antónimos WordNet. Um dos algoritmos elegantes é proposto em [16]. O método define a distância d (t1, t2), entre os termos t1 e t2 como o comprimento do caminho mais curto entre t1 e t2 na *WordNet*. A orientação de t é definida como OS (t) = (d (t, péssimo) - d (t, bom)) / d (bom, péssimo). | OS (t) | é a força do sentimento de t, SO (t)> 0 implica que t é positivo, e t é negativo caso contrário. A principal desvantagem de qualquer algoritmo baseado em dicionário é que o léxico adquirido é independente do domínio e, portanto, não capta as peculiaridades específicas de qualquer domínio. Abordagens baseadas em Dicionários mais avançados [4, 29]. O léxico sentimental é o recurso mais importante para a maioria dos algoritmos de análise de sentimento. Se queremos criar um léxico sentimental de domínio específico, temos que usar algoritmos baseados em corpus. Em

[11] o trabalho clássico neste domínio introduziu o conceito de consistência sentimental que permite identificar adjectivos adicionais que têm uma polaridade consistente como um conjunto de adjectivos encontrados. Um conjunto de conectores linguísticos (*AND, OR, NEITHER-NOR, EITHER-OR*) foram utilizados para encontrar adjectivos que estão ligados aos adjectivos com polaridade conhecida. Considere a frase "o telefone é poderoso e versátil." Se sabemos que 'poderoso' é uma palavra positiva, podemos supor que, utilizando o conector e a palavra "versátil " é positivo. A fim de eliminar o ruído do algoritmo cria-se um gráfico de adjectivos usando ligações induzidas pelo corpus e depois faz-se um *clustering*, de positivos e negativos.

4.6.7 Abordagem baseada em Corpus

A abordagem chamada de propagação duplo para aquisição simultânea de um léxico sentimental de um domínio específico e um conjunto de aspectos foi introduzido em [31]. Esta abordagem utiliza o analisador minipar [19] para analisar as sentenças no corpus e encontrar aspectos associados e expressões de sentimento. O algoritmo começa com um conjunto de expressões de sentimentos e usa um conjunto de regras de dependência pré-definidas e o analisador minipar para encontrar aspectos que estão ligados às expressões de sentimento. Ele então usa os aspectos achados pertinentes para encontrar as expressões mais sentimentais que por sua vez levam a encontrar mais aspectos. Este processo de *bootstrapping* mútuo pára quando não há mais aspectos ou expressões de sentimento que podem ser adicionadas. A migração um léxico sentimental de um domínio para outro domínio foi estudado em [5].

Entre as abordagens apresentadas a comumente utilizada é a nível de documento, pois por exemplo o grau de dificuldades em realizar o processo a nível de frase ou atributos é maior, porque se trabalha com detalhes relativamente ínfimos. Embora actualmente o processo é facilitada por ferramentas que categorizam a polaridade de forma automática p.e. o *SentiWordnet*. Que é uma lista com mais aproximadamente 110 mil termos ou mais em língua inglesa onde cada palavra é associada a três pontuações que distinguem a polaridade negativa, positiva e neutra.

4.7 Áreas de Aplicação

Análise de sentimentos oferece às organizações a capacidade de monitorizar as diferentes informações em sítios de médias sociais em tempo real e agir em conformidade. Gestores de marketing, empresas de relações públicas, os gestores de campanha, políticos e até mesmo investidores de capital e compradores on-line são os beneficiários directos da tecnologia de análise de sentimento.

- Aplicações inteligentes: para identificar as razões de preferência ou não dos consumidores de um determinado produto p.e. *laptop HP*, de formas a conhecer dados concretos e.g. preços, especificação, concorrência etc. poderia ser interessante conhecer os dados subjectivos t.c. o design é péssimo, os serviços ao consumidor são condizentes. Percepções erradas também são importantes p.e. drivers disponíveis não são actualizados ainda que estejam. De facto a resposta a esta subjectividade é árduo na análise convencional o AS facilita as respostas a estes questionamentos. Entretanto podemos utilizar a AS para: pesquisar opiniões na Web e concorrentes para analisa-los criam-se versões condizentes aos pontos de consenso;
- Política ou Ciências Políticas: para numerosas aplicações e possibilidades t.c., análise de tendências, identificação ideológica, direccionamento de publicidade ou mensagens, reacções aferidas etc; Avaliação de eleitores ou público; Análise e discussões políticas; Leis ou formulação política;
- Sociologia: para Propagações de ideias através de grupos de conceitos importantes de sociologia, as reacções e opiniões para as ideias são relevantes para adaptação de novas ideias; A análise de sentimento em *blogs* pode dar discernimento para este processo;
- Psicologia: para aumento potencial para investigação psicológica ou experiências com dados extraídos em "Linguagens Naturais" (LN).

4.8 Desafios de AS

Entre os maiores desafios propensos a realização da AS podemos destacar os seguintes:

- Pessoas expressam opiniões de maneira complexa;
- No texto a opinião pode ser incompreensível;
- Inversão intratextual, subfrases, negação, mudança de tema são comuns;
- Retóricas, modos como sarcasmo, ironia, implicação etc.;

- Alguns jargões de análise de sentimento são: orientação semântica e polaridade.

4.9 Conclusões do Capítulo

A análise de sentimento veio evolucionar sobre que maneira a forma pela qual podem ser tirados proveitos ao conteúdo guardado em formato não estruturado. Haja visto o volume que esses dados apresentam para qualquer organização, apoiado pelas técnicas de Recuperação Informação, Aprendizagem Automática e Processamento de Língua Natural esta tarefa está cada vez mais a facilitar a aquisição de conhecimento em tempo hábil com auxílio do processo automático o que permite de que maneira a obtenção dos resultados com alta precisão e fiáveis apesar da heterogeneidade e a ambiguidade que caracteriza os dados disponíveis para o efeito, os algoritmos proporcionam mecanismos eficientes que vão desde a recuperação, extracção de características relevantes que revestem-se na detecção da subjectividade no documento cujas propriedades nos permitiram realizar a análise em função da polaridade obtida nas frases sentimentais associados ao documento. Este processo é realizado de forma a obter resultados que permitam aos tomadores de decisão direccionar as suas acções de forma a alcançar a eficiência.

O nível de análise a ser explorado depende fundamentalemente dos propósitos e a profundidade da análise a que se pretende, por outro lado existem uma serie de desafios que tem que ser ultrapassados para a actividade de AS.

Capítulo 5

Protótipo de Análise de Sentimentos em textos

Este capítulo apresenta o protótipo elaborado objecto da presente tese, os testes e os resultados alcançados com a realização de experiências, com a utilização dos algoritmos de classificação. São também descritos neste capítulo todos os recursos e conjunto de dados utilizados, além das ferramentas e linguagens no contexto prático.

Para a realização da tarefa de análise como já abordado no capítulo anterior os documentos a serem analisados passam por um processo de transformação ou conversão cuja pretensão é criar uma representação vectorial ou bolsa de palavras, onde cada palavra é uma dimensão e indica uma determinada característica do texto. Através da utilização de modelos estatísticos podemos aprender as estruturas complexas da linguagem presente em um corpus do texto.

5.1 Introdução

Com a implementação do protótipo formalmente evidenciamos uma estrutura que alinha as tarefas que vão desde a captação de dados disponíveis num disco local, ou qualquer suporte conexo, desde que estejam no formato digital, através de utilização de ferramentas de recuperação de texto. A classificação é um processo como referenciado anteriormente, que é feita através de entradas constituídas por característica extraídas de ficheiros de textos ou corpus nas fases anteriores para a realização da tarefa e como

produto desta gera-se um classificador de documentos ou texto, de acordo com os objectivos a serem alcançado e com auxílio de algoritmos escolhidos com base nos propósitos evidenciados.

5.2 Protótipo

Para o processo de Análise de Sentimentos, desenvolvemos um protótipo apoiado na Linguagem Python e a biblioteca NLTK (*Natural Language Toolkit*), para a manipulação de texto e classificação de dados textuais.

Para a realização do processo de aquisição de conhecimento a abordagem contempla as principais fases da Arquitectura do processo de Análise de Sentimentos, apresentada na figura 4.1. As fases do protótipo de análise de sentimento são: Recuperação de Texto; Preparação do Texto; Gerador de Dados; Classificação do Texto e Análise dos Resultados.

- O módulo de Recuperação de Dados tem como principais funções recuperar documentos de diversas fontes através de utilização de mecanismos de buscas associados a motor de pesquisa com a técnica de orientação semântica das palavras pesquisadas e submete-los a fase seguinte onde serão pré- processados;
- O módulo Preparação de Texto procede com o pré-processamento do texto e a classificação de subjectividade em função de extracção de características;
- O módulo gerador de Dados faz a divisão do corpus em dois conjuntos de dados i.e. treinamento e testes com apoio de algoritmos de classificação;
- O Módulos classificador tem a função de gerar um classificador em função das entradas recebidas;
- O módulo visualização fornece os relatórios e sumários gerados na fase anterior, que servirão para testar o desempenho e a qualidade do modelo.

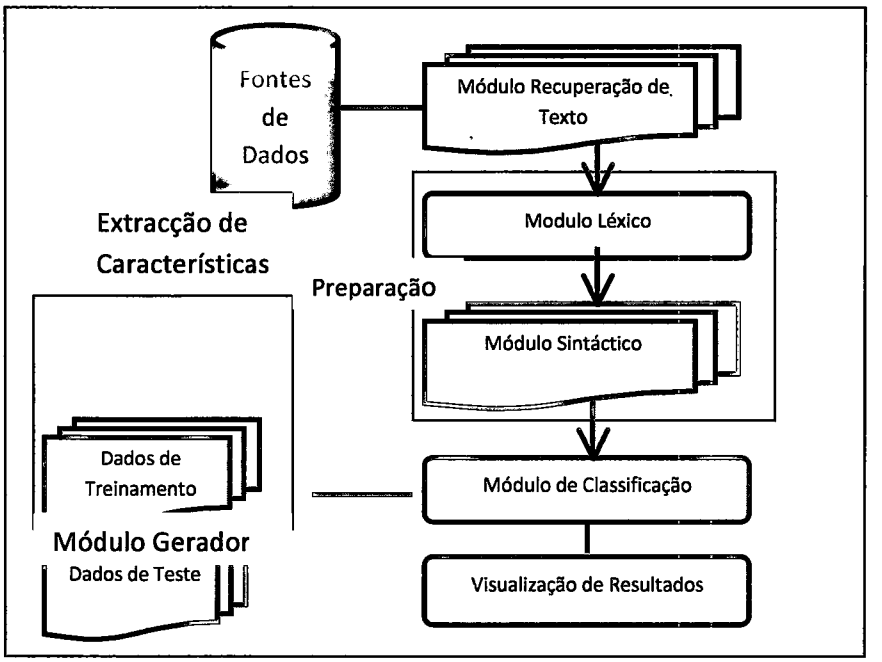

Figura 5.1: Estrutura do Protótipo

Conforme a figura 5.1 ficheiros de textos ou documentos recuperados na sequência são guardados numa base de dados textual ou submetidos a fase seguinte, na segunda fase do processo faz-se o tratamento do texto com vistas a construção do corpus e extracção de características a serem utilizadas para gerar o classificador, em tese esta fase desdobra-se no pré-processamento i.e. análise sintáctica e morfológica do texto. Construído o corpus a fase seguinte utiliza os algoritmos de classificação para extrair características de interesse e dividir o conjunto de dados em dados de treinamento e dados de teste, com os dados de treinamento gera-se o modelo de treinamento que servirá para realização dos testes na fase de classificação, os sumários, tabelas, gráficos e relatórios são gerados e seus resultados podem ser interpretados e analisados para a tomada de decisões.

