MEGI

MESTRADO
Estatística e Gestão de Informação

Text Mining: Análise de Sentimentos na classificação de notícias

Helder Joaquim Carvalheira Gomes

Trabalho de Projecto apresentado como requisito parcial para obtenção do grau de Mestre em Estatística e Gestão de Informação

UNIVERSIDADE NOVA ISEGI Estatística e Gestão de Informação

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

TEXT MINING: ANÁLISE DE SENTIMENTOS NA CLASSIFICAÇÃO DE NOTÍCIAS

por

Helder Joaquim Carvalheira Gomes

Trabalho de Projecto apresentado como requisito parcial para a obtenção do grau de Mestre em Estatística e Gestão de Informação, Especialização em Gestão do Conhecimento e *Business Intelligence*.

Orientador: Professor Doutor Miguel Neto
Co-orientador: Professor Doutor Roberto Henriques

Novembro 2012

RESUMO

Nos últimos anos, em consequência do aparecimento das redes sociais, a interacção entre o cliente e a empresa sofreu grandes alterações. Esta mudança, tal como outras, acarretou vantagens e desvantagens. Uma das maiores desvantagens que decorreu desta alteração é o facto de, actualmente, as organizações terem perdido o controlo sobre o que os clientes dizem acerca das mesmas, uma vez que estes facilmente publicam as suas opiniões negativas e estas são rapidamente propagadas. No entanto, algumas organizações rapidamente perceberam que poderiam retirar desta situação importantes vantagens competitivas, através da análise das opiniões que os clientes emitem sobre as mesmas, nos diversos canais.

Além disso, o crescente aumento da utilização da internet permitiu também que muita informação esteja disponível *online*, sendo exemplo disso o facto de, actualmente, a maioria dos jornais disponibilizarem diariamente as suas publicações, nos seus sítios, na internet. Consequentemente, o volume diário de dados disponíveis na internet cresce exponencialmente e toda a informação gerada através destes poderá ser relevante, se for tratada e utilizada correctamente. É, desta forma, que surge o desafio de gerar conhecimento através desta informação, de forma automatizada.

Assim, o objectivo deste trabalho consiste na construção de um modelo capaz de avaliar a polaridade (positiva, negativa ou neutra) de títulos de notícias de economia, disponíveis em endereços de *RSS Feeds*. Para a realização do mesmo, foi utilizado o *software* SAS e, por consequência, seguida toda a sua metodologia, cuja apresentação detalhada também constitui um objectivo.

PALAVRAS-CHAVE

Análise de Sentimentos, Análise de Opinião, *Text Mining*, Descoberta de Conhecimento em Textos, *Data Mining*, Descoberta de Conhecimento em Bases de Dados, Processamento de Linguagem Natural.

ABSTRACT

In the last few years, due to the emergence of social networks, the interaction between customers and companies has experienced major changes. This change, like others, has advantages but also disadvantages. One of the major disadvantages which arose from this modification is the fact that, currently, organizations have lost control over what customers say about them, since they can easily publish their negative opinions and spread them rapidly. However, some organizations have quickly realized this situation could promote important competitive advantages, through the analysis of what customers say about them in different communication channels.

Besides that, the increasing use of internet allowed that a lot of information is available online and an example of it is that, nowadays, the majority of newspapers make their publications daily available, on their websites, on the internet. Therefore, the data volume daily available on the internet grows exponentially and all of the information produced through this data might be important, if treated and used correctly. That is how the challenge of creating knowledge through this information in an automated way, emerges.

Thus, the goal of this project is to build a model able to evaluate the polarity (positive, negative or neutral) of economic news headlines, available on RSS Feeds addresses. In order to do that, software SAS was used and, consequently its methodology, whose detailed description is also a goal.

KEYWORDS

Sentiment Analysis, Opinion Analysis, Text Mining, Knowledge Discovery in Text, Data Mining, Knowledge Discovery in Databases, Natural language processing.

ÍNDICE

1. Introdução 1
   1.1. Objectivo do Estudo 1
   1.2. Motivação 2
   1.3. Organização do trabalho 4
2. Revisão da Literatura 5
   2.1. Fundamentos e Aplicações 5
      2.1.1. *Text Mining* 6
      2.1.2. Processamento de Linguagem Natural 7
      2.1.3. Aplicações 8
   2.2. Processo de Descoberta de Conhecimento em Textos 9
      2.2.1. Extracção 10
      2.2.2. Pré-processamento 11
      2.2.3. Indexação 15
      2.2.4. *Text Mining* 15
      2.2.5. Análise de Informação 16
   2.3. Análise de Sentimentos 16
      2.3.1. Terminologia 18
      2.3.2. Classificação de Sentimentos e Subjectividade 19
      2.3.3. Análise de Sentimentos baseada em Componentes 19
      2.3.4. Análise de Sentimentos através da comparação 19
      2.3.5. Classificação automática de documentos 20
3. Metodologia 29
   3.1. Etapas do Projecto 29
   3.2. Desenvolvimento do Projecto 30
      3.2.1. Selecção RSS Feeds 30
      3.2.2. Extracção das Notícias 31
      3.2.3. Armazenamento e Segmentação das Notícias 32
      3.2.4. Desenvolvimento dos Modelos 33
      3.2.5. Validação dos Modelos 37
4. Resultados e Discussão 38
   4.1. Selecção dos endereços de *RSS Feeds* 38

4.2. Segmentação de Notícias 39
4.3. Desenvolvimento dos Modelos 39
4.3.1. Modelo Estatístico 40
4.3.2. Modelo Baseado em Regras 42
4.3.3. Modelo Híbrido 45
4.4. Validação dos Modelos 46
5. Conclusões 48
6. Limitações e Recomendações para Trabalhos Futuros 50
7. Bibliografia 53
8. Anexos 56
8.1. Resultados obtidos para o modelo estatístico 56
8.2. Resultados para a fase de desenvolvimento do modelo BeR (negativas) 58
8.3. Resultados para a fase de desenvolvimento do modelo BeR (Positivas) 59
8.4. Resultados para a fase de desenvolvimento do modelo Híbrido (Negativas) 59
8.5. Resultados para a fase de desenvolvimento do modelo Híbrido (Positivas) 60
8.6. Estrutura de Regras implementada no modelo BeR 60

ÍNDICE DE FIGURAS

Figura 2.1 - Fases da análise de PLN 8
Figura 2.2 - Fases do processo de Descoberta de Conhecimento em Textos 9
Figura 2.3 - Criação do *training corpus* através da definição de regras 21
Figura 2.4 - Processo de concepção do classificador Naïve Bayes 23
Figura 3.1 - Etapas do projecto. 29
Figura 3.2 - Processo de Desenvolvimento do Projecto 30
Figura 3.3 - Arquitectura do *SAS Web Crawler* 31
Figura 3.4 - Arquitectura do *SAS Sentiment Analysis* 33
Figura 3.5 - Estrutura de Regras do Modelo BeR 36
Figura 4.1 – Exemplo de aplicação das regras a uma notícia 44

ÍNDICE DE TABELAS

Tabela 2.1 - Métodos de normalização de texto 25
Tabela 3.1 - Tabela com combinação de métodos utilizados pelo *SAS Sentiment Analysis* para o modelo estatístico 34
Tabela 4.1 - Endereços *RSS Feeds* 38
Tabela 4.2 - Informação das amostras seleccionadas para o treino e validação dos modelos 39
Tabela 4.3 - Indicadores de qualidade do modelo estatístico para as várias combinações de métodos 41
Tabela 4.4 – Exemplos de regras criadas para a classe de títulos positivos 42
Tabela 4.5 - Exemplos de regras criadas para a classe de títulos negativos 43
Tabela 4.6 - Resultados obtidos no modelo BeR 44
Tabela 4.7 - Indicadores de qualidade do modelo BeR 45
Tabela 4.8 - Testes efectuados para obter melhor combinação no modelo Híbrido 46
Tabela 4.9 - Tabela com os resultados da validação dos três modelos 46
Tabela 4.10 - Tabela com indicadores de qualidade dos modelos 47

LISTA DE SIGLAS E ABREVIATURAS

PSE	Produtos e Serviços de Estatística Lda.
TM	*Text Mining*
AS	Análise de Sentimentos
PLN	Processamento de Linguagem Natural
RSS	*Rich Site Summary*
DCBD	Descoberta de Conhecimento em Bases de Dados
DCT	Descoberta de Conhecimento em Textos
WWW	*World Wide Web*
BeR	Baseado em Regras

1. INTRODUÇÃO

Sentiment without action is the ruin of the soul.
Edward Abbey

O rápido crescimento da utilização da internet fez com que assistíssemos à aceleração da globalização e da aproximação dos povos. Este crescimento banalizou o uso da mesma e tornou a actualidade dependente dos serviços disponibilizados na *web*. Como consequência, grandes volumes de dados são gerados no dia-a-dia, seja através das redes sociais, *blogs* e fóruns, seja através dos meios de comunicação social na internet (Pang & Lee, 2008).

As organizações viram neste aumento uma grande oportunidade para ganhar vantagem competitiva, através da criação de conhecimento a partir destes dados. Para além disso, e dado o facto de a competitividade entre as organizações ter vindo a acentuar-se ao longo dos anos, este aumento fez com que a visão sobre o cliente fosse diferente, enfatizando a sua importância e dando, por sua vez, origem à fomentação das boas práticas e da consciencialização para a manutenção dos mesmos.

Assim, ter conhecimento do que os clientes pensam sobre a organização ou do que acontece diariamente na internet é um desafio que todas têm na luta pela sua sobrevivência no mercado. Tendo em conta que a tomada de decisão é um processo que resulta da prévia investigação e análise de dados, é essencial obter informação qualitativa que contenha alto valor acrescentado, de forma a criar diferenciação. Esta situação só é possível se existir uma monitorização constante da realidade, a qual terá sempre de ser feita recorrendo a processos automatizados, uma vez que a quantidade de dados gerados é, a cada segundo, gigantesca. É desta forma que o *Text Mining* (TM), em particular a Análise de Sentimentos (AS), ganhou nos últimos anos grande interesse, uma vez que permite de uma forma automatizada tratar e analisar grandes volumes de dados não estruturados e, daí, gerar conhecimento.

1.1. OBJECTIVO DO ESTUDO

O objectivo principal deste trabalho é apresentar um documento que proporcione às organizações portuguesas um exemplo de como se iniciarem na produção de conhecimento através de dados textuais, facilitando futuros projectos para a língua portuguesa. Para a concretização do mesmo, serão obrigatoriamente definidos outros objectivos mais concretos.

Assim, o segundo objectivo consiste na elaboração de uma revisão cuidada da literatura existente sobre o tema, de forma a mostrar quais os desenvolvimentos nesta área e apresentar uma metodologia considerada como pioneira para o mercado português.

O terceiro passará pela criação de um modelo de AS que avalie a polaridade (positiva, negativa ou neutra) de títulos de notícias presentes em endereços de *RSS Feeds* de economia. Esta classificação será feita de acordo com o reflexo que cada uma das notícias representa, do estado da mesma. Ou seja, pretende-se com este modelo obter uma classificação de notícias que, posteriormente a este projecto, torne possível monitorizar o estado da economia, através das notícias diárias. A estas será atribuída uma classificação positiva, negativa ou neutra, sendo possível saber consequentemente quantas notícias saem diariamente, para cada uma destas categorias.

Por fim, o último objectivo visa a utilização de um *software de* referência no mercado. A escolha da plataforma do *SAS* para *Text Analytics* surge para a concretização deste objectivo, uma vez que se trata de uma empresa bastante conceituada no mercado, relativamente aos seus produtos analíticos.

1.2. MOTIVAÇÃO

O desenvolvimento dos sistemas de informação tem permitido o armazenamento de grandes quantidades de dados e a automatização de tarefas que anteriormente tomavam muito tempo até estarem concluídas. Desta forma, análises de grande complexidade que envolvem enormes quantidades de dados passaram a ser possíveis de efectuar, ajudando no rápido crescimento de diversas áreas, nomeadamente da Descoberta de Conhecimento em Bases de Dados (DCBD).

Com a rápida expansão do comércio electrónico, mais produtos são vendidos na internet, aumentando o número de pessoas que compram produtos *online*. De maneira a aumentar a satisfação dos clientes e a experiência de compra, tornou-se uma prática comum os comerciantes pedirem aos clientes para expressarem opiniões sobre os produtos que compraram (Hu & Liu, 2004).

Actualmente, perante este ecossistema complexo onde a competição pelo cliente é global, a necessidade de saber como gerir as opiniões emitidas sobre os produtos e serviços, torna-se fundamental para a manutenção dos clientes, através da tomada de decisão em tempo útil. Assim, descobrir o que as pessoas pensam sobre determinado assunto, tem sido um importante motivo para o armazenamento de

informação. Com o aumento da disponibilidade e da popularidade em torno das fontes que armazenam opiniões, tais como *sites* de opinião de produtos e *blogs*, novas oportunidades e desafios têm aparecido na forma como as pessoas podem utilizar as tecnologias de informação para procurar e perceber a opinião dos outros (Pang & Lee, 2008).

Um estudo recente, que decorreu entre os meses de Janeiro e Fevereiro de 2011, elaborado pela PSE, revela que as empresas nacionais procuram cada vez mais saber a opinião dos consumidores sobre as mesmas (PSE - Produtos e serviços de Estatística, 2011). Neste estudo foram revelados alguns dados interessantes, dos quais destaco:

- 87,8% dos inquiridos referiu ser crucial esta capacidade de recepção, integração e análise imediata das respostas e interacções dos consumidores com a empresa;

- 82,8% das empresas que responderam já implementaram ou estão a implementar um sistema interno para ouvir a opinião dos seus clientes, 14,3 % está a pensar fazê-lo e só 2,9% não tem planos para tal;

- Das empresas que estão a pensar fazê-lo no futuro, 75% esperam fazê-lo no prazo de 1 ano.

Com o aumento deste interesse, a investigação em torno da AS tem vindo a crescer. Apesar de este ser o principal motivo para o crescimento do interesse nesta área, a AS poderá ser utilizada para mais situações, tal como é provado através deste projecto.

A acompanhar a situação anteriormente referida, têm surgido no mercado variadíssimas ferramentas, fruto do crescente interesse nesta temática. No entanto, estas têm sido essencialmente adaptadas à língua inglesa e, devido a esta situação, as organizações portuguesas começam a enfrentar um verdadeiro problema, uma vez que apercebem-se, cada vez mais, da real importância destas análises e deparam-se com a escassez de instrumentos que as ajudem.