5.2.1 Preparação e Extracção de texto

A preparação é uma fase inserida na etapa de pré- processamento conforme referenciado anteriormente e utiliza técnicas e ferramentas de extracção de caracacterísticas. Esta etapa visa gerar corpus que será processado com vista a proporcionar o conhecimento desejado. Assim com o processo podem ser reconhecidos entidades nomeadas e extrair as caracerísticas fundamentais para o processo. As entidades nomeadas podem

representar certas entidades, como nomes de empresas, localidades, acontecimentos etc. Outra actividade normalmente realizada nesta etapa é a sumarização.

Para o tratamento do texto, no que se refere a obtenção, extracção e a sua preparação utilizamos a Linguagem Python e a biblioteca NLTK. O Python é uma linguagem de programação multiparadigma, onde podem ser combinados conceitos de programação estruturada e orientada a objectos na implementação de construtores. A linguagem e as suas ferramentas auxiliaram-nos fundamentalmente na realização das seguintes tarefas:

- Extrair informações não estruturadas em qualquer média ou suporte digital;
- Analisar a estrutura linguística em termos sintácticos e semânticos;
- Utilização e criação de recursos léxicos para fazer a manipulação de diversas partes que constituintes do texto, p.e. unidades lexicais, palavras, frases etc;
- Remover *stopwords* ou filtrar unidades irrelevantes no texto;
- Encontrar a frequência de ocorrência de palavras no texto o que pode levar a descobrir o assunto ou contexto de interesse a abordagem;
- Especificação de padrões através de expressões regulares com vistas a mapear padrões no texto;
- Codificar soluções que sirvam para aceder ficheiros disponíveis nas médias ou em pastas do disco local;
- Construção ou geração de modelos de treinamento para a classificação de corpora.

A biblioteca inclui várias funcionalidades para o processamento de texto em Linguagem Natural seja no formato falado ou escrito, e portanto é vital para a realização da análise sintáctica e semântica.

5.2.2 Corpus

Corpus é um grande conjunto estruturado de textos utilizado sobretudo para a realização de análise sintática, assim para a verificação ou checagem de ocorrências e validação de regras gramaticais linguísticas considera-se um universo específico. Um corpus pode ter associado anotações que podem ser utilizados para pesquisa linguística, por exemplo, o uso de POS (*Part-of-Speech*), onde associa-se informação sobre a classe gramatical de cada palavra ou termo (i.e. verbo, substantivo, advérbio, pronomes etc.) é adicionada ao

Corpus. Outro exemplo seria a utilização de lemas que indicam a base de cada palavra ou anotações (gloss)[42]. Neste projecto as classes gramaticais são referenciados como categorias léxicas, em função disto a classificação de texto é feito através da utilização da técnica de etiquetagem ou marcação de texto.

5.2.3 Etiquetagem de texto

Consiste basicamente na transformação de lista de palavras (frase) numa lista de tupla, cujo formato é definido como tupla (palavra, etiqueta gramatical), a finalidade é identificar ou determinar e associar cada palavra a sua classe gramatical, p.e. substantivo, artigo, adjectivo etc. Os etiquetadores existentes são plenamente treináveis, o que significa que podem utilizar na sua maioria uma gama de frases já treinadas, em outras palavras modelos que servem de referência para etiquetagem ou testes. Neste processo cada palavra recebe uma tupla que inclui a palavra e etiqueta como forma de identifica-lo gramaticalmente p.e. a tupla (palavra, verbo) seria moldado da seguinte maneira tupla (palavra, V), onde a palavra representa qualquer termo ligado a um verbo e V representa a classe gramatical associada a palavra. Analogamente e de forma automática, este processo de mapeamento é feita para cada palavra contida no texto em questão no processo em referência.

```
import nltk
 nltk.corpus.mac_morpho.words() [u'Jersei', u'atinge', u'm\xe9dia', u'de', u'Cr$', ...]
 nltk.corpus.mac_morpho.tagged_words() [(u'Jersei', u'N'), (u'atinge', u'V'), ...]
nltk.SentiCorpus-PT.mac_morpho.words()
```

Figura 5.2:Trecho de código Exemplo de etiquetagem

O trecho de código apresentado na figura 5.2 representa a etiquetagem do texto é um processo que deve ocorrer subsequentemente a tokenização do texto, noutras palavras é a segunda actividade a ser realizada no processamento em PLN e tem uma capital importância porque através desta conseguimos obter a semântica das palavras ou significado do texto. Um dos etiquetadores de referência ou padrão muito utilizado para o efeito é o POS-TAGger (*part-of-speech tagger*). O NLTK Taggers é um pacote que contém classes e interfaces para etiquetagem. Uma etiqueta é um caso sensitivo ou *string* que especifica algumas propriedades do *token*, as etiquetas são codificadas como como tuplas (*tag,token*).

```
1. input = "POL=1 Paulo Portas POL=1 Sócrates refugiou-se no passado - como se
   ele próprio e o partido dele em 11 anos de 14"
2. tokens = nltk.word_tokenize(input)
3. tokens ['POL=1', 'Paulo', 'Portas', 'POL=1', 'S\xf3crates', 'refugiou-se', 'no',
   'passado', '-', 'como', 'se', 'ele', 'pr\xf3prio', 'e', 'o', 'partido', 'dele', 'em', '11', 'anos',
   'de', '14']
```

Figura 5.3: Trecho de código Tokenização da frase

A figura 5.3 apresenta um trecho de código que demonstra a tokenização do corpus, texto extraído do SentiCorpus-pt.

5.2.3 Classificação dos Corpus

Classificação é a tarefa de escolher a classe correcta para uma determinada entrada. Nas tarefas básicas de classificação, cada entrada é considerada separadamente das demais onde o conjunto de etiquetas é previamente definida, em termos gerais documentos podem ser classificados de diversas maneiras, p.e. por assunto, gênero, ou por sentimento contido neste. Envolve antes a Detecção de padrões que é uma parte central de PLN, p.e. as palavras que terminam em "ou"ou "vão" etc. tendem a ser verbos no passado. O uso frequente de palavra notícia é indicativo de texto de Jornal. Esses padrões observáveis ajudam a compreender a estrutura das palavras e as frequências de ocorrência de palavras de formas a correlacionar os aspectos particulares do significado, como tempo, tópicos etc. Mas como saber onde começar a procurar estes aspectos, quais são os aspectos que estão relacionados de forma a associa-los aos seus significados. As Técnicas de AM supervisionadas nomeadamente Naive Bayes, e Máxima Entropia foram adoptadas para a realização de experiências. Os fundamentos matemáticos e estatísticos a estas foram descritos na sessão 3.3, assim nos interessa por hora, apresentar a utilização ou aplicação prática.

A base da tarefa de classificação é o número de variáveis de interesse p.e. na classificação multi-classe, cada instância pode ser atribuído a várias etiquetas, na classificação aberta, o conjunto de etiquetas não é definida previamente, e na sequência da classificação, uma lista de entradas é classificada.

A estratégia adoptada para a análise concentra-se em classificar os sentimentos em duas classes principais, positivas e negativas. Os termos disponíveis no corpus são contados, na sequência faz-se o preprocessamento com as técnicas de PLN, com vista a obter a análise léxica e sintática do texto, com apoio de *stopwords* da linguagem e através da aplicação do modelo Bolsa de palavras, extrai-se as características mais importantes que serão utilizadas para a classificação dos termos. A análise do texto é considerada positivo se tem mas termos positivos do que negativos, e negativos caso o texto contém mas termos negativos do que positivos, e é considerada neutra se a análise dos termos resulta em igual número de positivos e negativos, esta última análise está fora do âmbito deste projecto. Isto é possível através da comparação dos termos, tags e rótulos contidos num dicionário da linguagem, neste caso o *WordNet.*

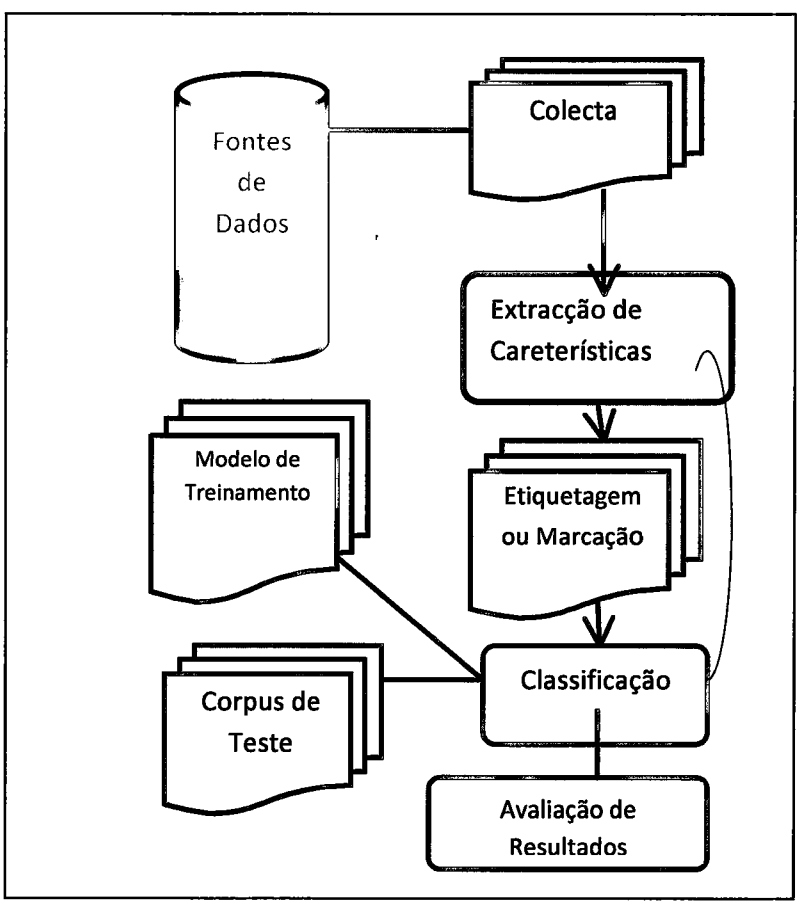

Figura 5.4: Principais actividades do Protótipo

Na figura 5.4 os dados são colhidos a partir da Web ou biblioteca NLTK, e podem ser etiquetados no passo seguinte caso não estejam, os não etiquetados passam pelo processo de etiquetagem, os já etiquetados são submetidos a fase de pré-processamento e classificação. No processo de classificação são subdivididos em duas partes:

Treinamento; e Testes. O corpus de treinamento gerará um modelo que servirá para testar os dados que posteriormente serão submetidos ao algoritmo. Durante o processo de treinamento as características extraídas do texto são utilizadas para converter cada valor de entrada para um conjunto de características. Esse conjunto de características tem como objectivo capturar informações básicas das entradas que devem ser utilizadas na classificação. Características e etiquetas são inseridas no algoritmo para gerar o modelo. Durante a predição algumas características extraídas são utilizadas para converter entradas ocultas de formas a obter um conjunto de características, este conjunto de características são então inseridos no modelo que gera as etiquetas de predição.

5.2.4 Descrição do Experiências

Dois corpura foram utilizados para a realização de experências nomeadamente *Movie-Rewiew e SentiCorpus_PT*. A metodologia adotada para a realização deste experências é a classificação supervisionada, onde o corpus é treinado com base em exemplos cujas classes são previamente conhecidas. Com base nas regras geradas e conhecimentos adquiridos pelos algoritmos, na sequência quaisquer exemplos submetidos serão treinados ou testados de acordo com o modelo criado, nas nossas experiências os corpos foram subdivididos de forma aleatória.

Os resultados da análise ou avaliação são eficientes quando os dados de treinamento e testes são diferentes, i.e. os dados que são utilizados para a geração do modelo ou treinar o algoritmo não podem ser os mesmos para fazer testes de precisão do modelo, em termos quantitativos, mas tem que referênciar o mesmo contexto ou serem da mesma natureza em contraste o mesmo algoritmo de treinamento pode ser utilizado para gerar classificadores de diferentes domínios. É interessante notar que o classificador gerado num determinado contexto é independente do algoritmo de apredizado, logo para análise de outro tipo de contexto o algoritmo deverá ser treinado novamente para adequar-se aos novos propósito ou cenário a que se propõe. Qualquer método de AM utilizado na realização de experências necessitam de corpus para a fazer a classificação no nosso caso e para capturar os elementos do estilo de subjectividade e a polaridade das frases são empregues as técnicas de PLN. São explorados corpus para a selecção das características numa primeira fase, utilizando a técnica de remoção de *stopword*, e numa segunda fase a bolsa de palavras e outras técnicas para a selecção das melhores

caracteracteristicas i.e. aqueles que contribuíram para aumentar os ganhos da informação e optimizaram a classificação e os resultados obtidos. Todos os classificadores do género trabalham com estrutura de características que pode ser um simples dicionário que mapeia o nome de atributos para os seus valores, no nosso caso utilizamos o modelo Bolsa de Palavras por ser o mais simples para levar a cabo a referida actividade. Quando se classifica modelos com milhares de características, p.e. categorização de documentos ou texto, é preciso observarmos que várias características podem ser irrelevantes ou insignificantes embora estas características sejam comuns a todas as classes, porém contribuem pouco para o processo de classificação, individualmente são inofensivos mas em conjunto podem diminuir extremamente a performance do modelo. A eliminação ou remoção de características irrelevantes torna o modelo mais claro, em suma, a alta dimensionalidade do espaço de termos pode representar um grande problema na análise, por isso a dimensionalidade deve ser reduzida o que é importante porque permite obviamente a redução de superajuste. Os algoritmos que fazem superajuste dos dados são bons na reclassificação dos dados utilizados, porém deficientes nos dados que ainda não foram utilizados. A performance é comprometida quando a quantidade de atributos é expressiva para criação de modelos, quando se utiliza apenas características mais relevantes pode-se aumentar a performance, e ao mesmo tempo diminuir o tamanho do modelo, o que resulta em menos memória e treinamento e classificação mais eficientes.

O ganho de informação é uma ferramenta utilizada em AM como recurso para a representatividade de termos, formalmente consiste em medir através de predição a categoria associada a um determinado termo em relação a outras classes. O termo que ocorre com maior frequência nos texto positivos, raramente ocorre nos negativos, logo é relevante para a informação, p.e. a ocorrência da palavra excelente no *review movie* é um indicador forte de positividade, então excelente é um termo relevante no texto, nota que as características mais relevantes não se alteram, isso faz sentido porque a tendência na análise é a utilização das palavras relevantes ignorando as demais.

Uma das melhores métricas para obter o ganho de informação é a utilização do módulo *Chi_Square*. Para efeito utilizamos o Bigrama, para calcular e especificar a frequência global ou seja a frequência de cada classe, a frequência global de termos é calculada utilizando a função *FreqDist* e especifica com a função *ConditionalFreqDist*, onde as condições são os rótulos das classes, com estes números podemos fazer *pontuações* de

palavras com a função Bigrama, em seguida classificar as palavras por *pontuações,* e depois colocar as palavras num conjunto e utilizar o teste associado ao conjunto de características a função de selecção, para seleccionar as palavras que aparecem no conjunto, logo cada ficheiro é classificado com base na presença destas palavras relevantes o que torna o classificador eficiente.

5.2.5. Estruturas de Dados

Uma estrutura estabelece a representação dos componentes de dados de um determinado corpus ou texto. A seguir discutiremos a representação dos corpus Movie Review e SentiCorpus_pt.

5.2.5.1 Movie-Rewiew

É o corpus padrão e mais utilizado para análise de sentimentos actualmente, possui 2000 ficheiros com opiniões sobre diversos filmes dos quais 1000 da classe positiva e 1000 da classe negativa, disponibilizados por [57]. Considera também que a análise de sentimentos constitui um caso particular de classificação, na qual sentimentos podem ser agrupados em apenas duas classes, na classe positiva aqueles cuja polaridade resulta em positiva e classe negativa onde as características extraídas do texto ou a análise leva a polaridade negativa.

Dividimos o corpus em dois conjuntos de dados inicialmente um de treinamento correspondendo a ¾ e obtemos como resultado 1500 instâncias de treinamento e 500 de testes, com o modelo gera-se uma lista de tokens na forma de [atributos, classes], o atributo é característica do dicionário e a classe é *Tag* da classificação. Atributos são da forma {palavra: verdadeiro}, e a classe será positiva ou negativa para avaliar a precisão com a utilização do classificador de desempenho que testa o conjunto.

5.2.5.2 SentiCorpus_PT

O corpus sentiCorpus_PT [86] é um ficheiro único que contém opiniões de diversos actores, expressa em função de análise de debate eleitoral feito por três políticos portugueses em termos quantitativos tem 887134 caracteres. Inclui a polaridade em classes Positivos e Negativos.

- Positivos: aquelas que contêm propriedades ou características que expressam opiniões com polaridade (pol= 1);
- Negativos: aquelas que contêm propriedades ou características que expressam opiniões com polaridade (pol= -1).

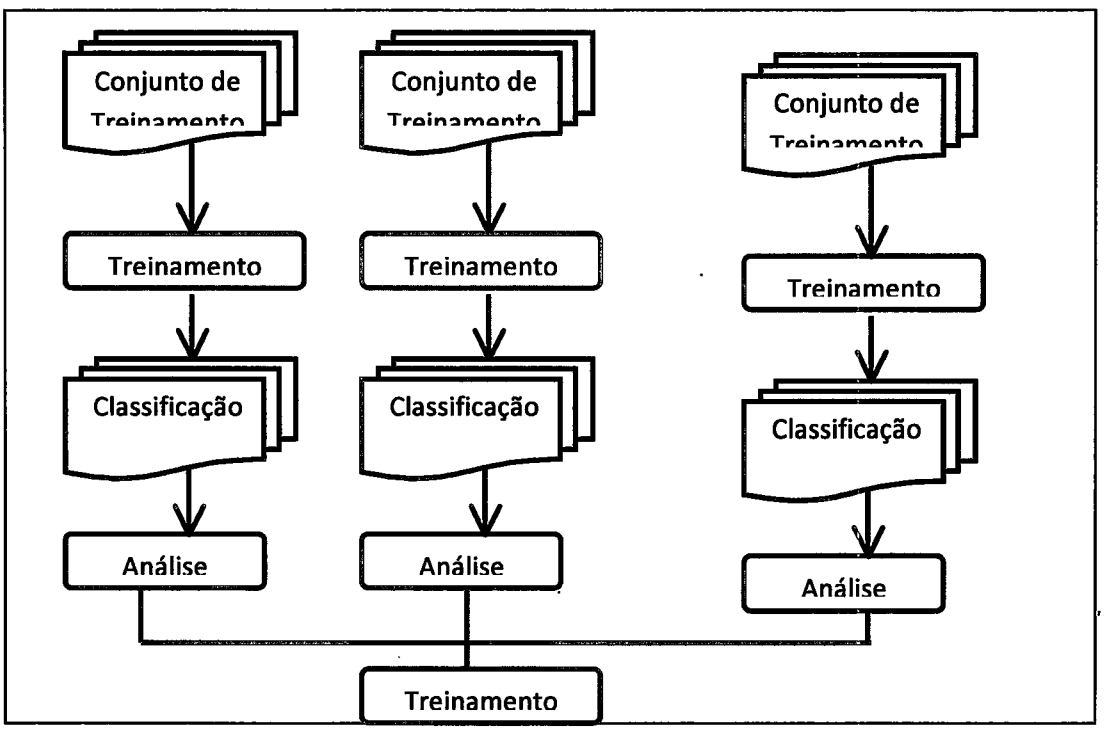

Figura 5.5: Treinamento e teste do corpus

A figura 5.5 apresenta o modelo de treinamento gerado com aplicação do algoritmo de classificação ou *clustering*, a qual na fase subquente é testado com os dados de testes, para obter a avaliação propriamente dita, geralmente faz-se a divisão dos dados em função de uma percentagem arbitrária ou randómica, uma parte servirá para treinar o algoritmo e gerar o classificador e a outra para teste, consequentemente em função da precisão obtida com a avaliação extraímos a conclusão e desta auferimos se o modelo é eficiente ou não. Conforme se pode verificar o modelo pode ser testado várias vezes com uma percentagem determinada dos dados da mesma natureza, os resultados de testes serão eficientes sempre que os dados utilizados para gerar ou treinar o algoritmo forem respeitantes ao mesmo contexto ou em suma representarem a mesma realidade, caso contrário os resultados serão péssimos.

5.3 Resultados e Modelo

Apresentamos a seguir os resultados e os classificadores gerados com aplicação de técnicas na tentativa de obtermos o melhor desempenho.

5.3.1 Modelo

A utilização do classificador Naive Bayes e Máxima Entropia permitiu obter um modelo com boa precisão, inicialmente com a aplicação do modelo Bolsa de Palavras para extrair características a serem utilizados para treinar o algoritmo e posteriormente com aplicação de outras técnicas.