Deste modo, a motivação para a elaboração deste trabalho surgiu da necessidade emergente de informação na área de AS para a língua portuguesa. Neste sentido e uma vez que não foi possível elaborá-lo no contexto empresarial, este permitirá gerar conhecimento nesta área para a língua portuguesa, testar o *software* do *SAS* para a mesma e ainda possibilitar a utilização do modelo em futuros projectos.

Assim, pretende-se desenvolver um caso de estudo para organizações que pretendam dar os primeiros passos nesta área. A escolha pelo *software* da SAS está relacionado com o facto de que a maioria das empresas utiliza-o nas suas tarefas

analíticas, tornando mais fácil a sua expansão para as tarefas de análise textual. Deste modo, conjugando as duas vertentes, o *software* e o conhecimento em AS para a língua portuguesa, será possível ajudar organizações ou indivíduos que pretendam utilizar estas técnicas.

1.3. ORGANIZAÇÃO DO TRABALHO

No capítulo 1 foi introduzido o tema para que o leitor perceba o enquadramento deste trabalho. Foram ainda apresentados os objectivos a concretizar com a realização do mesmo, bem como as motivações que levaram à escolha do tema.

O objectivo do capítulo seguinte consiste na discussão dos fundamentos que estão subjacentes à realização deste trabalho, com a apresentação de cada uma das áreas que está presente neste projecto e ainda das técnicas e métodos que serão utilizados ou referenciados.

No capítulo 3 será apresentada toda a metodologia que virá a ser utilizada, descrevendo pormenorizadamente todas as fases do projecto.

Relativamente ao capítulo 4, pretende-se apresentar e discutir os resultados obtidos. Por fim, no capítulo 5 serão apresentadas as limitações que ocorreram no desenvolvimento do projecto e sugestões de trabalho que poderá ser desenvolvido no futuro.

2. REVISÃO DA LITERATURA

O principal objectivo deste capítulo consiste na apresentação das técnicas e metodologias existentes actualmente na literatura e que serão utilizadas ou referenciadas no âmbito deste trabalho projecto.

2.1. FUNDAMENTOS E APLICAÇÕES

Nos últimos anos vários tipos de dados têm sido armazenados, fruto do desenvolvimento dos sistemas e tecnologias de informação, dos negócios e das bases de dados das organizações, variando também de acordo com a área de actuação e estrutura da organização (Konchady, 2006).

Por outro lado, a proliferação dos equipamentos digitais e a sua utilização como meio de comunicação continua a proporcionar um aumento da procura de sistemas e algoritmos capazes de descobrir conhecimento através de dados. Assim, o desenvolvimento de técnicas para descoberta de conhecimento em dados não estruturados, semi-estruturados e estruturados tem-se revelado bastante importante, quer para a indústria, quer para o meio académico (Berry & Kogan, 2010).

Neste sentido, devido ao facto do TM procurar extrair informação útil de dados não estruturados ou semi-estruturados e estes serem difíceis de tratar (Ronen Feldman, 2006), o seu desenvolvimento ocorreu mais tarde face ao *Data Mining* (DM), uma vez que este último é utilizado na análise de dados estruturados (Gharehchopogh, 2010). Assim, o TM é a extensão natural do DM, uma vez que se trata da extracção de padrões, informação útil ou conhecimento através de dados não estruturados (Inniss et al., 2006).

Para além das técnicas e metodologias do DM que são utilizadas no TM, outras áreas são igualmente importantes para a extracção de conhecimento de dados textuais. Exemplo disso é o Processamento de Linguagem Natural (PLN), uma vez que explora a maneira como os computadores podem ser utilizados para compreender e manipular linguagem natural (Chowdhury, 2003), proporcionando um pré-processamento de dados não estruturados e, ainda, a Extracção e Recuperação de Informação que visa a procura e extracção de informação textual presente em diversas fontes (Manning, Raghavan, & Schütze, 2008).

Apesar destas outras áreas fazerem parte integrante do processo de Descoberta de Conhecimento em Textos (DCT), o foco deste trabalho assenta na construção do modelo de AS, ou seja, dum modelo de classificação. Deste modo, a contextualização

que será feita para as restantes áreas será mais superficial do que a feita para os métodos e técnicas utilizadas na AS.

2.1.1. *Text Mining*

A DCBD, também conhecida como DM, é uma área que centra o seu trabalho na exploração automática e na descoberta de padrões interessantes em grandes quantidades de dados. A maioria do trabalho desenvolvido em DCBD tem-se concentrado nas bases de dados estruturadas (R Feldman et al., 1998).

No entanto, os dados não estruturados expressam uma quantidade vasta e rica de informação, ainda que este formato codifique a informação de uma forma difícil de decifrar automaticamente. Esta será, porventura, a razão para o trabalho desenvolvido nesta área ter aparecido mais tarde face aos dados estruturados (Hearst, 1999).

A área que lida com este tipo de dados designa-se de TM, também conhecido por *Text Data Mining* ou DCT. Este refere-se ao processo de extracção de padrões interessantes e não triviais a partir de documentos de texto não estruturados (Tan, 1999). Isto engloba tudo, desde Recuperação de Informação até à Classificação ou *Clustering* de documentos textuais (Kao & Poteet, 2010).

O TM aplica as mesmas funções analíticas do DM, mas, neste caso, para o domínio da informação textual, baseando-se em técnicas sofisticadas de análise textual que extraem informação a partir de documentos de texto. Entretanto, o DM contempla uma parte muito limitada dos dados contidos pelas empresas. Provavelmente, mais de 90% dos dados das organizações encontra-se em formato não estruturado, ou seja, em cartas dos clientes, *e-mails*, gravações de telefone, contratos, informação técnica, patentes, etc. (Dörre, Gerstl, & Seiffert, 1999).

Esta grande quantidade de dados não pode simplesmente ser utilizada pelos computadores, uma vez que estes os tratam como simples sequências de caracteres. Deste modo, diferentes métodos e algoritmos são necessários de forma a obter uma estruturação dos dados com o objectivo de melhorar o processo de extracção de padrões úteis. Como referido, o TM refere-se normalmente ao processo de extracção de informação e conhecimento a partir de dados textuais, no entanto, tem sido muitas vezes referido noutras áreas de investigação, devido à necessidade que este tem de recorrer a essas no seu desempenho. Ou seja, o TM, para além de ser associado aos métodos e técnicas utilizadas em *Machine Learning* e Estatística, em determinadas situações, pode ser também ligado à Extracção e Recuperação de informação, ao PLN,

ou ainda, ao processo que conjuga todas estas áreas (Hotho, Nürnberger, & Paaß, 2005).

2.1.2. Processamento de Linguagem Natural

Chama-se PLN ao conjunto de técnicas teórico-computacionais que analisam e representam dados textuais, com o objectivo de processar linguagem humana para vários tipos de tarefas e aplicações (Liddy, 2003).

Como referido anteriormente, o PLN desempenha um papel fundamental para o TM, uma vez que é utilizado na etapa de pré-processamento, possibilitando um primeiro nível de estruturação de dados. No entanto, importa referir que existem algumas técnicas dentro da área de PLN que não são aplicadas ao TM, tais como traduções automáticas de texto e correctores ortográficos (Junior, 2008).

Os investigadores que se dedicam a esta área de investigação tentam reunir conhecimento sobre a forma como os seres humanos utilizam a linguagem, de modo a que possam ser desenvolvidas ferramentas e técnicas apropriadas, possibilitando a compreensão e manipulação de línguas naturais por computadores (Chowdhury, 2003).

Tipicamente, o PLN utiliza conceitos linguísticos, como partes do discurso (substantivo, verbo, adjectivo, etc.) e estrutura gramatical. A linguagem natural é de grande complexidade, fazendo com que este processamento tenha de lidar com diversas situações complexas, tal como ambiguidades, tornando-o de extrema dificuldade, mas ao mesmo tempo de grande relevância. Para tal, o PLN faz uso de diversas representações de conhecimento, tais como léxico, semântica, propriedades e regras gramaticais e ainda outros recursos, como ontologias de entidades e acções ou dicionários de sinónimos e abreviaturas (Kao & Poteet, 2010).

Como se pode observar na figura seguinte, para o processo de PLN, existem vários níveis que precisam de ser entendidos e diferenciados. Nomeadamente, o morfológico que lida com tratamento das palavras, o léxico que se refere à análise do significado das palavras e de partes do diálogo, o sintáctico que trabalha a gramática e a estrutura das frases, o fonético que lida com a pronúncia, o semântico que traduz o significado das palavras e frases, o discurso que lida com a estrutura de diferentes tipos de texto e, por fim, o pragmático que introduz conhecimento presente nas pessoas (Feldman, 1999). Todos estes níveis poderão ser utilizados no processo de PLN, no entanto, para este projecto apenas serão utilizados os dois primeiros, uma vez

que, como referido anteriormente, é apenas necessário conceder um primeiro nível de estruturação aos dados.

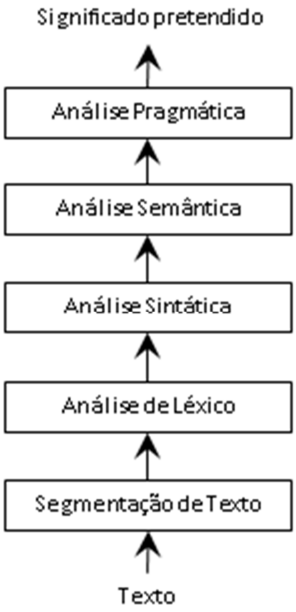


Figura 2.1 - Fases da análise de PLN
Fonte: Adaptado de Indurkhya & Damerau (2010)

As origens do PLN estão em inúmeras disciplinas, tais como engenharia informática, eléctrica e electrónica, matemática, línguas, inteligência artificial, ciência da comunicação, robótica, psicologia, etc. (Chowdhury, 2003).

2.1.3. Aplicações

Como anteriormente referido, as aplicações do PLN incluem um grande número de domínios de estudo. Entre estes estão a tradução automática de textos, o processamento de textos em linguagem natural e respectiva sumarização, a recuperação de informação em diferentes línguas, o reconhecimento de discurso e a inteligência artificial, entre outros (Chowdhury, 2003).

Muitas técnicas de PLN são utilizadas em diversas áreas, das quais destacam-se Reconhecimento de Entidades, Segmentação de palavras e frases, Remoção de *stopwords*, *Part-of-Speech Tagging*, Segmentação de Texto, Normalização de palavras (*stemming* e *lemmatization*) e Desambiguação (Indurkhya & Damerau, 2010).

Quanto ao processo de TM, este engloba um grande conjunto de aplicações que permite a exploração de informação escondida em dados textuais. Normalmente, é hábito dizer-se que o TM pode ser utilizado em qualquer sítio, bastando apenas existir um grande conjunto de dados textuais que necessitem de ser analisados. De todas as

aplicações possíveis de se realizar recorrendo a TM, destacam-se a extracção de informação e eventos, procura e recuperação de informação escondida em documentos de texto, sumarização, seguimento de tópicos, *clustering* (aprendizagem não supervisionada) e classificação (aprendizagem supervisionada) de documentos, visualização de informação, resposta a questões e ligação de conceitos (Fan, Wallace, Rich, & Zhang, 2006).

As principais aplicações são frequentemente associadas a diferentes tipos de sectores de actividade, dos quais destacam-se publicidade, telecomunicações, internet e tecnologias de informação, bancos, seguros, instituições políticas e empresas ligadas à investigação de fármacos (Gupta & Lehal, 2009).

2.2. Processo de Descoberta de Conhecimento em Textos

Neste subcapítulo serão apresentadas e discutidas todas as etapas do processo de DCT apresentadas em Aranha (2007). Esta metodologia é a que melhor se enquadra no âmbito deste projecto, uma vez que permite obter os resultados pretendidos e também reflecte a metodologia presente na solução utilizada.

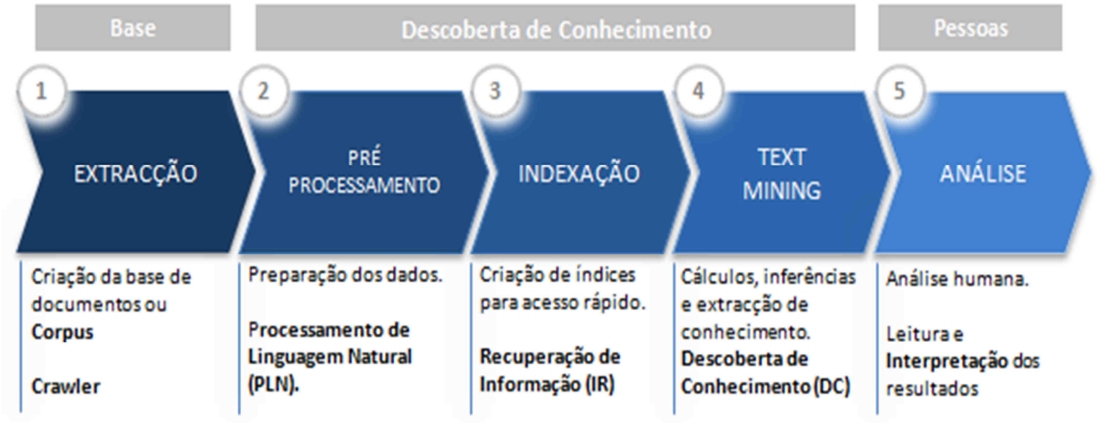


Figura 2.2 - Fases do processo de Descoberta de Conhecimento em Textos
Fonte: Aranha (2007)

Aranha (2007) sugere que o processo de DCT passa por cinco fases distintas, a primeira na aquisição dos dados textuais necessários para a análise, a segunda por um pré-processamento desses mesmos dados com o objectivo de lhes conferir um primeiro nível de estruturação, a terceira na criação de índices que possibilitem uma melhor recuperação ou procura destes, quando necessário, a quarta na extracção de conhecimento e por fim uma quinta fase para análise dos resultados obtidos.

É importante referir que, para problemas que utilizem abordagens de aprendizagem não supervisionada, ou seja de *clustering*, este processo é aplicado tal como é ilustrado na imagem anterior. No entanto, para problemas semelhantes ao deste trabalho, ou seja de classificação, uma vez que utilizam técnicas baseadas em aprendizagem supervisionada, este processo terá que ser percorrido várias vezes. A primeira para criação do modelo, a segunda para validação do mesmo e, por fim, as restantes para utilização de dados futuros. Para este trabalho, apenas serão percorridas as duas primeiras, uma vez que o objectivo é a criação do modelo.

Em seguida, serão detalhadas cada uma das fases acima descritas.

2.2.1. Extracção

Na DCT, tal como no DM, quando se está perante um problema como o apresentado neste projecto, ou seja, de classificação, existe sempre a necessidade de obter um conjunto de treino, uma vez que este tipo de problema recorre a técnicas estatísticas com aprendizagem supervisionada.

Assim, o processo de extracção de dados tem como função a recolha de dados que servirão como base do trabalho. Esta base pode ser recolhida apenas uma vez ou então ser constantemente alterada, substituindo os dados mais antigos pelos recentes ou simplesmente adicionar os novos aos antigos (Aranha, 2007).

Para a extracção dos textos que permitirão a criação do conjunto de treino são utilizados, em muitos casos, os chamados *crawlers*. Um *web crawler* é um programa que visita sítios da internet e que automaticamente extrai os dados destes (Jackson & Moulinier, 2002).

O objectivo do processo de *crawling* é, de uma forma rápida e eficiente para um conjunto de endereços *web* previamente disponibilizados, recolher automaticamente a informação presente nestes (Manning et al., 2008).

Depois de recolher o conjunto de textos pretendidos para a análise, será possível construir o conjunto de treino que servirá como base para as técnicas a aplicar no processo de DCT. Na DCT, chama-se ao conjunto de treino *training corpus*. Um *corpus* define-se com um conjunto de textos, possíveis de ser interpretados pelo computador, que representam uma ou um conjunto de linguagens naturais. A criação de *corpora* (plural de *corpus*) revela-se uma tarefa difícil uma vez que na maioria dos casos a sua criação implica processos manuais à base de *expert judgment* (Indurkhya & Damerau, 2010).

2.2.2. Pré-processamento

O pré-processamento dos dados trata-se de uma etapa muito importante, uma vez que proporciona uma primeira fase de estruturação dos mesmos. Para tal, são aplicadas algumas técnicas de PLN. Neste sentido, esta secção tem como objectivo descrever as principais técnicas necessárias para a elaboração deste projecto. Relembrar apenas que, para este projecto, apenas serão aplicadas algumas das muitas técnicas de PLN, pelo que só essas serão detalhadas nesta revisão.

2.2.2.1. Segmentação de Texto

A segmentação de texto é o processo que converte um texto livre em unidades únicas com significado. Esta revela-se extremamente importante, uma vez que permite obter uma estruturação dos dados, convertendo-os em dimensões (Jackson & Moulinier, 2002).

A segmentação de texto pode ser dividida em: 1) **segmentação de palavras** e 2) **segmentação de frases**. A primeira divide a sequência de caracteres achando as fronteiras de divisão das palavras, enquanto a segunda fá-lo localizando as fronteiras das frases. A segmentação de palavras é muitas vezes referida como *Tokenization,* que significa divisão de um conjunto de caracteres em *tokens.* Na maioria das línguas Europeias o delimitador de *tokens* é o espaço. No entanto, muitas línguas, como o Japonês e o Chinês, não utilizam o espaço como delimitador. Desta forma, a *tokenization* é dividida em duas abordagens: para as linguagens em que o espaço é o delimitador e para as que o delimitador de *tokens* não é o espaço. (Indurkhya & Damerau, 2010) Ainda assim, utilizar unicamente o espaço como delimitador não é suficiente, uma vez que não tomaria em conta sinais de pontuação, como pontos finais, de exclamação, etc. (Jackson & Moulinier, 2002).

Relativamente à segmentação de frases, esta procura dividir um texto em frases. Como a maioria das línguas usam pontuação, as fronteiras de divisão são normalmente sinais de pontuação. Ainda assim, existem tarefas auxiliares que permitem detectar a existência de pontuação que não significa um final de frase. Exemplo disso são as abreviaturas que frequentemente utilizam um ponto final. Desta forma, a segmentação de palavras e frases são duas tarefas que não podem acontecer em separado (Indurkhya & Damerau, 2010).

Pegando num exemplo, a frase "Amanhã vamos à praia!", pode ser dividida em cinco *tokens*. Ou seja, [Amanhã] trata-se do primeiro, [vamos] o segundo, [à] o terceiro, o quarto é [praia] e por fim, o quinto [!]. No caso da segmentação de frase, é

relativamente fácil de identificar que a fronteira da frase se trata do ponto de exclamação, ou seja, o último *token*.

O principal objectivo de um *tokenizer*, sendo ele de palavra ou de frase, é o de traduzir um texto em dimensões possíveis de analisar, obtendo um conjunto de dados estruturados (Jackson & Moulinier, 2002).

No fundo, a segmentação de texto consiste na forma como os computadores organizam os dados não estruturados, de modo a que possam utilizá-los como fonte de extracção de conhecimento. Obviamente que esta segmentação provoca um problema de dimensionalidade, uma vez que a divisão em *tokens* leva à criação de um grande número de dimensões para análise. Nas subsecções 2.2.2.4, 2.2.2.5 e 2.2.2.6 serão apresentadas algumas técnicas de redução de dimensionalidade.

2.2.2.2. Part-of-Speech Tagging

Trata-se de um técnica relativamente simples mas que permite categorizar cada *token* na respectiva categoria sintáctica. Essas categorias são verbos, nomes, adjectivos, advérbios, preposições, conjunções, pronomes e determinantes. Esta técnica apenas atribui uma categoria a cada termo. Isto é verdade para línguas Indo-Europeias, não acontecendo o mesmo para linguagens com estruturas morfológicas mais complexas (Indurkhya & Damerau, 2010).

A identificação da categoria sintáctica é encarada na literatura como um problema de classificação. Tal como em todos os problemas de classificação é atribuída a cada *token* uma probabilidade de pertencer a uma categoria específica. Para tal é necessário existir um conjunto de treino, para que se possa construir um modelo capaz de classificar o *token* com uma categoria sintáctica. Esta tarefa revela-se uma importante ajuda para o reconhecimento de entidades nomeadas (Junior, 2008).

2.2.2.3. Identificação de Entidades Nomeadas

A identificação de Entidades Nomeadas, como o próprio nome diz, é a identificação de diferentes tipos de nomes próprios, tais como nomes de pessoas, locais e organizações (Indurkhya & Damerau, 2010). Para esta identificação, torna-se importante que outras tarefas tenham anteriormente sido realizadas, tais como segmentação de frases e *POS tagging*. A utilização destas tarefas em conjunto com a busca de palavras cuja primeira letra está em maiúscula é uma importante ajuda no reconhecimento de entidades (Junior, 2008).

Num problema de classificação como o apresentado neste documento, a identificação de entidades presentes no texto acrescenta um enorme valor, uma vez que é através desta que é possível identificar a quem se refere determinado texto. A título de exemplo, é importante identificar que a palavra "BCE" se trata de uma entidade, uma vez que, caso não o seja feito, esta palavra não passa disso mesmo, enfrentando o risco de classificar erradamente notícias que contenham esta palavra a uma categoria específica (negativa ou positiva). Esta situação não seria de todo correcta, dado que as notícias que saem sobre esta entidade podem pertencer actualmente a uma categoria mas, terminada a crise, poderão pertencer a outra.

2.2.2.4. Remoção de Palavras Não Discriminantes (stop words)

Seguindo o que foi referido anteriormente, um dos grandes problemas das tarefas relacionadas com processamento de linguagem natural é o elevado número de dimensões de análise. Neste sentido, a redução de dimensionalidade é um factor de muita importância na diminuição do tempo de processamento, sendo a remoção de palavras não discriminantes, um exemplo disso.

Muitas palavras, normalmente as que aparecem com maior frequência nos textos, adicionam pouco valor à análise. Estas são frequentemente chamadas de palavras não discriminantes ou, em inglês, *stop words*. Exemplos destas são o "de", "para", "por", "que", "em", "no", etc. A criação destas listas de palavras ocorre normalmente através de tabelas de contingência que, depois, dão origem à sua remoção (Manning et al., 2008). Normalmente as palavras que constam nas listas de palavras não discriminantes são proposições, substantivos e determinantes. Nesta fase, os documentos são frequentemente pré-processados, removendo-as, seguindo-se a normalização para as palavras que restam (Jackson & Moulinier, 2002). Esta última fase será detalhada em seguida.

2.2.2.5. Normalização de palavras

Apesar de não ser utilizada qualquer técnica de normalização de dados no âmbito deste projecto, é importante apresentá-las uma vez que serão referidas a quando da apresentação das conclusões.

Manning et al. (2008) refere que existem diversas técnicas a aplicar de forma a normalizar os dados. Essas técnicas são aplicadas à normalização de:

Acentuação - Em inglês, a remoção de acentos das palavras não tem grande impacto, uma vez que ao fazê-lo, a compreensão da palavra mantém-se. No entanto,

para outras línguas essa remoção pode ser problemática porque a acentuação confere significados diferentes às palavras. O Espanhol é um exemplo de língua que removendo a acentuação poderá alterar o significado das palavras.

Letras Maiúsculas - É prática comum converter todas as letras maiúsculas em minúsculas. Esta conversão permite que palavras começadas por letra maiúscula não sejam compreendidas pelo *software* como palavras diferentes quando estão escritas completamente em minúsculas.

Stemming* e *Lemmatization - O objectivo das duas técnicas é precisamente o mesmo, ou seja, reduzir palavras que se encontram em formas derivadas para a sua forma base. Exemplo disso é a transformação das formas "comi", "comeram" para a sua forma base "comer". A diferença reside no facto do *stemming* se tratar de um processo heurístico que simplesmente corta as extremidades das palavras na tentativa de alcançar na maioria das vezes o objectivo pretendido. Quanto à *Lemmatization*, este procura atingir o objectivo com o uso de vocabulário e análise morfológica das palavras.

A utilização destas técnicas torna-se muito relevante na criação de modelos baseados em regras, que serão explicados mais pormenorizadamente no decorrer do trabalho. Se os dados forem correctamente normalizados evitam-se a duplicação de regras. Utilizando o exemplo anterior, se não existisse a normalização para a forma base “comer” teriam de ser criadas regras para cada um dos tempos verbais.

No caso do *software* utilizado neste trabalho, esta normalização ocorre de uma maneira diferente. O algoritmo utilizado faz a **expansão morfológica das palavras**. Ou seja, uma vez que esta normalização será bastante útil para o modelo Baseado em Regras (BeR), estas regras poderão recorrer à esta expansão, de forma a captarem um maior número de situações. Na prática, trata-se de uma normalização que ocorre na fase de aplicação das regras e que permite criar regras que possam captar todas as formas de uma palavra, sejam verbais, sejam nominais.

2.2.2.6. Selecção de Características

A selecção de características é uma importante técnica para a redução de dimensionalidade, uma vez que esta procura seleccionar um conjunto de dimensões que melhor descrevem o problema, excluindo as restantes.

Junior (2008) apresenta um conjunto de métricas que definem o quão importante é cada *token* para o léxico: *information Gain, Chi Square* e Frequência de

documentos. No entanto, o autor da metodologia adoptada neste trabalho para a classificação de notícias, sugere que na selecção de características um termo é seleccionado ou não sem qualquer tentativa de utilizar o respectivo grau de importância, que poderá ser calculado de várias formas. Ou seja, todas as palavras seleccionadas por este método são consideradas igualmente importantes. Para tal, são utilizadas as métricas referidas anteriormente, mas neste caso para considerar a relevância de cada termo, ao contrário da selecção de características que apenas exclui as menos relevantes ou considera as mais importantes, tornando-se um método de selecção binário (Kim, Han, Rim, & Myaeng, 2006). Assim, o autor sugere um método que considera a situação descrita anteriormente. Este método, bem como os métodos de normalização de texto, serão discutidos mais à frente quando for abordada a metodologia de classificação adoptada neste trabalho.

2.2.3. Indexação

As técnicas utilizadas para a criação de índices são amplamente referenciadas na área de Recuperação de Informação. Estas são muito úteis, uma vez que o volume de informação textual aumentou exponencialmente. Permitem que a procura/recuperação de informação seja efectuada de uma forma mais rápida e eficaz. Ficaram muito conhecidas depois do aparecimento do *Google*, que recorre a estas técnicas para encontrar informação presente na internet (Aranha, 2007).

Esta etapa tem uma grande relevância no processo de descoberta de conhecimento, uma vez que criar índices para os documentos que foram recolhidos na primeira etapa dá um grande contributo no momento em que se pretenda encontrar tais documentos. Para este trabalho esta fase não será importante, uma vez que o objectivo é criar um modelo de AS. Assim, esta etapa não terá um detalhe tão aprofundado como as restantes.

2.2.4. *Text Mining*

Após estarem concluídas as etapas que permitem a criação e o pré-processamento dos dados que servirão como base à análise que se pretende realizar, é tempo de definir quais os métodos a aplicar.

A decisão de qual a abordagem a seguir deverá ter em conta qual o objectivo do estudo. Tal como no DM, se o objectivo passa apenas por obter o relacionamento entre documentos, optando por uma técnica de aprendizagem não supervisionada, a classe de técnicas a utilizar deverá ser *clustering.* No caso de se desejar obter um

modelo que seja capaz classificar novos documentos com base noutros, então a escolha deverá ser pela classe de técnicas de métodos com aprendizagem supervisionados, ou seja, algoritmos de classificação (Junior, 2008).

Para este trabalho será utilizado um método de classificação, uma vez que se pretende obter um modelo de AS capaz de classificar notícias, que será detalhado no capítulo seguinte. Para este, importa distinguir duas fases: a primeira que envolve a criação do modelo propriamente dito e a segunda que implica a utilização deste para dados futuros. No âmbito deste modelo está apenas a elaboração da primeira, no entanto, depois de terminado este projecto, será possível de uma forma simples aplicar o modelo aqui criado e assim ter a segunda fase também finalizada.

2.2.5. Análise de Informação

Esta é a etapa de avaliação dos resultados obtidos, tanto para a qualidade revelada das técnicas e dos algoritmos utilizados, como para o conhecimento extraído. As duas não podem estar completamente separadas, uma vez que o conhecimento extraído dos dados só poderá ser considerado válido se as avaliações feitas à qualidade das técnicas e algoritmos aplicados, para atingir o objectivo pretendido, se revelarem eficazes (Aranha, 2007).

Assim, na primeira importa avaliar os indicadores que possibilitam verificar se as técnicas e os algoritmos utilizados são suficientemente bons para a causa, através das respectivas medidas de qualidade. Para a segunda, importa avaliar se os objectivos propostos foram atingidos e, através do conhecimento obtido, agir em conformidade.