5.3.1.1 Naive Bayes

Para um tipo diferente de problema de classificação, estamos interessados na classificação do texto. Neste problema, os itens a serem classificados são documentos. Os mais conhecidos são conjuntos de dados que identificam cada documento como uma categoria temática (daí o nome categorização) mas vamos olhar para os documentos do *NLTK Movie Review corpus*, onde cada documento é rotulado como 'pos' para positivo ou 'neg' para negativo de acordo com opinião ou análise. As características de cada documento serão as palavras contidas no documento de um conjunto de palavras que são frequentes em toda a colecção de documentos. Para cada item a ser classificado, neste caso uma única palavra na NLTK construirmos as características desse item como um dicionário que mapeia o nome de cada características para o seu valor, uma string neste caso. Um conjunto destas características constitui o dicionário com a etiqueta do item a ser classificado, neste caso o POS Tag (ver apêndice A). Em tese é construído um modelo de bolsa de palavras com valores verdadeiros como método para extrair as características. Inicialmente começamos pela identificação de sufixos das palavras que podem constituir características que denominamos 'Endswith (s)', onde s pode ser qualquer sufixo. O valor das características será Verdadeiro ou Falso dependendo se a palavra termina com o sufixo. Por exemplo para obter um conjunto de sufixos das características que irão seleccionar as 100 palavras mais frequentes no nosso corpus, utilizando sufixos de comprimentos 1, 2 ou 3.

```
Suffix_fdist nltk.FreqDist ()
For palavra em marrom Palavras. (): word = word.lower ()
      suffix_fdist.inc (palavra [-1:])
      suffix_fdist.inc (palavra [-2:])
     suffix_fdist.inc (palavra [-3:])

Common_suffixes = suffix_fdist.keys () [: 100]
Common_suffixes impressão
```

Figura 5.6: trecho de código obtenção 100 Palavras mais Frequentes

Os documentos de análise de filme não são identificados ou categorizados individualmente, mas sim são separados em ficheiros de directorias em função da sua categoria. Primeiro criamos a lista de documentos onde cada documento é associado a sua categoria. Internamente os documentos são ordenados por categorias e depois são mesclados e em seguida separados em conjuntos de treinamento e testes. Na sequência definimos o conjunto de palavras ou termos que serão usadas como cracterísticas, isto é essencial em todas as palavras da colecção de documentos, excepto se limitamos por exemplo para 2000 as palavras mais frequentes.

```
from nltk.corpus import  movie_reviews
import random
movie_reviews.categories()
documents = [(list(movie_reviews.words(fileid)), category)
for category in movie_reviews.categories()
for fileid in movie_reviews.fileids(category)]
random.shuffle(documents)
documents[0]
```

Figura 5.7: Trecho de código obtenção das Palavras mais Frequentes

O primeiro documento é composto de todas as palavras na análise, seguido pela sua etiqueta. Desde que as marcações sejam independentes cada pessoa deve ter um poder diferente da análise, seguidamente são definidas as características de cada documento. O rótulo da categoria terá as palavras-chave, para cada palavra no conjunto de características o valor das características será booleano se a palavras estiver contido no documento, depois é definido o conjunto de características para cada documento. Podemos olhar para o primeiro documento, mas lembre-se que o corpus contém 2000 Documentos. O passo seguinte treinar e testar o algoritmo, tarefa realizada com o classificador Naive Bayes.

```
train_set, test_set = featuresets[100:], featuresets[:100]
classifier = nltk.NaiveBayesClassifier.train(train_set)
print nlt k.classify.accuracy(classifier, test_set)
classifier.show_most_informative_features(20)
```

Figura 5.8: Trecho de código treinamento e teste do Nives Bayes

No trecho 5.8 pesquisamos 20 palavras-chave em vários documentos, serão positivas aquelas que contém a palavra e negativas aquelas que não contém, então a taxa de probabilidade será exibida como 20,00: 1,00 pos: neg.

```
train_set, test_set = featuresets[100:], featuresets[:100]
classifier = nltk.NaiveBayesClassifier.train(train_set)
print nltk.classify.accuracy(classifier, test_set)
```

A precisão resulta em 0.81 porcentos, o que nos leva a considerar que o modelo apresenta resultados que são razoáveis. A tabela 5.1 mostra algumas características classificadas de acordo com a proporção da classe, i.e. a quantidade de instâncias de uma classe para outra p.e. se tivermos 20 documentos que contém a palavra "outstanding" p.e. a relação será 20,00:1,00.

Atributo	Instâncias	Classe	Taxa
whatsoever	True	neg : pos	14.3 : 1.0
inept	True	neg : pos	13.7 : 1.0
stupidity	True	neg : pos	11.7 : 1.0
ludicrous	True	neg : pos	11.7 : 1.0
nomination	True	pos : neg	11.7 : 1.0
vulnerable	True	pos : neg	10.3 : 1.0
views	True	pos : neg	10.3 : 1.0
fictional	True	pos : neg	9.7 : 1.0
exceptional	True	pos : neg	9.7 : 1.0
abilities	True	neg : pos	9.7 : 1.0

Tabela 5.1: Treinamento teste com Naïve Bayes

A tabela 5.1 apresenta o resutado da classificação com ½ dados de treinamento e ½ de testes na qual mostramos apenas informações de 10 características com vista a exemplificarmos associação do termo a classe p.e. a presença do termo "inept" tem a taxa de probabilidade de 14 vezes mais de ser negativo do que positivo, porém o termo "nomination" tem a probabilidade de ser 12 vezes mais de ser positivo do que negativo. Tanto o teste como treinamento contribuem com a mesma quantidade de instâncias em função da divisão ser 50 porcento para cada conjunto. As categorias neg:pos representa a quantidade de termos falsos positivos e pos:neg representa a verdadeiros negativos no documento.Todos os classificadores NLTK utilizam um dicionário para mapear as carecteristica para os seus valores, assim cada palavra encontrada é associada ao valor true.

Atributo	Frequência	Classe	Taxa/Razão
wonderful	2	pos : neg	14.5 : 1.0
boring	2	neg : pos	11.2 : 1.0
bad	5	neg : pos	10.8 : 1.0
stupid	2	neg : pos	10.8 : 1.0
outstanding	1	pos : neg	10.2 : 1.0
we	8	pos : neg	9.1 : 1.0
learns	2	pos : neg	8.5 : 1.0

Tabela 5.2: Extracção de Características com Bolsa de Palavras

A tabela 5.2 mostra a bolsa de palavras, onde as palavras são associadas a frequência de ocorrência, dados utilizados para a realização da análise.

Classificador	Corpus	Bolsa de Palvras	Precisão	Precisão Pos	Cobertura Pos	Precisão Neg	Cobertura Neg
Naïve Bayes	Rewiew Movie	Presença	0.81	0.76	0.92	0.90	0.70

Tabela 5.3: Avaliação de Características Extraídas

A tabela 5.3 mostra a avaliação do classificador com as características extraídas do corpus resulta em 81% de precisão.

Classificador	Corpus	Stopword	Precisão	Precisão Pos	Cobertura Pos	Precisão Neg	Cobertura Neg
Naïve Bayes	Rewiew Movie	Presença	0.81	0.76	0.92	0.90	0.70
Naïve Bayes	Rewiew Movie	Removidos	0.81	0.75	0.93	0.90	0.68

Tabela 5.3: Treinamento e teste Stopword

A tabela 5.3 mostra os resultados obtidos com a aplicação do filtro de remoção de *stopwords e* sem aplicação do mesmo. Vários ficheiros são classificados correctamente como positivos, com 76% aproximadamente e cobertura 92%. Isto significa que há muito poucos Falsos Negativos nas classes Positivas. Mais dos ficheiros classificados como positivos 76 % foram classificados correctamente o que não traduz uma boa precisão para 16% de Falsos Positivos para a classe Positiva. A precisão se comparada com a aplicação do filtro baixou em 0.6%, com base nisto podemos afirmar que provavelmente existem poucos stopwords no corpus, porque o resultado não se alterou significativamente com a utilização do filtro. Muitos ficheiros que são identificados como Negativos aproximadamente 90% foram classificados correctamente o que representa uma boa precisão, isto significa que existem poucos Falsos Positivos na classe Negativa. A explicação que se pode dar em torno destes resultados é a seguinte: normalmente as pessoa utilizam palavras positivas e negativas sempre na realização da avaliação, as palavras negativas são precedidas obrigatoriamente do termo não p.e. "não é excelente". No modelo de balsa de palavras assume-se que as palavras são independentes umas das outras, e desta forma é difícil relacionar ou associar "não é excelente" como Negativo. Se este for o problema então estas métricas melhorarão se for treinado o algoritmo com uma grande quantidade palavras. Outra possibilidade é a grande quantidade de palavras neutras, mais o classificador trata todas palavras e assigna-as a classe Positiva e Negativa, talvez sejam consideradas como Positivas porque o classificador não as conhece, logo a sua eliminação melhora as métricas, classificar as melhores características apenas de um conjunto que expressam sentimentos é traduzido no conceito de ganho de informação.

Classificador	Corpus	Stopword e Bigrama	Precisão	Precisão Pos	Cobertura Pos	Precisão Neg	Cobertura Neg
Naïve Bayes	Rewiew Movie	Sem a utilização	0.81	0.76	0.91	0.90	0.70
Naïve Bayes	Rewiew Movie	Com a utilização	0.84	0.82	0.88	0.87	0.80

Tabela 5.4: Treinamento e teste com Bigrama e stopword

Na tabela 5.4 vemos claramente que a inclusão de Bigrama melhora substancialmente a precisão da classificação. A hipótese é que as pessoas dizem coisas como "não é grande", que é uma expressão negativa que o modelo bolsa de palavras poderia interpretar como positivo, uma vez que vê o termo "grande" como uma palavra separada. Para mapear bigramas utilizam-se as funções *Finder e Meansure, o Finder* mantém duas frequências internas, uma para palavras individuas e outra para bigrama, assim com esta distribuição de frequência pode criar *pontuações* para *bigrama*, utilizando a função *Meansures*, que ajudará a medir a relação de duas palavras.

Classificador	Corpus	Bigrama e melhores caracterís ticas	Precisão	Precisão Pos	Cobertura Pos	Precisão Neg	Cobertura Neg
Naïve Bayes	Rewiew Movie	Sem	0.73	0.65	0.98	0.96	0.48
Naïve Bayes	Rewiew Movie	Melhores Caracterí sticas	0.92	0.91	0.93	0.93	0.91

Tabela 5.5: Treinamento e teste com utilização de Bigrama e melhores características

A tabela 5.5 mostra que a utilização do método Bigrama não é diferencial quando se usa apenas atributos relevantes. Neste caso a melhor maneira e que constitui uma mais-valia é a sua utilização através de cálculo de precisão e cobertura, com os bigramas, obtém-se um desempenho mais uniforme em cada classe, sem este método a precisão e cobertura são menos equilibradas. Mas as diferenças podem depender de seus dados particulares. Para selecionar das melhores características o método bigrama foi empregue cujos resultados diferem-se com os primeiros resultados da análise onde foram utilizadas todas as características. A precisão é 19% superior quando se utiliza os 1000 melhores

termos e a precisão da classe Pos aumentou em 26%, enquanto cobertura negativo melhorou em torno de 43%, trata-se de um aumento expressivo sem a redução dos pos cobertura e também vemos um ligeiro aumento na precisão dos resultados para a classe neg. Se for melhorada a selecção de características melhorará a classificação. A redução de dimensionalidade é importante para melhorar o desempenho do classificador, também para inutilizar os dados que não contribuem para o ganho de informação o que é recomendável quando os dados pioram o desempenho do modelo.

5.3.1.1 Máxima Entropia

Representa o cálculo de ganho de informação com base na medida utilizada, por tanto mede-se a falta de heterogeneidade entre os dados submetidos a classificação, caso o conjunto de dado seja heterogéneo a entropia máxima é igual a 1, é discriminativo i.e. dado um conjunto de características binárias ele aprende diretamente, aplica probabilidade condicional p (c|d). O classificador considera todas as distribuições de probabilidade que são empiricamente consistente i.e. a frequência estimada de uma característica ocorrer numa classe do conjunto de dados de treinamento é igual a frequência real. Uma vez que tenhamos que calcular a entropia do conjunto original de de valores de entrada, podemos determinar o quanto mais organizados é o modelo. Os resultados de testes podem ser visualizados parcialmente no apêndice A ou na tabela 5.6.