2.3. ANÁLISE DE SENTIMENTOS

Actualmente, a *WWW* contém uma vasta quantidade de documentos que expressam opiniões e que contêm observações, comentários, críticas, etc. Estes, têm informação muito valiosa para ajudar as pessoas na sua tomada de decisão, como por exemplo, análises de produtos que podem ajudar as organizações a promovê-los melhor; comentários sobre política que podem ajudar os políticos a clarificarem a sua estratégia; ou mesmo críticas a um evento que pode ajudar as partes envolvidas a reflectir sobre as suas actividades. Contudo, a quantidade deste tipo de documentos é enorme e são, geralmente, expressos em linguagem natural, impossibilitando a leitura e análise pelos seres humanos. Assim, o trabalho que ajuda a determinar automaticamente a direcção do sentimento (positivo ou negativo) em textos é frequentemente designado de AS ou *Opinion Mining* (Li & Liu, 2010).

Embora a área de AS e do *Opinion Mining* tenham tido recentemente uma grande explosão no que diz respeito à actividade de investigação, tem existido um constante interesse há já algum tempo. O ano de 2001 marca o início de uma maior sensibilização dos problemas de investigação e oportunidades que estes trazem (Pang & Lee, 2008).

Assim, a AS ou *Opinion Mining* é o estudo computacional de opiniões, sentimentos e emoções expressadas em texto. A informação textual pode ser classificada em dois tipos principais: factos e opiniões. Os factos são expressões objectivas sobre entidades, eventos e as suas propriedades. As opiniões são geralmente expressões que descrevem os sentimentos e avaliações das pessoas em relação a determinadas entidades, eventos e suas respectivas propriedades (Liu, 2010).

Apesar de grande parte da literatura apresentar a AS como o estudo computacional de sentimentos, esta pode ser utilizada para muitos outros casos, tal como mostra este projecto. É importante referir que a AS se trata de um problema de classificação e que, como tal, pode ser utilizada para classificar informação textual de acordo com a sua polaridade, independentemente da frase denotar algum sentimento ou não. A frase "*Taxa Euribor mantém forte queda*" apenas descreve um facto, no entanto poderá ser classificada como positiva ou negativa para a economia.

Para que sejam entendidos os conceitos que estão adjacentes a esta área, seguir-se-á um exemplo que mostra algumas frases típicas normalmente utilizadas na literatura para apresentar a terminologia da AS.

" (1) Comprei umas sapatilhas há uns dias atrás. (2) Eram fantásticas. (3) Tinham um amortecimento e flexibilidade nunca vistos. (4) Melhores que as minhas anteriores. (5) No entanto, aqueciam-me demasiado os pés. (6) A minha mãe reclamou comigo porque não lhe disse antes de as comprar. (7) Obrigou-me a devolve-las porque eram demasiado caros. (8) Disse-lhe que os meus outros sapatos estragaram-se em dois dias"

Como é possível verificar no exemplo anterior, existem muitas opiniões dadas numa única análise. As frases (2) e (3) expressam opiniões favoráveis, no entanto a (2) revela uma opinião geral, enquanto a (3) refere-se a determinadas características do produto. A frase (5) expõe uma opinião negativa em relação ao tecido do produto, uma vez que este aquece demasiado. As frases (6) e (7) referem-se a opiniões expressas pela "mãe".

2.3.1. Terminologia

Liu, Hu, & Cheng (2005) referem-se a um conjunto de termos utilizados na AS e que são importantes apresentar em seguida.

Objecto – Trata-se do alvo de análise. Pode referir-se a uma entidade, produto, serviço ou pessoa. No exemplo dado, o objecto trata-se das sapatilhas.

Componente – Refere-se às características do objecto. Ou seja, uma opinião pode ser dada sobre um determinado produto, mas ao mesmo tempo sobre uma característica do mesmo. A frase (3) expõe uma opinião em relação a algumas características das sapatilhas.

Componentes implícitas ou explícitas – As componentes podem ser explícitas ou implícitas. Na frase (3) podemos verificar que as componentes "amortecimento" e "flexibilidade" são explícitas, uma vez que estão presentes na frase. Na (5), a componente é implícita uma vez que não se encontra na frase. O facto de referir que aqueciam demasiado os pés não especifica que o tecido das sapatilhas era quente.

Opinião – A expressão, atitude ou emoção emitida por alguma entidade.

Titular – Trata-se da entidade que expressa a opinião. Na frase (3) o titular da opinião é o sujeito que comprou o produto e na (7), a opinião pertence à "mãe".

Polaridade/Orientação – Determina se a opinião é negativa, positiva ou neutra.

Modelo de opinião - As opiniões podem ser directas ou por comparação. A frase (4) emite uma opinião por comparação, uma vez que compara as sapatilhas com as anteriores. A frase (3) possui uma opinião directa uma vez que se refere claramente às componentes sem as comparar com outras.

Subjectividade da frase – Uma frase pode ser objectiva ou subjectiva. As objectivas expressam factos reais (1). As subjectivas manifestam crenças e sentimentos pessoais (3).

Opinião implícita ou explícita – A frase (8) tem implícita uma opinião negativa embora ela não esteja presente. A frase (3) contém uma opinião explícita, uma vez que ela está directamente presente.

Tipos de avaliação de produto/serviço/entidade - Geralmente, para um dado produto ou serviço, existem frequentemente dois tipos de avaliação na Web. O primeiro é escrito num estilo semi-estruturado. O *Epionions.com* e *cnet.com* utilizam este formato. O segundo tipo é escrito num formato livre e não há separação dos prós e dos contras, podendo o cliente opinar livremente. O amazon.com usa este formato.

2.3.2. Classificação de Sentimentos e Subjectividade

Esta tem sido a área mais investigada, tratando a AS como um problema de classificação de texto. Dois subtemas têm sido amplamente estudados: (i) classificar as opiniões expressas num documento como negativas ou positivas e, (ii) classificar uma frase como objectiva ou subjectiva e, para as frases subjectivas, classificar as opiniões como negativas, positivas ou neutras. O primeiro subtema é vulgarmente conhecido como *sentiment classification* ou *document-level sentiment classification*. O segundo procura determinar em frases individuais se a mesma expressa uma opinião ou não (frequentemente chamada de *subjectivity classification*) e, no caso de expressar, verificar se essa opinião é ou não positiva (*sentence-level sentiment classification*) (Liu, 2010).

2.3.3. Análise de Sentimentos baseada em Componentes

Embora seja importante perceber se um determinado texto contém opiniões positivas ou negativas, em determinadas situações isso não chega. Em muitas condições é relevante perceber quais as componentes ou objectos sobre as quais foram expressas opiniões.

Este modelo foca-se em primeiro lugar no alvo das opiniões e posteriormente, em determinar qual a orientação das mesmas. Os alvos são os objectos e as suas componentes. Um objecto pode ser um produto/serviço, uma organização ou um evento (Liu, 2010).

Através deste modelo, consegue-se distinguir quais os objectos ou componentes que foram alvo de opiniões, contrariamente à classificação, que apenas fornece informação geral sobre uma frase ou documento.

2.3.4. Análise de Sentimentos através da comparação

Como referido anteriormente, o modelo de opinião pode ser directo ou por comparação. Esta é uma área bastante importante, uma vez que é frequente serem dadas opiniões, com base noutros objectos semelhantes. Em geral, as frases com comparações expressam uma relação baseada nas similaridades e diferenças de mais que um objecto. A comparação é normalmente transmitida usando a forma comparativa ou superlativa dum adjectivo ou advérbio. A comparativa é usada para evidenciar que um objecto é melhor do que outro. Por outro lado, a superlativa é utilizada para determinar se um objecto é o melhor ou pior (Liu, 2010).

2.3.5. Classificação automática de documentos

Recentemente, tem-se assistido a um crescente interesse na identificação e extracção automática de atitudes, opiniões e sentimentos em textos. Esta motivação aparece fruto da necessidade de fornecimento de ferramentas para os analistas de informação controlarem de uma forma automática sentimentos que são depositados nos mais variados locais da internet, quer sejam notícias, quer sejam fóruns, *blogs* ou redes sociais (Wiebe & Riloff, 2005).

A classificação automática refere-se ao processo no qual um classificador determina qual a classe a que um documento pertence. O objectivo principal da classificação é assignar uma determinada classe a um conjunto de documentos (Prabowo & Thelwall, 2009). No caso da AS, trata-se de assignar automaticamente um conjunto de documentos às classes positivas e negativas.

No âmbito deste trabalho serão utilizados três classificadores, um BeR, um outro estatístico e, por fim, um híbrido.

Para que seja possível construir o classificador híbrido, será necessário desenvolver o seguinte conjunto de actividades:

- Criação de regras para definir um conjunto de documentos para o treino do modelo estatístico;
- Criação do modelo estatístico;
- Construção de um modelo que verá as regras definidas no primeiro ponto implementadas;
- Criação do modelo híbrido através da combinação dos dois modelos referidos nos pontos anteriores.

2.3.5.1. Modelo Baseado em Regras

O modelo BeR torna-se importante porque não necessita de dados para a criação do classificador, necessitando apenas de um conjunto de regras que permitam classificar os documentos (Wiebe & Riloff, 2005), ou seja, poderão ser criadas por alguém que domine o assunto que se pretende tratar. Estas regras poderão ser importantes, tanto para a construção dos *corpora* a utilizar no treino e validação do modelo estatístico, como para a construção do modelo BeR.

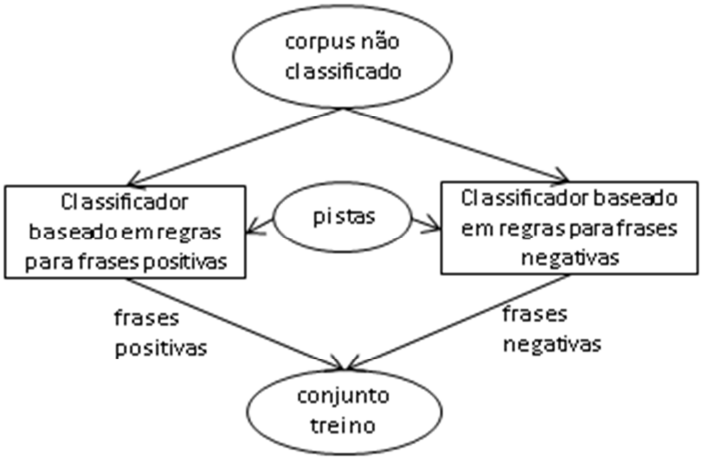


Figura 2.3 - Criação do *training corpus* através da definição de regras
Fonte: Adaptado de Wiebe & Riloff (2005)

A imagem anterior reflecte a forma de criação do *training corpus,* tal como do *testing corpus*, este último será muito importante na fase de validação dos modelos.

Relativamente à tarefa de criação das regras, esta deverá contar com a possibilidade de efectuar consultas à base de dados, de maneira a que o analista "aprenda" com os dados e, consequentemente, possa ser incorporado conhecimento no modelo que apenas pode ser fornecido por pessoas.

Quanto às regras, Prabowo & Thelwall (2009) referem que uma regra consiste num antecedente e o seu respectivo consequente, ou seja, uma relação normalmente chamada de *if-then.* Relativamente ao antecedente, este define a condição e trata-se de um *token* ou de um conjunto de *tokens* que dão origem ao respectivo consequente. O consequente trata-se da classe a que o antecedente dá origem. O autor referido anteriormente apresenta alguns exemplos que mostram como uma regra poderá ser construída e que serão apresentados de seguida.

$$\{token1 \wedge token2 \wedge \ldots \wedge tokenn\} \rightarrow \{+|-\}$$

Os dois exemplos que se seguem mostram duas regras, apenas com a definição de uma palavra para cada uma das regras, em que o consequente "+" significa classe dos positivos, e o "-" classe dos negativos.

$$\{excelente\} \rightarrow \{+\}$$
$$\{absurdo\} \rightarrow \{-\}$$

Um *token* poderá ser visto tanto como uma palavra, como um "?" que representa um substantivo, ou como um "#" que caracteriza um termo alvo. Usando a frase "O computador A é mais caro que o Computador B", o alvo trata-se do "Computador A". A regra que deriva desta frase é a seguinte:

{# ^ mais caro ^ do que ^ ?} → {−}

Esta regra explicita que a palavra alvo é menos favorável que o outro computador devido ao seu preço. O foco desta frase é definitivamente o atributo "preço".

No entanto, muitos outros tipos de regras poderão ser criados recorrendo a outro tipo de operadores booleanos, à expansão morfológica, à ordem e distância das palavras, etc.

A criação de classificadores baseados em regras é extremamente relevante, seja para a construção de conjuntos de treino, seja para a criação de classificadores que permitam em conjunto com classificadores estatísticos criar classificadores híbridos. Com estes será possível adicionar conhecimento presente nas pessoas e que de outra forma não estaria presente.

2.3.5.2. Modelo Estatístico

Quanto aos classificadores estatísticos, Pang, Lee & Vaithyanathan (2002) referem três métodos bastante utilizados para a classificação de textos, *Naïve Bayes*, *Maximum Entropy* e *Support Vector Machines*, no entanto, para este projecto apenas será utilizado o primeiro. A imagem seguinte mostra o processo de aprendizagem e classificação dos documentos com base naquilo que foi dito anteriormente.

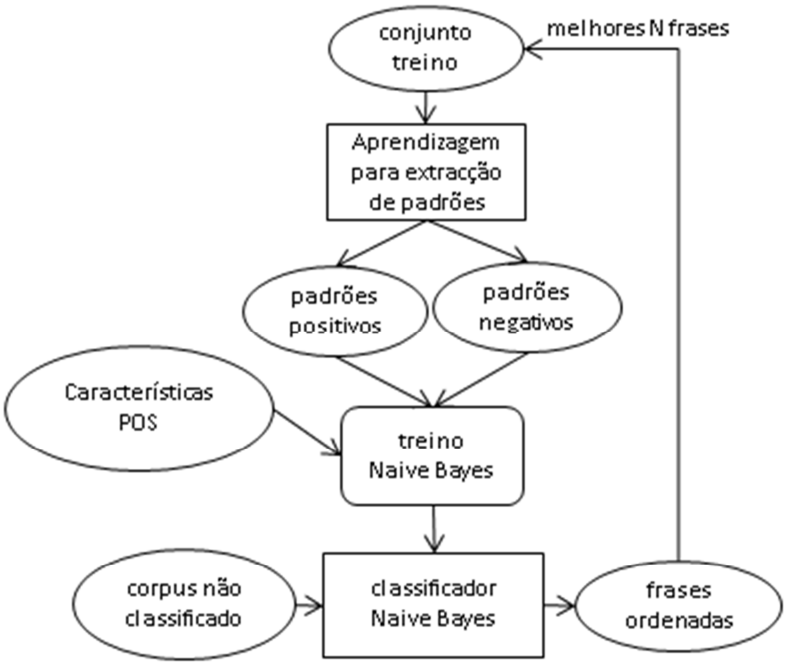


Figura 2.4 - Processo de concepção do classificador Naïve Bayes
Fonte: Adaptado de Wiebe & Riloff (2005)

Na fase da classificação, neste caso na criação do modelo de AS, é frequentemente utilizado o classificador de *Bayes*, pela facilidade de implementação e também pelos resultados positivos que se obtém nas tarefas de classificação.

Como classificador, é considerado um dos mais eficientes em questões relacionadas com processamento e precisão na classificação de novas amostras. *Thomas Bayes* desenvolveu-o em meados do século XVIII e é normalmente chamado de fórmula de probabilidade condicional de um determinado evento, muito utilizado em *Machine Learning* (Junior, 2008).

Neste trabalho será utilizada uma abordagem que normaliza o cumprimento por documento através da introdução de um modelo multivariado de *Poisson* para o classificador de texto *Naïve Bayes* em conjunto com um método de cálculo de pesos que pretende melhorar a performance das palavras onde os parâmetros dos modelos não são credíveis devido à baixa frequência dessas palavras, propostos em Kim et al. (2006).

No contexto de classificação, a probabilidade de d_j pertencer a uma classe c é calculado pelo teorema de *Bayes* da seguinte forma:

$$P(c|d_j) = \frac{P(d_j|c)\mathrm{p(c)}}{\mathrm{p}(d_j)} = \frac{P(d_j|c)\mathrm{p(c)}}{P(d_j|c)\mathrm{p(c)} + P(d_j|\bar{c})\mathrm{p}(\bar{\mathrm{c}})} = \frac{\frac{P(d_j|c)}{P(d_j|\bar{c})} \cdot \mathrm{p(c)}}{\frac{P(d_j|c)}{P(d_j|\bar{c})} \cdot \mathrm{p(c)} + \mathrm{p}(\bar{\mathrm{c}})}$$

Se for definida uma nova função $\mathrm{z_{jc}}$, tem-se:

$$\mathrm{z_{jc}} = \log \frac{P(d_j|c)}{P(d_j|\bar{c})}$$

Assim, pode-se reescrevê-la como:

$$P(c|dj) = \frac{\mathrm{e}^{\mathrm{z_{jc}}}.\mathrm{p(c)}}{\mathrm{e}^{\mathrm{z_{jc}}}.\mathrm{p(c)} + \mathrm{p}(\bar{\mathrm{c}})}$$

onde, d_j é um documento com texto, p(c) é a probabilidade de pertencer à classe positiva e $p(\bar{c})$ é a probabilidade de pertencer à classe negativa.

Ambas as probabilidades são calculadas durante o processo de treino do modelo. Por exemplo, se o conjunto de treino tiver um número de documentos positivos igual ao número de negativos então as probabilidades são 0.5 para cada uma.

A abordagem utilizada neste projecto utiliza técnicas para normalização do comprimento dos documentos, bem como métodos para considerar a importância de cada termo através dos seus pesos. Desta forma, o processo de classificação será dividido em etapas, de acordo com o referido e que serão descritas em seguida.

I. Normalização de texto

O objectivo destes métodos é obter a frequência normalizada de todos os *tokens* para cada documento. Primeiro são normalizadas as frequências dos termos em cada documento de acordo com o comprimento mesmo. Posteriormente, as frequências normalizadas serão combinadas linearmente de acordo com a probabilidade de cada documento pertencer a cada classe.

Os métodos propostos pelo autor são apresentados na tabela seguinte.

Método	$\hat{f}_{ij}$
Relative Frequency (RF)	$\frac{tf_{ij}}{(\alpha\ .\ avdl)\ +\ (1\ -\ \alpha)\ .\ dl_j}$
Smoothed Relative Frequency (SRF)	$\frac{tf_{ij}\ +\ \theta}{((\alpha\ .\ avdl)\ +\ (1\ -\ \alpha)\ .\ dl_j)\ +\ (\theta.\lvert T\rvert)}$
Okapi BM25 (BM25)	$\frac{(k_1+1)\ .\ tf_{ij}}{tf_{ij}\ +\ k_1\ .\ \frac{(\alpha\ +\ (1\ -\ \alpha)\ .\ dl_j)}{avdl}}$
Pivoted Length Normalization	$\frac{\frac{1\ +\ ln\ tf_{ij}}{1\ +\ ln\ avtf}}{(1\ -\ \alpha)\ .avdu\ +\ \alpha\ .\ du_j}$

Tabela 2.1 - Métodos de normalização de texto
Fonte: Kim et al. (2006)

Onde, tf_{ij} é a frequência to *token i* no documento *j*, α é um parâmetro (ex. 0,2), $avdl$ é a média do cumprimento dos documentos expresso em número de *tokens*, dl_j é o comprimento do documento *j*, expresso em número de *tokens*, θ é outro parâmetro (ex. 0,0001), $|T|$ é o número total de total de *tokens* no conjunto de treino, k_1 é outro parâmetro (ex. 1,2), $avtf$ é a média da frequência de todos os *tokens*, $avdu$ é a média do número de *tokens* distintos num documento, du_j é a média do número de *tokens* distintos no documento *j.*

No final será obtido e $\hat{f}_{ij}$, que se trata da frequência normalizada do *token i* no documento *j.*

Depois de calculado $\hat{f}_{ij}$ para cada um dos métodos, é calculada probabilidade de cada documento pertencer a cada classe da seguinte forma:

$$\lambda_{ic} = \frac{1}{|D_c|} \cdot \sum_{j=1}^{|D_c|} \hat{f}_{ij} \quad \text{e} \quad \mu_{ic} = \frac{1}{|D_{\bar{c}}|} \cdot \sum_{j=1}^{|D_{\bar{c}}|} \hat{f}_{ij}$$

onde, $|D_c|$ é o número total de documentos positivos no conjunto de treino e $|D_{\bar{c}}|$ é o número total de documentos negativos no conjunto de treino

Relativamente a estes métodos, enquanto a RF e a SRF só normalizam a frequência do termo dividindo pelo factor de normalização, as outras duas transformam a frequência do termo baseado em algumas razões teóricas ou empíricas, com resultados demonstrados em trabalhos realizados. A fórmula do BM25 deriva do modelo probabilístico de *Poisson* e a *Pivoted Length Normalization* resulta da investigação da relação entre o cumprimento do documento e a importância deste para várias categorias de textos.

II. Cálculo dos pesos de cada token

Como referido anteriormente, na abordagem utilizada neste trabalho, em vez de serem utilizadas técnicas de selecção de carecterísticas (*tokens*), excluindo as características que se revelem insignificantes, são utilizadas técnicas para atribuíção de pesos a cada *token* de acordo com a sua importância.

O objectivo é obter um peso para cada *token* de acordo com a sua importância para a classificação em determinada classe. Para tal, serão utilizados os seguintes métodos:

Chi Square - Esta métrica mede a falta de independência entre o termo *i* e classe *c* e pode ser comparada à distribuição χ2 com um grau de liberdade para julgar extremos. Usando uma tabela de contingências de um termo *i* e uma categoria *c*, onde

w é o número de vezes que *i* e *c* co-ocorrem, *y* é o número de vezes que *i* ocorre sem *c*, *x* é o número de vezes que *c* ocorre sem *i*, *z* é o número de vezes que nem *i* nem *c* ocorrem.

$$fw_{ic} = \frac{(wz - xy)^2}{(w+x)(w+y)(x+z)(y+z)}$$

Risk Ratio - O *Risk Ratio* é uma métrica simples mas bastante utilizada na investigação biomédica. A sua fórmula é a seguinte:

$$fw_{ic} = \frac{\lambda_{ic}}{\mu_{ic}} + \frac{\mu_{ic}}{\lambda_{ic}}$$

Information Gain - Define a importância de determinado termo para a discriminação entre classes de documentos previamente conhecidos, verificando o nível de correlação existente com cada classe. A fórmula é a seguinte:

$$fw_{ic} = H(C) - H(C|W_i) = \sum_{c_s \in \{c,\bar{c}\}} \sum_{w_t \in \{w_i,\bar{w}_i\}} p(c_s, w_s) \log \frac{p(c_s, w_t)}{p(c_s)p(w_t)}$$

onde, por exemplo, *p(c)* é o número de documentos que pertencem à classe *c* dividido pelo número de documentos e $p(\overline{w})$ é o número de documentos sem o termo *w* dividido pelo número total de documentos, etc.

III. Cálculo de z_{jc}

Este modelo, que utiliza o peso das características (*tokens*), pode ser visto como uma versão generalizada do modelo que utiliza a selecção das características. Ou seja, se forem atribuídos um ou zero a todas as palavras seleccionadas ou não seleccionadas, respectivamente, então o classificador de texto acaba por seleccionar um subconjunto de palavras.

Para minimizar o impacto de utilizar este método binário de classificação de palavras, o método sugerido pelo autor e que é utilizado no *software* que suporta este trabalho, calcula o peso de cada palavra, em vez de as seleccionar, da seguinte forma:

$$z_{jc} = \sum_{i=1}^{|V|} \frac{fw_{ic}}{FW_c} . f_{ij} . log \frac{\lambda_i}{\mu_i}$$

onde, |V| é o número total de *tokens* no documento d_j, f_{ij} é a frequência do *token i* no documento d_j e fw_i é o peso do *token i*.

Durante o treino, os quatro esquemas de cálculo de pesos (fw_{ic}) são avaliados e o que melhor se comportar é o escolhido.

FW é o factor de normalização, e é calculado da seguinte forma:

$$FW_c = \sum_{i=1}^{V} fw_{ic}$$

onde, V é o número total de *tokens* na classe c.

2.3.5.1. Modelo Híbrido

Por fim surge o terceiro classificador de texto, que recorre à combinação dos dois classificadores anteriormente explicados, ou seja, um classificador híbrido. Trabalhos realizados anteriormente mostram que os resultados obtidos através de abordagens híbridas revelam-se melhores em tarefas de classificação (Jiang, 2006).

No caso deste trabalho, o modelo híbrido é construído recorrendo ao cálculo da probabilidade de cada documento pertencer a uma determinada classe, através da combinação linear dos classificadores estatístico e BeR. Na fase de elaboração do modelo híbrido, a percentagem a atribuir a cada um dos classificadores é medida

através de *test and learn*, de forma a obter a combinação que melhores resultados produz. Estes serão medidos através de métricas de precisão e de erro do modelo.

Este classificador, uma vez que se trata de uma situação muito específica deste projecto, será detalhado mais à frente quando for apresentada a metodologia do trabalho.

3. METODOLOGIA

A elaboração deste trabalho propõe, como referido anteriormente, desenvolver um modelo que avalie a polaridade dos títulos de notícias de economia disponíveis em *RSS feeds*, através de um modelo de AS. O planeamento de todo o processo revela-se de uma importância extrema, uma vez que permite definir com rigor todas as etapas do projecto, bem como o seu o cumprimento no tempo disponível. Deste modo, nesta secção serão apresentadas todas as suas etapas, assim como toda a metodologia adjacente ao desenvolvimento do mesmo.

3.1. ETAPAS DO PROJECTO

O trabalho de projecto aqui descrito estará enquadrado em quatro fases, como ilustra a seguinte figura.

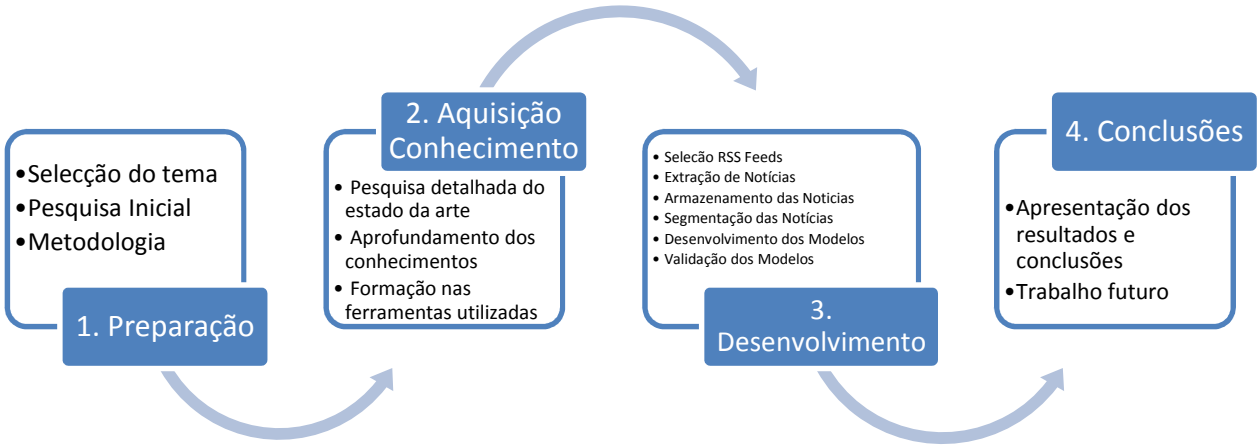


Figura 3.1 - Etapas do projecto.

A primeira fase correspondeu ao início da pesquisa, permitindo definir o tema de trabalho, bem como identificar a metodologia adjacente, a qual será explicada em detalhe neste capítulo. Por sua vez, a segunda fase será constante, ou seja, decorrerá ao longo de todo o projecto, uma vez que é essencial acompanhar os desenvolvimentos em torno desta área e adquirir conhecimentos nas ferramentas que serão utilizadas. Ainda assim e antes do início da mesma, foi elaborada uma profunda revisão da literatura de forma a aprofundar os conhecimentos, bem como apresentar o estado da arte.

A terceira fase será a etapa mais importante do projecto, uma vez que é nesta que assenta a base de todo o trabalho e que será descrita em detalhe no seguinte subcapítulo. Na última fase do projecto serão documentados todos os trabalhos

realizados na etapa de desenvolvimento, bem como apresentados os resultados e respectivas conclusões.

3.2. DESENVOLVIMENTO DO PROJECTO

Como já foi referido anteriormente, este projecto pretende construir um modelo que permita classificar títulos de notícias em três categorias distintas: **positivas**, **negativas** ou **neutras**. Como tal, neste subcapítulo são apresentadas todas as tarefas a elaborar no decorrer do desenvolvimento do modelo, bem como da sua validação.

Como mostra a imagem abaixo, o processo descrito a seguir será percorrido duas vezes, em tempos distintos, uma primeira vez para criar o *training corpus* que servirá para o desenvolvimento do modelo e, uma segunda para criar o *testing corpus*, para a respectiva validação. Esta imagem mostra a metodologia referida na revisão da literatura mas adaptada a este projecto.