Interação	Probabilidade	Precisão
1	-0.69315	0.500
2	-0.16114	0.905
3	-0.11627	0.903
4	-0.10458	0.905
5	-0.09762	0.908
6	-0.09169	0.915
7	-0.08795	0.916
8	-0.08664	0.920
9	-0.08678	0.919

10	-0.08824	0.917
11	-0.09051	0.915
12	-0.09371	0.913
13	-0.09822	0.907
14	-0.10222	0.902
15	-0.10605	0.900
16	-0.10977	0.898
17	-0.11311	0.893

Tabela 5.6: Treinamento e teste com Máxima Entropia

A tabela 5.6 mostra os resultados parciais do treinamento e teste com o algoritmo Máxima Entropia. Usa o teste estátisticos não parametrizado para análise da ordem a precisão a cada interacção tenda convergir ou aproximar-se de 1. O classificador considera todas as distribuições de probabilidades que são empiricamente consistentes no conjunto de dados de treinamento e escolhemos a distribuição que maximiza a entropia. A distribuição de probabilidade é empiricamente consistente quando o conjunto de dados de treinamento de cada documento é representado por um vector de palavras, e assim constrói-se a distribuição de probabilidade que representa o documento.

5.4 Conclusões do Capítulo

Os resutados obtidos com os experiências realizados nos dois algoritmos i.e. Naive Bayes e Máxima Entropia mostram de forma clara que para obtenção de melhor modelo ou resultados dos testes não depende apenas da utilização do melhor conjunto de dados, mais sim, de outros parâmetros e factores preponderantes t.c. a remoção de stopwords, a extracção e utilização das melhores características, a utilização das técnicas t.c bigramas simples ou combinado com outros métodos de filtração que convergem para proporcionar ganhos de informação, porque solucionam os problemas ligados a questão da dimensionalidade.

Quando se classifica modelos com milhares de características, p.e. categorização de documentos ou texto, é preciso observarmos que várias características podem ser irrelevantes ou insignificantes, embora estas características sejam comuns a todas as classes, contribuem muito pouco para o processo de classificação, individualmente são inofensivos mas em conjunto podem diminuir a performance do modelo. Os algoritmos que fazem o superajuste dos dados são bons na reclassificação dos dados utilizados, porém deficientes nos dados que ainda não foram utilizados. A performance é comprometida quando a quantidade de atributos é expressivo para criação de modelos, quando se utiliza apenas características mais relevantes pode-se aumentar a performance, e ao mesmo tempo diminuir o tamanho do modelo, o que resulta em menos memória e treinamento e classificação mais eficientes.

O ganho de informação é uma ferramenta utilizada em AM como recurso para a representatividade de termos, formalmente consiste em medir através de predição, a categoria associada a um determinado termo em relação a outras classes.

Capítulo 6

Considerações Finais

6.1 Considerações

Conforme referido ao longo dos capítulos a mineração de texto ou opiniões é uma área de inteligência artificial recente e está a ganhar terreno e importância nos últimos anos dado à dinâmica e a governação das medias sócias e a novas formas de fazer e divulgar negócios e actividades em geral, normalmente facilitados ou coadjuvados pelas tecnologias de informação e comunicação. O processo tem como principal objectivo a avaliação e análise de um volume expressivo de informações e dados não estruturadas ou semi-estruturadas para obtenção de conhecimento que servirá para a tomada de decisão, assim urge realçar, por conseguinte que se trata de uma tarefa que requer a combinação de várias técnicas e métodos quase sempre para a sua realização, o que era impossível de imaginar há alguns anos atrás, tornou-se possível graças às pesquisas e ferramentas produzidas para o tratamento de dados não estruturadas, sobretudo na área de processamento de Linguagens Naturais e Aprendizagem automática. Com os métodos de classificação supervisionada obtemos a ideia de automatização do processo como um todo e criam-se condições para viabilizar a interpretação dos resultados do processamento.

O benefício de mineração de texto está fundamentalmente associado a grande quantidade de informações valiosas latentes em textos que não estão disponíveis em formatos de dados estruturados clássicos, por várias razões: o texto foi sempre a forma padrão de armazenar informações por centenas de anos, e, principalmente questões associados a tempo, constrangimentos pessoais e custo proíbem-nos de transformá-los em formatos bem estruturados (p.e. *frames* ou tabelas de dados), visando o seu tratamento e de formas a tirar o máximo proveito das informações e dados contidos nestes.

A mineração de textos de maneira geral actualmente é imprescindível para a inferência ou dedução em pesquisas ou peritagem para detecção de autenticidade de documentos por exemplo no contexto estatístico é amplamente utilizada em pesquisas e negócios inteligentes fundamentalmente para propósitos tais como: a utilização de técnicas de análise semântica em bioinformática, a utilização de métodos de estatística para automação de investigação jurisdicional o que auxilia na tomada de decisões e acelera o processo como um todo, a detecção de plágios nas universidades e editoras como mecanismo para preservar os direitos de autores, detecção de spam através da utilização da inferência estatística como medida para evitar prejuízos e todos os constrangimentos causados por estes códigos, a medição de preferências dos clientes por meio da análise o que proporciona ganhos em termos de divulgação e alinhamento das acções da organização a entidades e parceiros que proporcionam vantagens competitivas e sobretudo o apredizado que se obtém com os resultados dos estudos e interpretação dos dados e informações disponíveis.

Em suma é interessante realçar a complexidade da actividade da análise de sentimentos haja vista, a gama de limitações que a caracterizam advindas da diversidades de fontes e dados produzidos quotidianamente nas organizações e por individualidades em escala expressiva e termos linguísticos, cuja qualidade dos mesmos está ligada directamente a factores tais como: a dinâmica da interacção inerente a sua produção, o tipo de comunicação em que são produzidos, que pode ser síncrono ou assíncrono, o tipo de linguagem quase sempre pouco literário e uniforme, admitindo amplamente o empregue de abreviações, gírias e outras formas de escrita e fala próprias do contexto das médias, outrossim a diversidade de fontes e a facilidade inerente a sua produção.

Com base nestes pormenores a fase de pré- processamento de dados é extremamente árdua. Porém é realizável, pois a análise de sentimento conta com diversas técnicas e sistemas interessantes e métodos para a extracção e normalização e padronização do texto de maneira automática. E dado o grau de importância deste processo e do contexto, várias pesquisadores e organizações nos últimos anos têm-se dedicado e buscado investir em mecanismos e ferramentas de excelências com finalidade de encontrar mecanismos que criam modelos para servirem de base para testes de vários exemplos que proporcionam resultados aceitáveis para os objectivos ou finalidade que se pretende esclarecer, como forma de acelerar e apoiar a tomada de decisões.

Para a classificação dos dados numa vertente mais ampla e pelos desafios da era moderna as ferramentas de classificação automática são imprescindíveis uma vez que são de importância capital e exploram técnicas de processamento em Linguagem Natural que agilizam de forma eficiente e eficácia a extracção de dados, a análise léxical, sintático e semântico do texto e auxiliam na classificação do documento com a aplicação de diversidade de algoritmos. Estes algoritmos contribuem sob maneiras no propósito de facilitar a actividade do homem na realização das suas actividades ou tarefas, a produção de algoritmos mais eficazes para a detecção da subjectividade nas frases do texto ou documentos, visa principalmente separar de formas mais clara os factos das opiniões, uma vez que para o contexto da análise o fundamental assenta-se nas frases ou palavras subjectivas, as técnicas também devem explorar o contexto de detecção de sarcasmos que defacto é um desafio enorme, dado o volume cada vez maior de documentos que o envolvem e a tendência para o seu avanço. Outro aspecto relevante é associado as referências múltiplas de itens numa frase com opiniões diferentes também é um desafio enorme porque pode confundir o classificador.

A análise de sentimento veio evolucionar sobre que maneira a forma pela qual podem ser tirados proveitos os conteúdo guardado em formato não estruturado. Haja visto o volume que esses dados apresentam para qualquer organização, entretanto esta actividade somente é possível com apoio das técnicas de Recuperação Informação, Aprendizagem Automática e Processamento em Linguagem Natural. Assim esta tarefa cada vez mais está a facilitar a aquisição de conhecimento em tempo hábil através de um processo automático, o que permite a obtenção dos resultados com alta precisão e qualidade e fiáveis, embora esteja suscetível a algumas limitações dada a

heterogeneidade e a ambiguidade que caracteriza os dados disponíveis para o efeito. Actualmente as ferramentas e algoritmos proporcionam mecanismos eficientes que vão desde a recuperação, extracção de características relevantes e melhores características, a detecção da subjectividade no documento cujas propriedades nos permitiram realizar a análise em torno da polaridade obtida nas frases sentimentais no documento. Estes mecanismos oferecem um processo cujas etapas não são totalmente inclusivas. i.e. ainda não existem ferramentas que fornecem uma interface integrada para a realização de todas as fases de mineração de texto e a análise consequentemente a título do que acontece com mineração de dados, o que torna a análise de documentos ou textos ainda uma tarefa árdua.

É uma área que fundamentalmente e pela sua importância deve proporcionar ganhos substanciais e um vasto capital em termos de retorno de investimento, logo qualquer organização ou entidade particular deve aproveitar esta tecnologia para segmentar o conhecimento em torno do seu negócio sobretudo a nível de aceitação ou avaliação ou visão ou juízo que os clientes e os parceiros de maneira geral tem em relação aos seus produtos, suas actividades quando se tratar de um político, estilo musical, acontecimento ou um determinado evento ou qualquer assunto em geral. Logo os dados antes tidos como fontes apenas de consultas mecânicas, com a área tornaram-se de grande valia para a obtenção de conhecimento. Outro aspecto importante a frisar com esta tecnologia é a detecção de spam e outros fontes de riscos que comprometem o bom funcionamento e a credibilidade nas transacções electrónicas e assim como a violação de direitos de autoria com a apropriação indébita de trabalhos de outrens, prática conhecido como plágio.

As limitações são enormes o que ainda requer um caminho enorme a percorrer para a solidificação e a facilitação da realização desta tarefa, por exemplo existem poucos corpora que podem ser utilizados para extrair subjectividade e realizar a actividade de análise de sentimentos, p.e. em Inglês os estudos mais bem-sucedidos foram feitos em torno de filmes, em língua portuguesa não existe um corpus que retrata a mesma questão, já referente a políticos existe um corpus porém sem sombras a duvidas é ínfimo para o processo, e não se adequa a sintaxe dos mecanismos de PLN e algoritmos de classificação, logo a tarefa dos pesquisadores é árduo na realização de testes e produção de resultados em função da falta de opções para construção de modelos.

A análise dos dados textuais para extracção de informações úteis no documento ou texto depende fundamentalmente de certos aspectos chaves tais como entender a estrutura morfológica e sintáctica da língua a utilizar como referência, pois textos extraídos como já referenciado podem conter certas anomalias t.c. erros de grafia, abreviações, pontuação incorrecta, gírias etc, o que dificulta imensamente o processo na sua plenitude.

Como principal contributo é fornecer um método e interface integrada que permita ou facilita o processo, em todas as fases i.e. desde a obtenção dos textos a geração dos resultados, de formas a suprir deficiências na aplicação das técnicas e métodos de análise de sentimentos. Fornece um material guia para incentivar a utilização ou adopção de técnicas de PLN para a realização dos mais diversas propósitos.

O protótipo fornece uma estratégia de suporte a análise de sentimentos, cujo método acopla os módulos de extracção, etiquetagem, análise de subjectividade e análise de sentimento do documento ou texto de entrada.