Figura 3.2 - Processo de Desenvolvimento do Projecto

3.2.1. Selecção RSS Feeds

A correcta selecção das fontes de extracção de notícias é de grande importância para o sucesso do projecto, uma vez que será através desta que serão compiladas as notícias para o *training* e *testing corpus*. Os critérios utilizados para a escolha dos endereços serão:

- **Periocidade de actualização:** Serão escolhidos apenas os endereços que tenham periocidade de actualização diária, ou seja, que tenham actualizações ao longo do dia e que permitam uma constante monitorização. Este critério é fundamental se, posteriormente ao desenvolvimento do modelo, este passar a produção de forma a

avaliar diariamente as notícias que saem. Se os *RSS Feeds* escolhidos não disponibilizarem notícias diariamente, então não será possível a utilização do modelo.

- **Conteúdos**: Serão privilegiados endereços que contenham notícias que reflictam o estado da economia, tais como notícias de indicadores económicos, de mercados e de opiniões relativamente ao país. Endereços que contenham muitas notícias com conteúdo político serão excluídos.

3.2.2. Extracção das Notícias

Uma vez seleccionados os endereços de *RSS feeds* a utilizar no âmbito do projecto, seguir-se-á o período de extracção de notícias. Como referido anteriormente, nesta fase existirão dois períodos de extracção: um primeiro para compilar o *training corpus* e um segundo para o *testing corpus.* Em ambos, o procedimento a seguir será igual, pelo que apenas será descrito uma vez.

Nesta fase entrará pela primeira vez a utilização de *software SAS*, mais concretamente o *SAS Information Retrieval Studio* versão 1.3 e o *SAS Web Crawler* versão 2.1*,* que funcionarão interligados. A imagem que se segue ilustra a arquitectura do *SAS Web Crawler*.

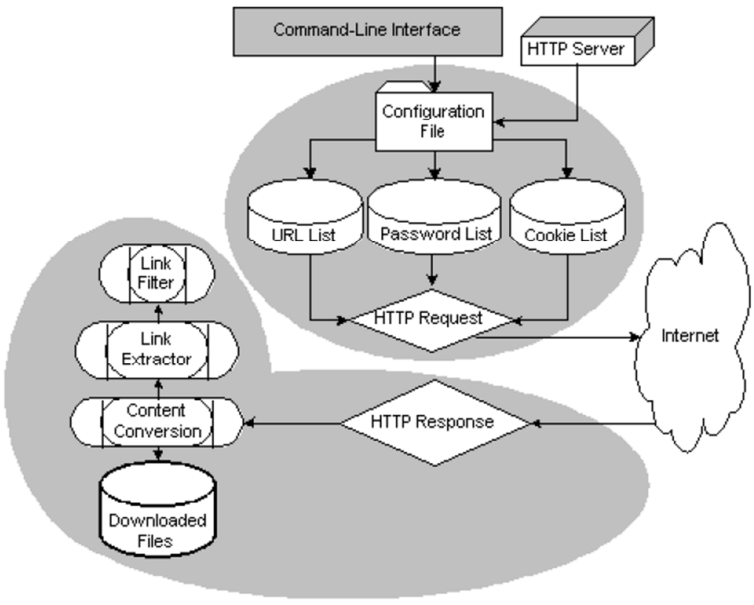


Figura 3.3 - Arquitectura do *SAS Web Crawler*

Ao *SAS Web Crawler*, apenas terá de ser disponibiliza a lista de URL para que este possa extrair as notícias desejadas. Como se pode ver pela imagem anterior, o *SAS Web Crawler*, depois de importar todas as notícias disponíveis no URL, armazená-las-á

em ficheiros de texto únicos. No final deste processo, estará disponível uma pasta com um ficheiro de texto por cada notícia extraída.

Para o *training corpus* e *testing corpus* serão extraídas apenas notícias uma vez por dia durante os períodos de extracção. Para o *training corpus* o processo de extracção foi iniciado a 15 de Março e concluído a 15 de Maio de 2012. Quanto ao *testing corpus*, o período da sua extracção decorreu entre 1 de Junho e 15 de Julho de 2012.

3.2.3. Armazenamento e Segmentação das Notícias

Após a extracção das notícias estar concluída será construído um processo que importará e guardará em *Microsoft SQL* todos os títulos de forma a facilitar a etapa seguinte, ou seja, a segmentação entre positivas, negativas ou neutras. O armazenamento numa base de dados torna-se realmente importante uma vez que, caso não o fosse feito, seria necessário abrir ficheiro a ficheiro para conseguir segmentar as notícias, uma vez que, como referido, as notícias serão disponibilizadas pelo *SAS Web Crawler* em ficheiros de texto. Para além disso, o armazenamento possibilitará a exploração dos dados, de forma a obter estatísticas de palavras importantes para a análise dos dados, bem como aplicação das regras definidas para a criação do *corpus* e para o modelo BeR.

Relativamente à segmentação das notícias, a definição do contexto em que se pretende actuar e do objectivo pretendido é realmente importante nesta fase do trabalho. Esta necessita de um bom planeamento prévio, uma vez que a má definição do contexto e do objectivo pode levar à má segmentação das notícias e, por consequência, à má definição do *training corpus.* Um *training corpus* desajustado do objectivo faz com que todo o trabalho seja posto em causa. Neste sentido, no momento de segmentação das notícias será muito relevante ter a definição de contexto e objectivo bem presente.

Quanto aos critérios de segmentação utilizados, estes estarão em consonância com as regras aplicadas no modelo BeR, que se encontra detalhado na fase seguinte. As notícias a utilizar para o treino e validação dos modelos serão classificadas de acordo com a sua influência para o estado da economia. É importante referir que esta segmentação será elaborada à base de *expert judgment*, já que não existe outra maneira de obter conjuntos de treino e teste classificados automaticamente. Assim, poderá haver sempre um erro associado a esta segmentação.

Para terminar, após a segmentação estar elaborada, serão novamente exportadas para ficheiros de texto para as respectivas pastas, de forma a poderem ser utilizadas pelo *SAS Sentiment Analysis*.

3.2.4. Desenvolvimento dos Modelos

Uma vez construídos os *training* e *testing corpus*, será a vez de desenvolver os três modelos propostos, seguindo a metodologia implementada no *SAS Sentiment Analysis*.

A metodologia implementada no *software* do SAS é, como a imagem seguinte mostra, baseada em três modelos, o primeiro estatístico, o segundo BeR e por fim a junção dos dois primeiros num modelo híbrido onde que se define uma ponderação a cada um deles.

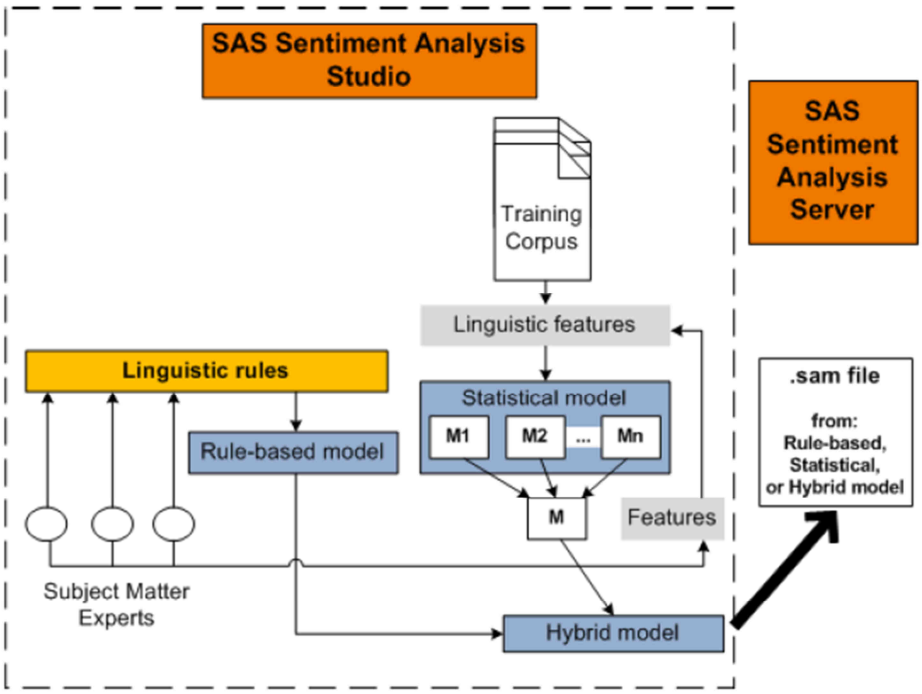


Figura 3.4 - Arquitectura do *SAS Sentiment Analysis*

Como se pode observar na figura anterior, o *SAS Sentiment Analysis Studio* necessita de um *training corpus* para a criação das regras linguísticas e, posteriormente, para a criação do modelo estatístico. Depois desta etapa será necessário reproduzir as regras definidas, de forma a criar o modelo BeR. Por fim, dá-se a criação do modelo hibrido, combinando linearmente os dois modelos anteriores. Terminada a criação destes modelos finais, será possível implementar no *SAS*

Sentiment Analysis Server o modelo vencedor, de forma a utilizá-los para dados futuros.

Referir ainda que as tarefas de NLP apresentadas no capítulo 2 e que precedem o desenvolvimento dos modelos, tais como segmentação de texto, *POS Tagging*, remoção de *stopwords* e selecção de características são efectuadas pelo *SAS Sentiment Analysis* em *background*, às quais não existe acesso.

3.2.4.1. Modelo Estatístico

No que diz respeito ao modelo estatístico, como podemos observar na imagem anterior, o *SAS* aplica a combinação de determinados métodos, anteriormente explicados, e escolhe o modelo vencedor com base na precisão dos mesmos. Estes modelos serão gerados através de aprendizagem supervisionada, uma vez que utilizam o *training corpus* na sua construção. A tabela seguinte mostra as diferentes combinações de métodos possíveis para o modelo que sairá vencedor, todos eles explicados no capítulo anterior.

Classificador	Normalização de Texto	Selecção Características
Naive Bayes	Smoothed Relative Frequency	No Feature Ranking
	Relative Frequency	Risk Ratio
	Okapi BM25	Chi Square
	Pivoted Length Normalization	Information Gain

Tabela 3.1 - Tabela com combinação de métodos utilizados pelo *SAS Sentiment Analysis* para o modelo estatístico

3.2.4.2. Modelo Baseado em Regras

Relativamente ao modelo BeR, como o nome indica, serão construídas regras por parte do analista de acordo com a sua percepção em relação ao que poderão ser notícias positivas, negativas ou neutras. Tal como na segmentação das notícias, as regras a criar serão baseadas em *expert judgment*.

Estas deverão, quanto possível, ajustar-se a um maior número de situações, de forma a que não sejam geradas em demasia. A criação de regras que abranjam poucas situações faz com que sejam geradas regras situação a situação. Consequência disto é o grande ajustamento do modelo ao *training corpus*, provocando o chamado *overfitting*, que mais tarde será visível na validação do modelo.

Além disto, a estrutura de regras a implementar neste modelo é a tarefa mais importante a definir para este tipo de modelos. O planeamento destas permite que as mesmas sejam criadas abrangendo um grande conjunto de situações, bem como

definir a melhor estrutura a seguir. Assim, este modelo deverá ter em atenção que as mesmas palavras em determinadas situações representam uma polaridade positiva e para outras situações uma polaridade negativa.

Com base neste pressuposto, as regras a aplicar no modelo seguirão a estrutura seguinte:

I. Regras para palavras/expressões, que isoladas, denotam alguma polaridade: Este tipo de palavras/expressões têm uma tendência que, por si só, independentemente do contexto em que estão inseridas, denotam alguma polaridade. Um exemplo muito concreto é o verbo “animar”. Este, independentemente de ter algum conceito na mesma frase, denota sempre polaridade positiva. Assim, a estrutura de regras terá uma secção para contemplar todas estas situações.

II. Regras para palavras/expressões que necessitam de um objecto para denotar alguma polaridade, sendo que na presença deste poderão haver regras de acordo com o tipo de objecto. O tipo de objecto poderá ser:

- Objectos que em conjunto com palavras/expressões que têm tendência crescente, denotam uma notícia com polaridade positiva e vice-versa;
  ex.: "Crescimento económico sobe para 3%" - Na presença de um objecto, neste caso "Crescimento económico", e duma palavra com tendência crescente, neste caso "sobe", a polaridade é positiva.
- Objectos que em conjunto com palavras/expressões que têm tendência crescente, denotam uma notícia com polaridade negativa e vice-versa;
  ex.: "Taxa de desemprego sobe para 3%" - Na presença de um objecto, neste caso "Taxa de desemprego", e duma palavra com tendência crescente, neste caso "sobe", a polaridade é negativa. Ou seja, para o caso anterior a palavra "sobe" tinha uma polaridade positiva enquanto que neste caso, a mesma palavra tem uma polaridade negativa.
- Objectos da categoria de combustíveis: têm uma estrutura de polaridade diferente e difícil de avaliar porque as mesmas palavras em determinadas situações podem conter polaridade positiva e noutras situações polaridade negativa. Um exemplo desta situação é a queda da cotação do petróleo que, em situação de guerra, poderá significar um abrandamento das tensões e por consequência um reflexo positivo na economia. Mas, em situações normais, a queda da cotação tem um reflexo negativo. Desta forma, as notícias relativas aos combustíveis estão muito mais

sujeitas às variações existentes na economia mundial. Assim, será criada uma secção na estrutura exclusivamente para os combustíveis de forma a ser mais fácil de alterar, em função da situação actual.

Resumindo, a imagem seguinte mostra a estrutura a utilizar no âmbito do modelo BeR.

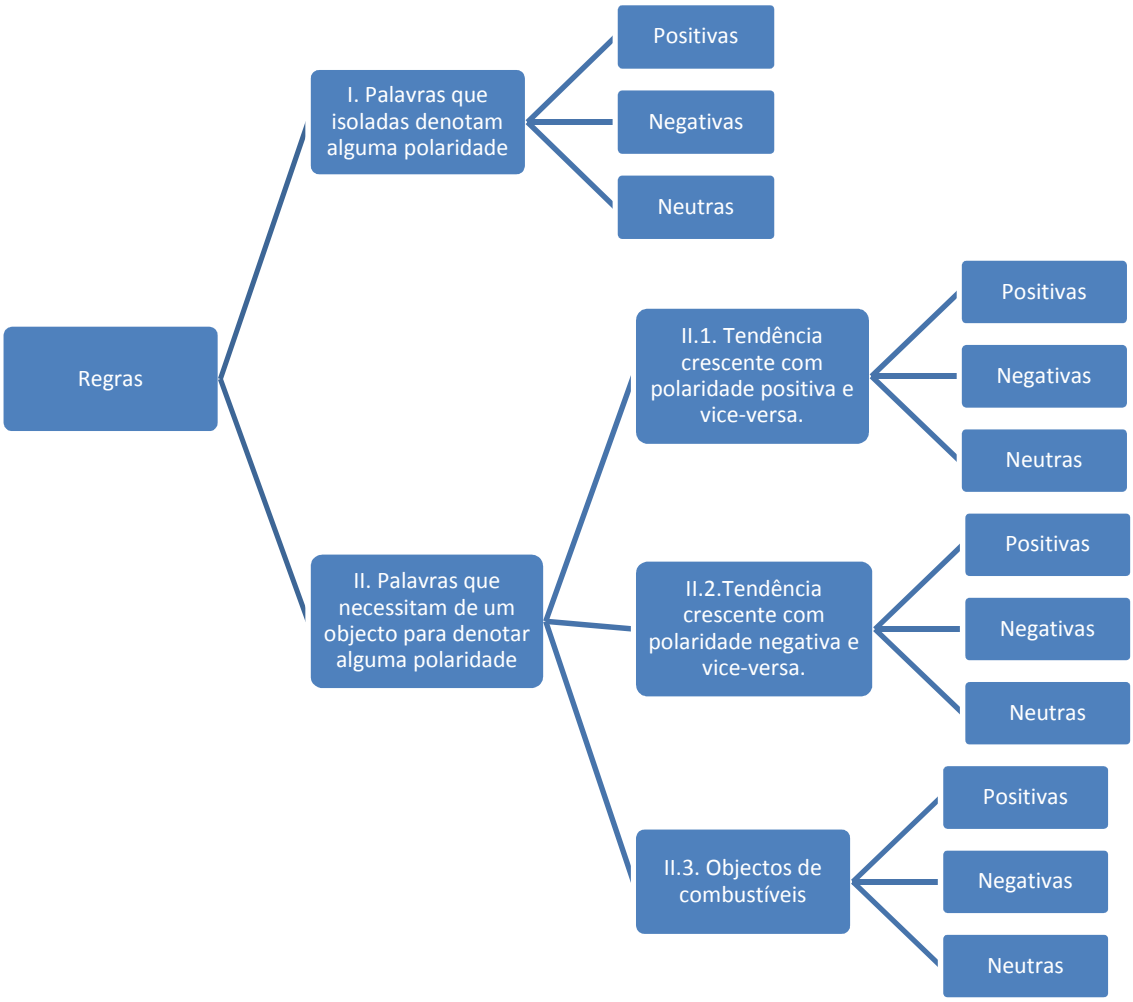


Figura 3.5 - Estrutura de Regras do Modelo BeR

3.2.4.3. Modelo Híbrido

No que respeita ao modelo híbrido, apenas terão de ser dados pesos a cada um dos modelos descritos anteriormente. De realçar que, à medida que o modelo BeR atinge uma maior maturidade, maior será o peso a dar a este na construção do modelo híbrido. No fim do projecto, eventualmente, poderá ser opção apenas usar um dos

modelos, o estatístico ou o BeR, se existir evidência que a utilização de apenas um se torna mais vantajosa face à utilização do modelo híbrido.

Apesar deste projecto ter um tempo limitado, o desenvolvimento deste tipo de modelos deverá ser contínuo, uma vez que deverá ser adaptado a novos tipos de títulos que vão saindo. Um modelo desenvolvido com base em notícias de hoje poderá não ser suficientemente preditivo quando a crise acabar, uma vez que neste momento apenas existem títulos associados à crise. Para além disso, uma notícia que hoje poderá ser considerada negativa, num futuro pós-crise poderá ser positiva. Desta forma, posteriormente ao desenvolvimento destes modelos, é aconselhado que para a utilização destes, exista uma equipa que trate da sua actualização de uma forma permanentemente.

3.2.5. Validação dos Modelos

Por fim, surge a fase de validação dos modelos construídos. Como explicado anteriormente, para esta fase será necessário um *testing corpus* e, para tal, foram percorridos novamente todos os passos seguidos na construção do *training corpus*. Uma vez construído o *testing corpus*, apenas será verificada a precisão de cada um dos modelos, de modo a poder observar qual o modelo que mais se ajusta à causa deste projecto. Esta etapa revela-se muito importante porque é nesta que poderemos observar a qualidade dos modelos construídos, bem como se um modelo construído com base nas notícias de um determinado período se ajusta a notícias de outro. Neste sentido, existirá uma diferença no tempo entre a extracção do *training corpus* e do *testing corpus*.

4. RESULTADOS E DISCUSSÃO

Este capítulo visa apresentar resultados obtidos com o desenvolvimento deste projecto e respectiva discussão, de acordo com a metodologia que foi proposta no capítulo anterior.

4.1. SELECÇÃO DOS ENDEREÇOS DE *RSS FEEDS*

Depois de serem verificados muitos dos endereços *RSS Feeds* que estavam disponíveis no início do projecto, a decisão inicial recaiu nos endereços que são indicados na tabela seguinte. Como se pode ver foram utilizados 9 endereços para a extracção de notícias. De referir que os critérios considerados relevantes foram tidos em conta, ou seja, para a escolha dos endereços foi avaliada a periocidade de actualização destes, bem como o respectivo conteúdo das notícias publicadas.

Endereço *RSS Feed*	Decisão
http://economico.sapo.pt/rss/mercados	Utilizar
http://economico.sapo.pt/rss/economia	Utilizar
http://feeds.dn.pt/DN-Economia	Utilizar
http://feeds.feedburner.com/PublicoEconomia	Utilizar
http://feeds.feedburner.com/PublicoPolitica	Não Utilizar
http://expresso.sapo.pt/static/rss/economia_23413.xml	Utilizar
http://expresso.sapo.pt/static/rss/mercados_25511.xml	Não Utilizar
http://expresso.sapo.pt/static/rss/dinheiro_25283.xml	Utilizar
http://www.oje.pt/rss	Utilizar
http://www1.ionline.pt/rss/dinheiro.xml	Não Utilizar
http://feeds.feedburner.com/iol/agenciafinanceira	Não Utilizar
http://feeds.controlinveste.pt/DV-mercados	Utilizar
http://feeds.controlinveste.pt/DV-economia	Utilizar

Tabela 4.1 - Endereços *RSS Feeds*

Assim, com o início das extracções foi possível verificar que alguns dos endereços escolhidos não continham actualizações frequentes ou continham maioritariamente notícias de conteúdo politico. Deste modo, a coluna "Decisão" mostra quais os endereços que foram utilizados para a extracção dos títulos que possibilitaram mais tarde a criação do *training* e *testing corpus*.

De salientar que durante os dois períodos de extracção foram recolhidos mais de 15000 títulos de notícias dos endereços referidos, apesar de muitos deles terem sido excluídos da análise, uma vez que segmentar um número tão elevado de notícias seria uma tarefa muito demorada e não acrescentaria mais valor ao desempenho dos

modelos. Além disso, algumas notícias continham informação positiva e negativa na mesma frase, tendo sido eliminadas da análise, uma vez que não seria correcto atribuir unicamente uma categoria para estas frases. Algumas delas foram divididas em duas notícias, de maneira a que pudessem ser classificadas para cada uma das categorias.

4.2. SEGMENTAÇÃO DE NOTÍCIAS

Relativamente à segmentação de notícias para o *training* e *testing corpus*, a tabela seguinte ilustra os períodos de extracção correspondentes, tal como o número de notícias por cada uma das amostras.

As duas amostras foram seleccionadas aleatoriamente da população de notícias, extraídas nos períodos indicados, tendo sido seleccionadas 5495 notícias para o conjunto de treino e 500 para o conjunto de teste.

Corpus	Período de Extracção	Amostra	Positivas	Negativas	Neutras
Treino	15-Mar > 15-Mai-2012	5495	1008	1623	2864
Teste	1-Jun > 15-Jul-2012	500	110	120	270

Tabela 4.2 - Informação das amostras seleccionadas para o treino e validação dos modelos

É importante salientar que a contextualização do objectivo do trabalho revelou-se muito importante nesta fase. A segmentação das notícias só foi possível porque era sabido claramente que o objectivo pretendido seria classificar as notícias de acordo com o reflexo que estas tinham do actual estado da economia.

4.3. DESENVOLVIMENTO DOS MODELOS

O objectivo deste subcapítulo é dar a conhecer os resultados obtidos com o desenvolvimento dos modelos, tal como relatar as decisões tomadas face a esses resultados. Assim sendo, este subcapítulo está dividido por modelo, de forma a apresentar estruturadamente os desenvolvimentos efectuados em cada uma das fases.

Salientar ainda que, no decorrer do projecto, foi verificado que o *software* utilizado não tem em conta as notícias neutras no módulo do modelo estatístico. Ou seja, apesar de pedir um conjunto de treino de notícias neutras, este não as utiliza na fase de classificação. Além disto, o modelo classifica as notícias como "Positiva" por defeito. Isto é, mesmo que a probabilidade de pertencer às notícias positivas ou negativas seja igual a zero, o modelo classifica-as sempre como positivas, significando que para o modelo estatístico não existirá qualquer notícia neutra. Ainda neste

sentido, no módulo correspondente aos modelos BeR, este também ignora os pesos atribuídos às regras para notícias neutras. Assim sendo, neste módulo, o modelo só classifica a notícia como neutra se não existir nenhuma regra positiva ou negativa que seja captada.

Deste modo, uma vez que o modelo híbrido se trata da combinação dos outros dois, acaba também ele por ser penalizado por esta situação. Aquando da apresentação das limitações do projecto, este assunto será novamente referenciado.

4.3.1. Modelo Estatístico

Na elaboração do modelo estatístico foi seleccionada a opção do *software* que executa todas as combinações possíveis de métodos para a normalização e dos métodos de selecção de características. Foi, então, possível obter um conjunto de combinações, de modo a poder seleccionar a combinação que apresentasse um resultado mais favorável. Consequentemente, a tabela abaixo mostra os resultados para as várias combinações. É visível que a solução vencedora é a combinação do método de normalização *Relative Frequency* com o método de selecção de características *Chi Square*. Com estes, foi possível obter uma precisão para o modelo de mais de 81%. A precisão associada às notícias negativas é bastante boa uma vez que se trata de 90%. Aqui já é possível observar que o modelo se consegue adaptar muito mais a estas notícias face ao resultado de 66% de precisão para as positivas.

Método de Normalização	Selecção de Características	Precisão		
		Global	Positivas	Negativas
Smoothed Relative Frequency	No Feature Ranking	79.17%	65.35%	87.73%
Smoothed Relative Frequency	Risk Ratio	80.68%	64.36%	90.80%
Smoothed Relative Frequency	Chi Square	78.03%	62.38%	87.73%
Smoothed Relative Frequency	Information Gain	76.89%	61.39%	86.50%
Relative Frequency	No Feature Ranking	80.68%	65.35%	90.18%
Relative Frequency	Risk Ratio	74.24%	42.57%	93.87%
Relative Frequency	Chi Square	81.06%	66.34%	90.18%
Relative Frequency	Information Gain	78.79%	66.34%	86.50%
Okapi BM25	No Feature Ranking	79.17%	62.38%	89.57%
Okapi BM26	Risk Ratio	74.62%	42.57%	94.48%
Okapi BM27	Chi Square	80.68%	64.36%	90.80%
Okapi BM28	Information Gain	79.55%	66.34%	87.73%
Pivoted Length Normalization	No Feature Ranking	45.45%	66.34%	43.56%
Pivoted Length Normalization	Risk Ratio	45.45%	48.51%	43.56%
Pivoted Length Normalization	Chi Square	47.73%	63.37%	38.04%
Pivoted Length Normalization	Information Gain	48.86%	67.33%	37.42%

Tabela 4.3 - Indicadores de qualidade do modelo estatístico para as várias combinações de métodos

Estes indicadores, tal como o classificador, também não têm em consideração os falsos positivos e negativos que provêm da classificação da categoria de neutros. Por conseguinte, existe um erro que não é visível nesta tabela, mas que também não pôde ser calculado, uma vez que o *software* não o permite. Ou seja, para o cálculo destes indicadores é necessário obter os falsos positivos e os falsos negativos. Neste caso, os títulos neutros que são incorrectamente classificados como negativos ou positivos não estão a ser considerados para o cálculo do indicador, o que provoca um erro que não é observável.

Além disso, depois de construído o modelo e verificados os pesos associados a cada *token*, verificou-se a inexistência de um método que identifique as entidades presentes no texto. Esta situação faz com que as entidades que apareçam no *training corpus* tenham associado a elas a probabilidade de serem positivas ou negativas. Esta enfoca mais a necessidade de o modelo ser reajustado regularmente, dado que algumas entidades podem aparecer neste momento maioritariamente em notícias negativas e, terminada a crise, essa situação mudar.

4.3.2. Modelo Baseado em Regras

No decorrer do desenvolvimento deste modelo foram criadas cerca de 1300 regras. As tabelas seguintes apresentam alguns exemplos de regras criadas para este modelo: a tabela 4.4 com regras para a categoria das positivas e a 4.5 com regras para a das negativas.

#	Regra
1	(DIST_4,(OR,"_a{mantém}","_a{manter@}"),"_b{triplo A}")
2	(DIST_3, "_a{crescimento}", "_b{reactivar@}")
3	(DIST_3, "_a{registar@}", "_b{crescimento}")
4	(DIST_4, "_a{aliviar@}","_b{pressão}")
5	(DIST_4, "_a{elo}", "_b{forte}")
5	(DIST_5, "_a{crescimento}", "_b{%}")
6	(ORDDIST_2, "_a{fazer@}", "bem")
8	(ORDDIST_3, "_a{negociar@}", "_b{acima}")
9	(ORDDIST_5, "_a{inverter@}", "_b{tendência de queda}")
10	(SENT, "_a{combater@}","desemprego")
12	(SENT, "_a{criar@}","emprego")
13	(SENT, "_a{euro}", (OR,"manter@","ficar@","permanecer@"))

Tabela 4.4 – Exemplos de regras criadas para a classe de títulos positivos

É importante realçar que, quanto mais regras negativas forem capturadas numa frase, maior será a probabilidade de esta ser classificada como negativa, acontecendo o mesmo para as positivas.

A primeira regra observada na tabela 4.4 contempla todas as situações em que o verbo “manter” surja em conjunto com a expressão “Triplo A” na mesma frase, a uma distância de 4 *tokens*, independentemente da ordem em que estas surjam. Para a sétima regra, acontece o mesmo para as palavras em causa, no entanto a ordem é considerada. Quanto à décima, são contempladas todas as situações em que todas as formas do verbo “combater” em conjunto com a palavra “desemprego” apareçam na mesma frase.