6.1.1 Trabalhos Futuros

A área é nova e reveste uma importância capital para o direccionamento das acções e das entidades e corporações em geral por conseguinte.

- Construção de corpora em Português que se adequa as tecnologias em exploração;
- Construção de uma gramática para reconhecer línguas nacionais angolanas, com vistas a realização da análise léxico, sintático e semântico
- Desenvolver um sistema de análise de sentimentos expressos em línguas nacionais, t.c. Kicongo e Kimbundu.

6.1.2 Recomendações

O projecto faz uma abordagem ampla sobre a importância e a aplicabilidade de análise de sentimentos de textos como instrumento eficiente na aquisição de conhecimento sobre várias temáticas e problemáticas.

Para pesquisadores este projecto deve servir de guia na realização deste processo e os desafios que encerra a sua realização, como um campo vasto que deve trazer benefícios tanto para organizações como para entidades individuais.

Para as organizações e a audiência em geral o material serve para a área de expecialidade de Inteligência Artificial no âmbito de aplicação de algoritmos para os mais diversos propósitos de actividades.

Referências Bibliográficas

[1]. Alexandra Balahur Dobrescu "Methods and Resources for Sentiment Analysis in Multilingual" Tesis University of Alicante 2011.Documents of Different Text Types.

[2]. Alexandra Balahur, Jesús M. Hermida, and Andrés Montoyo: Detecting implicit expressions of emotion in text: A comparative analysis. Decision Support Systems 53(4): 742-753,2012.

[3]. Andrew McCallum, Kamal Nigam A Comparison of event Models for Naive Bayes Text Classification, AAAI-98 workshop on learning for text categorization, 1998.

[4]. Andrew McCallum, Kamal Nigam, " Employing EM in pool-based active learning for text classification", Proceedings of ICML-98, 15th International Conference on Machine Learning, 1998.

[5]. António Alexandre Mello Ticom "Aplicação das Técnicas de Mineração de Textos e Sistemas Especialistas na Liquidação de Processos Trabalhistas" Dissertação de Mestrado, Universidade Federal do Rio de Janeiro (2007).

[6]. Bernardo J.M. And Smith " Bayesian Theory", Chichester Wiley, 1994;

[7]. Bernardo, José M. & Adrian F. SMITH (1994), Bayesian Theory. New York: Wiley.

[8]. Bo Pang "Automatic Analysis of Document Sentiment, Dissertation, Cornell University (2006).

[9]. Boiy, E. & Moens, M. (2009). A machine learning approach to sentiment analysis in multilingual Web texts. Information Retrieval 12(5): 526-558.

[10]. Boiy, E., Hens, P., Deschacht, K. , & Moens , M. F. Automatic sentiment analysis of on - line text. In Proceedings of the 11th International Conference on Electronic Publishing. Vienna, Austria, 2007.

[11]. Bosangit, C., McCabe, S. & Hibbert, S. (2009). What is Told in Travel Blogs? Exploring.

[12]. Brooke, J., Tofiloski, M. and Taboada, M. Cross-linguistic sentiment analysis: From English to Spanish. In Proceedings of RANLP (2009).

[13]. Carlo Strapparava, Rada Mihalcea. "Learning to identify emotions in text." Proceedings of the 2008 ACM symposium on Applied computing, March 16-20, 2008, Fortaleza, Ceara, Brazil.

[14]. Clark, Alexander, Fox, Chris E Lappin, Shalom. 2010. The Handbook Of Computacional Linguistics And Natural Language Processing. s.l.: Blackwell Publishing, 2010.

[15]. Cortes, Corinna; and Vapnik, Vladimir N.; "Support-Vector Networks", Machine Learning, 20, 1995.

[16]. Deborah Ribeiro Carvalho, Rio De Janeiro, Árvore De Decisão e Algoritmo Genético Para Tratar o Problema De Pequenos Disjuntos em Classificação De Dados, Rj – Brasil, Dezembro De 2005.

[17]. Ding, X., Liu, B. and Yu, P.S. A holistic lexicon-based approach to opinion mining. In Proceedings of the Conference on Web Search and Web Data Mining (2008).

[18]. Ding, X., Liu, B. and Zhang, L. Entity discovery and assignment for opinion mining applications. In Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2009).Dave et al. 2003

[19]. Dragut, E.C., Yu, C., Sistla, P. and Meng, W. Construction of a sentimental word dictionary. In Proceedings of ACM International Conference on Information and Knowledge Management (2010).

[20]. Du, W., Tan, S., Cheng, X. and Yun, X. Adapting information bottleneck method for automatic construction of domain-oriented sentiment lexicon. In Proceedings of ACM International Conference on Web Search and Data Mining (2010).

[21]. Erik Tromp "Multilingual Sentiment Analysis on Social Media", Master's Thesis, Department of Mathematics and Computer Science Eindhoven University of Technology (2011).

[22]. Esuli, A. and Sebastiani, F. Determining term subjectivity and term orientation for opinion mining. In Proceedings of Conf. of the European Chapter of the Association for Computational Linguistics (2006).

[23]. Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996) "Knowledge discovery and data mining toward a unifying framework.", In Proceeding of The Second Int. Conference on Knowledge Discovery and Data Mining, pages 82-88.

[24]. Feldman, R., Rosenfeld, B., Bar-Haim, R. and Fresko, M. the stock sonar—sentiment Analysis of Stocks Based on a Hybrid Approach. IAAI-12 (2011), 1642–1647.

[25]. Feldman, Ronen e Sanger, James. 2007. The Text Mining Handbook. s.l. : Cambridge, 2007.

[26]. Fellbaum, C.D. Wordnet: An Electronic Lexical Database. MIT Press, Cambridge, MA, 1998.

[27]. Feng, S., Bose, R. and Choi, Y. Learning general connotation of words using graph-based algorithms. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (Edinburgh, Scotland, UK, 2011). 1092–1103.

[28]. Gamon, Michael (2004).Sentiment classification on customer feedback data: Noisy data, large feature vectors, and the role of linguistic analysis.

[29]. IncProceedings of the 20th International Conference on Computational Linguistics,Geneva, Switzerland, 23–27 August 2004, pp.611–617

[30]. Gonçalves, L. A. O. Trabalho docente e subjetividade: embate teórico e novas perspectiva. Revista da Faculdade de Educação, São Paulo, USP, v. 22, p. 43-71, jul./dez. 1996.

[31]. Hai, Z., Chang, K. and Kim, J-j. Implicit feature identification via co-occurrence association rule mining. Computational Linguistics and Intelligent Text Processing (2011), 393–404.

[32]. HARRIS, Z. S. Computable Syntactic Analysis: the 1959 Computer Science-Analyzer. In: Papers in Structural and Transformational Linguistics. Holanda, D. Reidel Publishing Company, 1970[1959] p. 252-277

[33]. Hatzivassiloglou, V. and K. McKeown, Predicting the semantic orientation of adjectives. In Proceedings of the Joint ACL/EACL Conference (1997), 174–181.

[34]. Hu, M. and Liu, B. Mining and summarizing customer reviews. In Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (2004), 168–177.

[35]. Hu, M. and Liu, B. Mining opinion features in customer reviews. In Proceedings of AAAI (2004), 755–760.

[36]. Jackson, Peter e Moulinier, Isabelle. 2002. Natural Language Processing For Online Applications: Text Retrieval, Extraction And Categorization. s.l. : John Benjamins B.V., 2002.

[37]. Jakob, N. and Gurevych, I. Extracting opinion targets in a single-and cross-domain setting with conditional random fields. In Proceedings of Conference on Empirical Methods in Natural Language Processing (2010).

[38]. Jindal, N. and Liu, B. Identifying comparative sentences in text documents. In Proceedings of ACM SIGIR Conf. on Research and Development in Information Retrieval (2006).

[39]. Jurafsky, Daniel E Martin, James H. 2006. Speech And Language Processing: An Introduction to Natural Language Processing, Computacional Linguistics And Speech Recognition. 2006.

[40]. Kamps, J., Marx, M., Mokken, R.J. and de Rijke, M. Using WordNet to measure semantic orientation of adjectives. LREC, 2004.

[41]. Kim, Elsa and Sam Gilbert. "Detecting Sadness in 140 Characters: Sentiment Analysis and Mourning Michael Jackson on Twitter" Web Ecology Project, August 2009.

[42]. Kim, S.-M. and Hovy, E. Crystal: Analyzing predictive opinions on the Web. In Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (2007).

[43]. Kou, Z.; Cohen, W. W.; Murphy, R. F. (2005) High-cobertura protein entity recognition using a dictionary. Bioinformatics, v. 21, p. i266-273. Suppl. 1.

[44]. Krauthammer M, Kaufmann CA, Gilliam TC, Rzhetsky A. Molecular Triangulation: Bridging Linkage and Molecular-Network Information for Identifying Candidate Genes in Alzheimer's Disease. Proc Natl Acad Sci USA. 2004.2004 Oct 19;101(42):15148-53

[45]. Krauthammer, M.; Nenadic, G. (2004) Term identification in the biomedical literature. Journal of Biomedical Informatics, v. 37, n. 6, p. 512-526.

[46]. *Ku*, Lun-Wei, *Ke*, Kai-Jie and *Chen*, Hsin-Hsi "Opinion Analysis on CAW. 2.0 Datasets. Proceedings of Workshop on Content Analysis in the Song, 2007.

[47]. Lafferty, J., McCallum, A. and Pereira, F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proc. 18th International Conf. on Machine Learning. Morgan Kaufmann, San Francisco, CA, 2001, 282–289.

[48]. Lin, D. Minipar; http://Webdocs.cs.ualberta.ca/lindek/minipar.htm. 2007.

[49]. Liu, B. Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2012.

[50]. Liu, B., Sentiment analysis and subjectivity. Handbook of Natural Language Processing. N. Indurkhya and F.J. Damerau, eds. 2010.

[51]. Loughran, T. and Mcdonald, B. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. The Journal of Finance 66, 1 (2011), 35–65.

[52]. Marcelo Ribeiro, Manoel R. Neto, Ricardo B. C. Prudêncio, Local Feature Selection in Text Clustering. In: International Conference on Neural Information Processing, 2008, Auckland. Lecture Notes in Computer Science, 2008 (ICONIP-2008).

[53]. Meeting of the Association for Computational Linguistics, pp. 417–424, Philadelphia, 2002.

[54]. Mohammad, S.M. and Turney, P.D. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text (2010).

[55]. Monard, m. C.; Baranauskas, J. A. Conceitos sobre aprendizagem automática. In: REZENDE, S. O. (Ed.).

[56]. Sistemas inteligentes:fundamentos e aplicações. São Carlos:Manole, 2003. p. 89 -114. cap. 4.

[57]. Narayanan, R., Liu, B. and Choudhary, A. Sentiment analysis of conditional sentences. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (Singapore, 2009). Association for Computational Linguistics, 180–189.

[58]. Netzer, O., Feldman, R., Fresko, M. and Goldenberg, Y. Mine your own business: Market structure surveillance through text mining. Marketing Science, 2012.

[59]. Nigam, Kamal Nigam, Andrew McCallum, Tom M. Mitchell, and W. Cohen. (2000). "Text Classification from Labeled and Unlabeled Documents Using EM." In: Machine Learning. doi:10.1023/A:1007692713085.

[60]. Pang, B. and Lee, L. A Sentimental Education: Sentiment Analysis using Subjectivity Summarization based on minimum cuts. In Proceedings of the Association for Computational Linguistics (2004), 271–278Matos 2010

[61]. Pang, B. and Lee, L. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2, 1-2 (2008), 1–135.

[62]. Pang, B., Lee, L. and Vaithyanathan, S. Thumbs up? Sentiment Classification using machine learning techniques. In Proceedings of EMNLP-02, 7th Conference on Empirical Methods in Natural Language Processing (Philadelphia, PA, 2002). Association for Computational Linguistics, Morristown, NJ, 79–86.

[63]. PANG, BO. 2006. Automatic Analysis of Document Sentiment. 2006.

[64]. Paula Carvalho, Luís Sarmento, Jorge Teixeira, Mário J. Silva: Liars and Saviors in a Sentiment Annotated Corpus of Comments to Political Debates. ACL (Short Papers) 2011: 564-568.

[65]. Pedro Oguri," Aprendizado de Maquina para o Problema de Sentiment Classification, Dissertação de Mestrado, PUC-RIO, Rio de Janeiro, 2006.

[66]. Pellucci, Paula Silva, Ladeira, Utilização de Técnicas de Aprendizagem automática no Reconhecimento de Entidades Nomeadas no Português, Centro Universitário de Belo Horizonte, Belo Horizonte, MG 2011.

[67]. Peng, W. and Park, D.H. Generate adjective sentiment dictionary for social media sentiment analysis using constrained nonnegative matrix factorization. In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (2011).

[68]. Platt, J. C., "A Resource-Allocating Network for Function Interpolation," Neural Computation, 3(2):213-225, (1991)

[69]. Popescu, A.-M. and Etzioni, O. Extracting product features and opinions from reviews. In Proceedings of Conference on Empirical Methods in Natural Language Processing (2005).

[70]. Priscilla Inácia Neves " Uma Estratégia para Apoiar a Decisão Baseada em Mineração de Textos Livres, Dissertação de Mestrado, Rio de Janeiro,2012.

[71]. Qiu, G., Liu, B., Bu, J. and Chen, C. Opinion word expansion and target extraction through double propagation. Computational Linguistics 37, 1 (2011), 9–27.

[72]. Riloff, E. and Wiebe, J. Learning extraction patterns for subjective e Expressions. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (2003).

[73]. Rocchio, J. Relevance feedback in information retrieval Prentice Hall, Ing. Englewood Cliffs, New Jersey , 1971

[74]. Ruy Luiz Milidiú, Cícero Nogueira dos Santos, Julio Cesar Duarte. Portuguese Corpus-Based Learning Using ETL. Journal of the Brazilian Computer Society - Volume 14 - Número 4, 1996.

[75]. S. de Amo: Técnicas de Mineração de Dados, Programa de Mestrado em Ciência da Computação, Universidade Federal de Uberlândia, 2003.

[76]. Shang, W., Huang, H., Zhu, H., Lin, Y., Qu, Y., and Wang Z., " A Noval Feature Selection Algorithm for text catogorization." Elsevier, science Direct Expert system with application -2006, 33(1), pp.1-5, 2006.

[77]. Simon (1983).

[78]. Sivic, Josef (April, 2009). "Efficient visual search of videos cast as text retrieval". IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 31, NO. 4. IEEE. pp. 591–605.

[79]. Stone, P. The general inquirer: A computer approach to content analysis. Journal of Regional Science 8, 1 (1968).

[80]. Taboada, M., J. Brooke, J., Tofiloski, M., Voll, K. and Stede, M. Lexicon-based methods for sentiment analysis. Computational Linguistics 37, 2 (2011), 267–307.

[81]. Tan, A.-H. (1999). "Text mining: The state of the art and the challenges". In Proceddings,

[82]. Tanner M.A "Tools for Statiscal infere: Methods for the exploration of posterior distribution and linkihood fuctions (2 nd end.). New York: Springer, 1993.

[83]. Travel Blogs for Consumer Narrative Analysis. In W. Höpken, U. Gretzel & R. Law, eds. Information and Communication Technologies in Tourism 2009. Wien New York: Springer. pp.61-71.

[84]. Tsur, O., Davidov, D. and Rappoport, A. A great catchy name: semi-supervised recognition of sarcastic sentences in online product reviews. In Fourth International AAAI Conference on Weblogs and Social Media (2010).

[85]. Turney, P. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In Proceedings of the Association for Computational Linguistics (2002), 417 –424.

[86]. TURNEY,P.,"Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews", In: Proceedings of ACL-02, 40th Annual

[87]. Vapnik, V., The Nature of Statistical Learning Theory, Springer-Verlag, 1995.

[88]. Wan, X. Using bilingual knowledge and ensemble techniques for unsupervised Chinese sentiment analysis. In Proceedings of the 2008 Conference on Empirical

Methods in Natural Language Processing (Honolulu, Hawaii, 2008). Association for Computational Linguistics, 553 –561.

[89]. Wilson, T., Wiebe, J. and Hoffmann, P. Recognizing contextual polarity in phrase-level sentiment analysis. In Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (2005), 347–354.

[90]. Wiebe, Janyce, Theresa Wilson, and Matthew Bell. 2001. Identifying collocations for recognizing opinions. In Proceedings of the ACL-01 Workshop on

[91]. Collocation: Computational Extraction,Analysis, and Exploitation, Toulouse,France, July 7, pages 24–31

[92]. Wu, Y., Zhang, Q. Huang, X. and Wu, L. Phrase dependency parsing for opinion mining. In Proceedings of Conference on Empirical Methods in Natural Language Processing (2009).

[93]. Yessenov, Kuat, and Sasa Misailovic. "Sentiment Analysis of Movie Review Comments" Massachusetts Institute of Technology, Spring 2009.

[94]. Yu, H. and Hatzivassiloglou, V. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (2003).

Apêndice A

Esta parte traz informações sobre trechos de códigos referentes a aplicação de técnicas e consequentemente os resultados obtidos com a realização de experiências para analisar os dados.

Anexo A.1 - Scripts de Código

```
from nltk.corpus import movie_reviews
import random
movie_reviews.categories() [u'neg', u'pos']
documents = [(list(movie_reviews.words(fileid)), category) for category in
movie_reviews.categories()
for fileid in movie_reviews.fileids(category)]
random.shuffle(documents)
documents[0]
```

Trecho de código A.1: Categorizar texto

A.1 representa o exemplo de categorização de filmes nas classes positivas e negativas.

Resultado

([u'i', u"", u've', u'got', u'to', u'admit', u'it', u'.', u'.', u'.', u'i', u"", u'm', u'a', u'huge', u'jim', u'carrey', u'fan', u'.', u'i', u'loved', u'the', u'first', u'ace', u'ventura', u',', u'as', u'well', u'as', u'the', u'mask', u'and', u'dumb', u'and', u'dumber', u'--', u'and', u'even', u'in', u'batman', u'forever', u'(', u'which', u'was', u'a', u'pretty', u'awful', u'movie', u')', u',', u'carrey', u'was', u'one', u'of', u'the', u'few', u'people', u'to', u'come', u'off', u'looking', u'reasonably', u'good', u'.', u'until', u'i', u'saw', u'ace', u'ventura', u'2', u',', u'i', u'had', u'no', u'idea', u'how', u'people', u'could', u'find', u'the', u'guy', u'annoying', u'.', u'sadly', u',', u'ace', u'ventura', u'2', u'shows', u'just', u'how', u'irritating', u'and', u'annoying', u'carrey', u'can', u'be', u'.', u'carrey', u'goes', u'through', u'the', u'same', u'schtick', u'he', u'went', u'through', u'in', u'the', u'first', u'ace', u'movie', u',', u'but', u'this', u'time', u'it', u'is', u'no', u'longer', u'funny', u'--', u'it', u'is', u'just', u'a', u'rehash', u'of', u'many', u'of', u'the', u'same', u'jokes', u'used', u'in', u'ace', u'i', u'.', u'the', u'plot', u'sees', u'pet', u'detective', u'ace', u'ventura', u'retiring', u'after', u'failing', u'to', u'save', u'a', u'raccoon', u'(', u'in', u'a', u'reasonable', u'cliffhanger', u'spoof', u')', u'.', u'soon', u'he', u'is', u'called', u'out', *def*

```
document_features(document):document_words=set(document)
features = {}
```

```
for word in word_features:
features['contains(%s)' % word] = (word in document_words)
 return features
 featuresets = [(document_features(d), c) for (d,c) in documents]
featuresets[0]
```

Trecho de código A.2: extração de características

Resultado

Atributo	Valor	Atributo	Valor	Atributo	Valor
u'contains(waste)'	False	u'contains(lot)':	True	u'contains(*)':	False
u'contains(black)	False,	u'contains(rated)'	False	u'contains(potential)'	False
u'contains(m)'	True	u'contains(understand)	False	u'contains(drug)'	False
u'contains(case)':	False	u'contains(created)'	False	u'contains(kiss)'	False
u'contains(needed)' :	False	u'contains(c)'	False	u'contains(about)'	False
u'contains(toy)'	False	u'contains(longer)'	True	u'contains(ready)'	False
u'contains(certainly)'	False	u'contains(lame)'	False	u'contains(sadly)'	True
u'contains(anderso n)	False	u'contains(rent)'	False	u'contains(mulan)'	False
u'contains(Catherin e)	False	u'contains(fans)	False	u'contains(christopher)'	False
u'contains(them)'	False	u'contains(seen)	False	u'contains(fan)'	True
u'contains(robin)'	False	u'contains(moments)	True	u'contains(jim)	True
u'contains(cinemati c)'	False	u'contains(know)'	False,	u'contains(deal)'	False
u'contains(law)'	False	u'contains(started)'	False	u'contains(is)	True
u'contains(violence)	False	u'contains(tension)'	False	u'contains(focus)'	False
u'contains(suspects)':	False,	u'contains($)'	False	u'contains(images)'	False,
u'contains(driver)'	False	u'contains(feature)'	False	u'contains(knew)'	False,
u'contains(off)'	True	u'contains(fiction)'	False	u'contains(drawn)'	False,

u'contains(helen)'	False,	u'contains(road)	False,	u'contains(fi)'	False
u'contains(films)'	True	u'contains(troopers)'	False	u'contains(comedic)	False
u'contains(mediocre)	False,	u'contains(players)'	False	u'contains(attention)'	False

Tabela A.1: características extraidadas

Máxima Entropia

```
import nltk.classify.util
from nltk.classify import MaxentClassifier
from nltk.corpus import movie_reviews
def word_feats(words):
        return dict([(word, True) for word in words])
negids = movie_reviews.fileids('neg')
posids = movie_reviews.fileids('pos')
negfeats = [(word_feats(movie_reviews.words(fileids=[f])), 'neg') for f in negids]
posfeats = [(word_feats(movie_reviews.words(fileids=[f])), 'pos') for f in posids]
negcutoff = len(negfeats)*1/2
poscutoff = len(posfeats)*1/2
trainfeats = negfeats[:negcutoff] + posfeats[:poscutoff]
classifier = MaxentClassifier.train(trainfeats)
```

Trecho de código A.3: Máxima Entropia

Resultados de testes

Interação	Probabilidade	Precisão
18	**-0.11611**	**0.890**
19	**-0.11869**	**0.887**
20	**-0.12086**	**0.885**
21	**-0.12272**	**0.884**
22	**-0.12436**	**0.881**
23	**-0.12587**	**0.881**

24	-0.12735	0.881
25	-0.12877	0.880
26	-0.12996	0.880
27	-0.13097	0.878
28	-0.13190	0.878
29	-0.13282	0.877
30	-0.13373	0.876
31	-0.13465	0.875
32	-0.13555	0.873
33	-0.13642	0.873
34	-0.13724	0.871
35	-0.13799	0.869
36	-0.13868	0.868
37	-0.13929	0.868
38	-0.13984	0.868
39	-0.14034	0.868
40	-0.14080	0.868
41	-0.14124	0.868
42	-0.14166	0.868
43	-0.14206	0.867
44	-0.14245	0.867
45	-0.14283	0.867
46	-0.14321	0.867
47	-0.14357	0.867
48	-0.14393	0.866

49	-0.14427	0.865
50	-0.14460	0.864
51	-0.14490	0.864
52	-0.14519	0.864
53	-0.14545	0.864
54	-0.14569	0.864
55	-0.14592	0.865
56	-0.14613	0.865
57	-0.14634	0.866
58	-0.14654	0.865
59	-0.14673	0.864
60	-0.14691	0.864
61	-0.14709	0.864
62	-0.14726	0.864
63	-0.14742	0.864
64	-0.14758	0.864
65	-0.14774	0.864
66	-0.14789	0.864
67	-0.14803	0.863
68	-0.14817	0.863
69	-0.14830	0.863
70	-0.14844	0.863
71	-0.14856	0.863
72	-0.14869	0.863
73	-0.14881	0.863
74	-0.14893	0.863

75	-0.14904	0.862
76	-0.14916	0.862
77	-0.14928	0.862
78	-0.14940	0.861
79	-0.14951	0.861
80	-0.14963	0.861
81	-0.14975	0.861
82	-0.14988	0.861
83	-0.15000	0.861
84	-0.15013	0.861
85	-0.15026	0.860
86	-0.15040	0.860
87	-0.15054	0.860
88	-0.15068	0.860
89	-0.15082	0.860
90	-0.15096	0.860
91	-0.15111	0.860
92	-0.15125	0.860
93	-0.15140	0.859
94	-0.15154	0.859
95	-0.15169	0.859
96	-0.15183	0.859
97	-0.15198	0.859
98	-0.15212	0.858
99	-0.15226	0.858
100	-0.15240	0.858

Tabela A.2: Resultado de teste e treinamento

A tabela mostra os resultados de testes e treinamento do classificador Máxima Entropia com 50 interações.

Treinamento do Nives com dados Review Movie

```
import nltk.classify.util
from nltk.classify import NaiveBayesClassifier
from nltk.corpus import movie_reviews
def word_feats(words):
return dict([(word, True) for word in words])
negids = movie_reviews.fileids('neg')
posids = movie_reviews.fileids('pos')
negfeats = [(word_feats(movie_reviews.words(fileids=[f])), 'neg') for f in negids]
posfeats = [(word_feats(movie_reviews.words(fileids=[f])), 'pos') for f in posids]
negcutoff = len(negfeats)*2/4
poscutoff = len(posfeats)*2/4
trainfeats = negfeats[:negcutoff] + posfeats[:poscutoff]
testfeats = negfeats[negcutoff:] + posfeats[poscutoff:]
print 'train on %d instances, test on %d instances' % (len(trainfeats), len(testfeats))
```

Trecho de código A.4: Máxima Entropia

Resultados

Instâncias	Quantidade	Relação
Treinamento	1000	
Teste	1000	0.811

Características	Valor	Classe	Relação
whatsoever	**true**	neg : pos	**14.3 : 1.0**
inept	**true**	neg : pos	**13.7 : 1.0**
stupidity	**true**	neg : pos	**11.7 : 1.0**
ludicrous	**true**	neg : pos	**11.7 : 1.0**
nomination	**true**	pos : neg	**11.7 : 1.0**
vulnerable	**true**	pos : neg	**10.3 : 1.0**
Views	**true**	pos : neg	**10.3 : 1.0**

fictional	**true**	pos : neg	**9.7 : 1.0**
Exceptional	**true**	pos : neg	**9.7 : 1.0**
abilities	**true**	neg : pos	**9.7 : 1.0**

Tabela A.5: Treinamento com 2/4 de dados de treinamento e teste

Analogamente com ½ temos a mesma quantidade de instâncias de treinamento e testes, 10 características são mostradas

Características	**Valor**	**Classe**	**Relação**
whatsoever	**true**	neg : pos	**14.3 : 1.0**
inept	**true**	neg : pos	**13.7 : 1.0**
stupidity	**true**	neg : pos	**11.7 : 1.0**
ludicrous	**true**	neg : pos	**11.7 : 1.0**
nomination	**true**	pos : neg	**11.7 : 1.0**
vulnerable	**true**	pos : neg	**10.3 : 1.0**
Views	**true**	pos : neg	**10.3 : 1.0**
fictional	**true**	pos : neg	**9.7 : 1.0**
Exceptional	**true**	pos : neg	**9.7 : 1.0**
abilities	**true**	neg : pos	**9.7 : 1.0**

Tabela A.6: Treinamento com 1/3 de dados de treinamento e 1/7 teste

Resulta em 475 instâncias de treinamento e 1525 instâncias de teste, cuja precisão é de 88 porcentos

```
import collections
import nltk.classify.util, nltk.metrics
from nltk.classify import NaiveBayesClassifier
from nltk.corpus import movie_reviews
def evaluate_classifier(featx):
negids = movie_reviews.fileids('neg')
posids = movie_reviews.fileids('pos')
negfeats = [(featx(movie_reviews.words(fileids=[f])), 'neg') for f in negids]
posfeats = [(featx(movie_reviews.words(fileids=[f])), 'pos') for f in posids]
```

```
def evaluate_classifier(featx):
negids = movie_reviews.fileids('neg')
posids = movie_reviews.fileids('pos')
negfeats = [(featx(movie_reviews.words(fileids=[f])), 'neg') for f in negids]
posfeats = [(featx(movie_reviews.words(fileids=[f])), 'pos') for f in posids]
negcutoff = len(negfeats)*1/2
poscutoff = len(posfeats)*1/2
trainfeats = negfeats[:negcutoff] + posfeats[:poscutoff]
testfeats = negfeats[negcutoff:] + posfeats[poscutoff:]
classifier = NaiveBayesClassifier.train(trainfeats)
refsets = collections.defaultdict(set)
    testsets = collections.defaultdict(set)
    for i, (feats, label) in enumerate(testfeats):
        refsets[label].add(i)
        observed = classifier.classify(feats)
        testsets[observed].add(i)
print 'accuracy:', nltk.classify.util.accuracy(classifier, testfeats)
print 'pos precision:', nltk.metrics.precision(refsets['pos'], testsets['pos'])
print 'pos cobertura:', nltk.metrics.cobertura(refsets['pos'], testsets['pos'])
print 'neg precision:', nltk.metrics.precision(refsets['neg'], testsets['neg'])
print 'neg cobertura:', nltk.metrics.cobertura(refsets['neg'], testsets['neg'])
classifier.show_most_informative_features()
def word_feats(words):
return dict([(word, True) for word in words])
```

Trecho de código Balsa de Palavras

```
import collections
import nltk.classify.util, nltk.metrics
from nltk.classify import NaiveBayesClassifier
from nltk.corpus import movie_reviews
def evaluate_classifier(featx):
negids = movie_reviews.fileids('neg')
posids = movie_reviews.fileids('pos')
negfeats = [(featx(movie_reviews.words(fileids=[f])), 'neg') for f in negids]
```

```
posfeats = [(featx(movie_reviews.words(fileids=[f])), 'pos') for f in posids]
def evaluate_classifier(featx):
negids = movie_reviews.fileids('neg')
posids = movie_reviews.fileids('pos')
negfeats = [(featx(movie_reviews.words(fileids=[f])), 'neg') for f in negids]
posfeats = [(featx(movie_reviews.words(fileids=[f])), 'pos') for f in posids]
negcutoff = len(negfeats)*1/2
poscutoff = len(posfeats)*1/2
trainfeats = negfeats[:negcutoff] + posfeats[:poscutoff]
testfeats = negfeats[negcutoff:] + posfeats[poscutoff:]
classifier = NaiveBayesClassifier.train(trainfeats)
refsets = collections.defaultdict(set)
      testsets = collections.defaultdict(set)
      for i, (feats, label) in enumerate(testfeats):
         refsets[label].add(i)
         observed = classifier.classify(feats)
         testsets[observed].add(i)
print 'accuracy:', nltk.classify.util.accuracy(classifier, testfeats)
print 'pos precision:', nltk.metrics.precision(refsets['pos'], testsets['pos'])
print 'pos cobertura:', nltk.metrics.cobertura(refsets['pos'], testsets['pos'])
print 'neg precision:', nltk.metrics.precision(refsets['neg'], testsets['neg'])
print 'neg cobertura:', nltk.metrics.cobertura(refsets['neg'], testsets['neg'])
classifier.show_most_informative_features()
def word_feats(words):
   return dict([(word, True) for word in words])
from nltk.corpus import stopwords
stopset = set(stopwords.words('english'))
 def stopword_filtered_word_feats(words):
            return dict([(word, True) for word in words if word not in stopset])
```

Trecho de código Aplicação de Filtro Stopword

```
import collections
import nltk.classify.util, nltk.metrics
from nltk.classify import NaiveBayesClassifier
from nltk.corpus import movie_reviews
def evaluate_classifier(featx):
negids = movie_reviews.fileids('neg')
```

```
posids = movie_reviews.fileids('pos')
negfeats = [(featx(movie_reviews.words(fileids=[f])), 'neg') for f in negids]
posfeats = [(featx(movie_reviews.words(fileids=[f])), 'pos') for f in posids]
def evaluate_classifier(featx):
negids = movie_reviews.fileids('neg')
posids = movie_reviews.fileids('pos')
negfeats = [(featx(movie_reviews.words(fileids=[f])), 'neg') for f in negids]
posfeats = [(featx(movie_reviews.words(fileids=[f])), 'pos') for f in posids]
negcutoff = len(negfeats)*1/2
poscutoff = len(posfeats)*1/2
trainfeats = negfeats[:negcutoff] + posfeats[:poscutoff]
testfeats = negfeats[negcutoff:] + posfeats[poscutoff:]
classifier = NaiveBayesClassifier.train(trainfeats)
refsets = collections.defaultdict(set)
    testsets = collections.defaultdict(set)
    for i, (feats, label) in enumerate(testfeats):
        refsets[label].add(i)
        observed = classifier.classify(feats)
        testsets[observed].add(i)
print 'accuracy:', nltk.classify.util.accuracy(classifier, testfeats)
print 'pos precision:', nltk.metrics.precision(refsets['pos'], testsets['pos'])
print 'pos cobertura:', nltk.metrics.cobertura(refsets['pos'], testsets['pos'])
print 'neg precision:', nltk.metrics.precision(refsets['neg'], testsets['neg'])
print 'neg cobertura:', nltk.metrics.cobertura(refsets['neg'], testsets['neg'])
classifier.show_most_informative_features()
def word_feats(words):
  return dict([(word, True) for word in words])
from nltk.corpus import stopwords
stopset = set(stopwords.words('english'))
def stopword_filtered_word_feats(words):
        return dict([(word, True) for word in words if word not in stopset])
```

Trecho de código Aplicação de Filtro Stopword e uso de Bigrama