#	Regra
1	(DIST_6, "_a{sair@}", "_b{euro}")
2	(ORDDIST_3, "_a{apertar@}", "_b{cinto}")
3	(ORDDIST_3, "_a{negociar@}", "_b{vermelho}")
4	(ORDDIST_3, "_a{negociar@}", "_b{abaixo}")
5	(DIST_4, "_a{empregos}","_b{risco}")
5	(DIST_4, "_a{taxa de desemprego}","_b{escalada}")
6	(ORDDIST_3, "_a{não}", "_b{funcionar@}")
8	(DIST_4, "_a{pedido}", (OR, "_b{ajuda}", "_d{apoio}))
9	(SENT, "_a{menos}", (OR, "investimento","consumo"))
10	(SENT, "_a{receber@}","_b{pedidos}", "_d{ajuda}")
12	(SENT, "_a{resgate}", (OR,"_d{FMI}","_e{Troika}","_f{CE}", "_g{Comissão}"))
13	(DIST_5, "milhões","_b{perdas}")

Tabela 4.5 - Exemplos de regras criadas para a classe de títulos negativos

A tabela acima ilustra alguns exemplos de regras criadas para a classe de títulos negativos. Como se pode observar, as regras criadas para a classe dos títulos negativos utilizam as mesmas funções do *SAS Sentiment Analysis*, mudando apenas as palavras, e respectiva combinação, associadas à regra.

A figura seguinte exemplifica como as regras definidas são aplicadas no modelo BeR. Como se pode constatar, a notícia é bastante dúbia porque apresenta determinadas palavras que podem inverter o sentido negativo da mesma. Ainda assim, apesar dos "ganhos europeus" a notícia tem um grande enfase na situação de Portugal e da Grécia e, como tal, são capturadas mais regras negativas do que positivas.

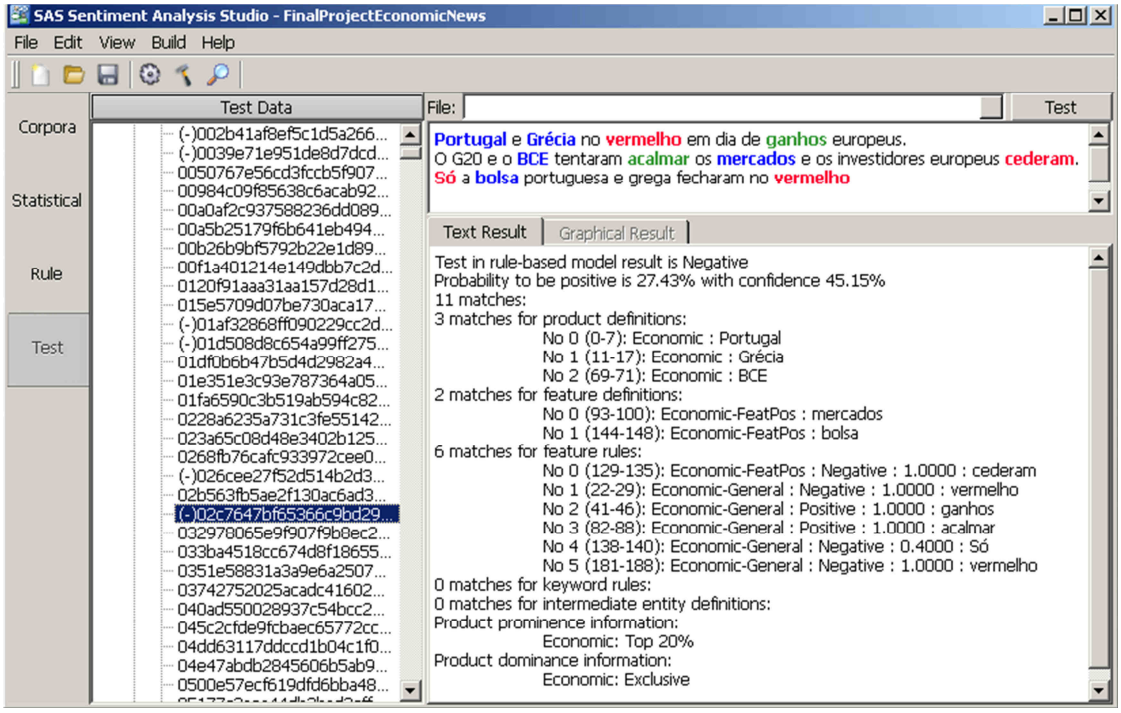


Figura 4.1 – Exemplo de aplicação das regras a uma notícia

Relativamente aos resultados do modelo, a tabela seguinte ilustra os números obtidos na fase de desenvolvimento.

			Sem Neutras			Com Neutras		
Verdadeiros Negativos	**Verdadeiros Positivos**	**Verdadeiros Neutros**	**Falsos Negativos**	**Falsos Positivos**	**Falsos Neutros**	**Falsos Negativos**	**Falsos Positivos**	**Falsos Neutros**
1470	853	1084	112	74	122	1015	951	122

Tabela 4.6 - Resultados obtidos no modelo BeR

Os resultados obtidos são muito satisfatórios. Como se pode constatar, através dos números de notícias bem classificadas, o modelo consegue classificar correctamente cerca de 91% das notícias positivas, 85% das notícias negativas e apenas 38% das neutras. Esta situação deve-se ao facto referido anteriormente, ou seja, o *sofware* não tem em atenção as notícias neutras, classificando-as apenas se não captar nenhuma regra de positivas e/ou negativas. Assim, e uma vez que o *software* não é capaz de as utilizar, não foram desenvolvidas regras para as mesmas. Deste modo, as notícias que este classifica como neutras são apenas aquelas que não têm probabilidade de pertencer nem às positivas nem às negativas, ou seja, que não captaram nenhuma regra relativa aos títulos positivos ou negativos.

A tabela seguinte revela os indicadores referentes à sua qualidade para o *training corpus*. Como constatado nos números anteriores, este modelo apresenta

uma precisão de 88% e um erro de 12%, sem as notícias neutras incluídas. Se incluídas, o modelo perde muita precisão e aumenta o erro consideravelmente.

			Sem Neutras		Com Neutras	
Precisão Negativas	Precisão Positivas	Precisão Neutras	Precisão Total	Erro	Precisão Total	Erro
83%	73%	78%	88%	12%	62%	38%

Tabela 4.7 - Indicadores de qualidade do modelo BeR

No entanto, trata-se do modelo que tem o melhor comportamento quando se pretende introduzir as notícias neutras na análise. Este modelo consegue obter uma precisão de 62%, mesmo considerando as notícias neutras, apesar do que foi dito anteriormente.

Realçar ainda que, nesta fase e no desenvolvimento das regras para as notícias positivas, foi observado que muitas regras tiveram de ser construídas. Isto significa que não é tarefa fácil obter uma regra que cubra muitas notícias positivas. Ou seja, contrariamente às notícias negativas, para estas notícias é mais comum aparecerem novas que não sejam captadas pelas regras já definidas. Esta situação ajuda a perceber a fraca performance do modelo face a estas notícias, podendo ser justificada porque é mais comum serem publicadas notícias negativas do que positivas e, pelo facto de o período que atravessamos ajudar na contribuição para a pouca existência de notícias positivas.

4.3.3. Modelo Híbrido

Uma vez desenvolvido o modelo estatístico e o BeR, chegou a fase de construir o modelo híbrido. Este trata-se da combinação linear dos modelos explicados anteriormente e o objectivo deste é, através dessa combinação, obter o melhor resultado.

A tabela seguinte mostra os testes efectuados no sentido de obter a melhor combinação.

				Sem Neutras		Com Neutras	
Importância do Modelo Estatístico	Precisão Negativas	Precisão Positivas	Precisão Neutras	Precisão Total	Erro	Precisão Total	Erro
20%	94%	94%	100%	94,0%	6,0%	45,2%	54,8%
17%	94%	95%	100%	94,2%	5,8%	45,3%	54,7%
15%	94%	96%	100%	94,5%	5,5%	45,4%	54,6%
14%	94%	96%	100%	94,5%	5,5%	45,4%	54,6%
12%	94%	96%	100%	94,3%	5,7%	45,4%	54,6%

Tabela 4.8 - Testes efectuados para obter melhor combinação no modelo Híbrido

Era expectável que a ponderação do modelo estatístico fosse bastante inferior à do BeR e, desse modo, a ponderação vencedora foi a dos 15% para o modelo estatístico e os restantes 85% para o modelo BeR. A tabela mostra-nos que esse é o ponto onde maximizamos a precisão e minimizamos o erro, tanto para a análise que contempla as notícias neutras, como para a análise que não as contempla. Se não forem consideradas as notícias neutras na análise, este é o modelo que apresenta melhores resultados. Com este é possível obter uma precisão de cerca de 95% e um erro inferior a 6%. No entanto, se quisermos introduzir as notícias neutras na análise, este modelo perde muita precisão e aumenta o erro. Deste modo, a ponderação de 15% para o modelo estatístico foi a escolhida para construir o modelo híbrido final. Esta será a ponderação a utilizar na validação deste modelo.

4.4. VALIDAÇÃO DOS MODELOS

Esta fase tem como principal objectivo validar os modelos que foram construídos. É nesta que será possível, através de outro conjunto de dados, avaliar a capacidade dos modelos de classificar correctamente os títulos de outras notícias. Na tabela seguinte pode-se observar os resultados obtidos para a amostra de teste.

				Sem Neutras			Com Neutras		
Modelo	Verdadeiros Negativos	Verdadeiros Positivos	Verdadeiros Neutros	Falsos Negativos	Falsos Positivos	Falsos Neutros	Falsos Negativos	Falsos Positivos	Falsos Neutros
Estatístico	104	82	4	28	16	0	200	110	0
Regras	97	93	141	10	8	15	71	76	22
Híbrido	113	88	3	22	7	0	196	100	0

Tabela 4.9 - Tabela com os resultados da validação dos três modelos

Os resultados que são apresentados têm em conta a existência de notícias neutras e também a sua inexistência. Isto porque apresentar apenas os resultados tendo em conta a existência das mesmas seria errado, uma vez que o modelo não as teve em consideração. Por outro lado, apresentar apenas os resultados sem a inclusão das neutras, faria com que não fosse considerado o erro que está adjacente à classificação destas notícias e, no momento de utilização do modelo, este teria que ser tido em consideração na análise dos resultados. Seguidamente, são apresentados os principais indicadores que avaliam a qualidade dos modelos desenvolvidos.

				Sem Neutras		Com Neutras	
Modelo	**Precisão Negativas**	**Precisão Positivas**	**Precisão Neutras**	**Precisão Total**	**Erro**	**Precisão Total**	**Erro**
Estatístico	79%	84%	100%	81%	19%	38%	62%
Regras	91%	92%	90%	83%	14%	66%	34%
Híbrido	84%	93%	100%	87%	13%	41%	59%

Tabela 4.10 - Tabela com indicadores de qualidade dos modelos

Como podemos observar através da tabela, os modelos apresentam muito boa qualidade se as notícias neutras não forem incluídas na análise. Caso sejam consideradas, a qualidade desce consideravelmente.

Na análise sem a inclusão das notícias neutras, o modelo híbrido é o que oferece os melhores resultados, uma vez que apresenta um erro inferior e uma precisão total superior, face aos outros dois. É importante referir que é um modelo que se comporta muito bem na classificação das negativas, com 94% de acerto, e também nas positivas, com cerca de 80%. No entanto, peca por classificar apenas 1% das notícias neutras. Relativamente às estatísticas finais, sem a inclusão das neutras apresenta uma precisão total de 87% e um erro de 13%. Se as incluirmos, o erro dispara para 59% e a precisão desce para 41%.

Se for pretendido introduzir as neutras na análise então a escolha do modelo cairá no modelo BeR, uma vez que se trata do modelo onde se minimiza o erro e se maximiza a precisão total. De referir que este é o modelo que mais acerta nas notícias neutras (52% das notícias neutras bem classificadas). Ainda assim, trata-se de um modelo com estatísticas muito baixas no caso de se pretender considerar estas notícias para a análise.

5. CONCLUSÕES

Neste projecto realizou-se um importante trabalho para o desenvolvimento da DCT para a língua portuguesa. Foram introduzidas as áreas que poderão ser utilizadas nesta descoberta de conhecimento. Este estudo destacou ainda a importância da utilização da DCT e, em particular, da AS, para o mundo actual e a necessidade de aperfeiçoamento do *software* utilizado. Além disso, foi possível descrever quais as razões que levam, cada vez mais, as empresas a procurar nestas técnicas importantes formas de diferenciação.

Foi explorada uma metodologia proposta noutros trabalhos de investigação e que é, ao mesmo tempo, utilizada na plataforma de *Text Analytics* do SAS. As técnicas utilizadas e referenciadas no âmbito da mesma foram retratadas tendo em conta o estado da arte, tendo ficado provado, através deste trabalho, que se trata de uma metodologia adequada para as organizações que pretendam utilizar diversas fontes de dados em formato textual para a descoberta de conhecimento.

Por fim, foram construídos os modelos considerados como um dos objectivos deste trabalho. Relativamente a estes, verificou-se que a impossibilidade de utilização das notícias neutras constitui um problema para a criação deste tipo de modelos. Tendo em conta que não foi possível considera-las na fase de aprendizagem do modelo, verificou-se a real importância das mesmas. Modelos que não as considerem aumentam a classificação dos falsos negativos e positivos, fazendo com que o erro aumente e poderá inviabilizar a utilização do modelo.

Ao nível das notícias, verificou-se que os modelos tendencialmente classificam mais correctamente as notícias com reflexo negativo na economia. Esta situação era expectável, uma vez que é mais frequente os meios de comunicação publicarem as notícias negativas do que as positivas devido, essencialmente, ao seu mediatismo. Apesar disso, a utilização apenas da informação fornecida pelos modelos do lado das notícias negativas é possível, uma vez que será perfeitamente possível avaliar o reflexo destas na economia através das oscilações diárias no número de notícias negativas que saem.

Neste sentido, concluiu-se que o modelo que apresenta melhores resultados para o objectivo pretendido trata-se do modelo BeR. Isto deve-se ao facto de o modelo estatístico não contemplar as notícias neutras. Desse modo, o modelo Híbrido fica sujeito também a esta situação por ser influenciado pelo modelo estatístico. Assim, o modelo que minimiza o erro tendo em consideração as notícias neutras, é o modelo BeR. Para além disto, como foi dito no decorrer deste documento, os modelos

baseados em regras tornam-se mais eficazes à medida que a sua maturidade vai evoluindo. Este modelo é a prova disso, uma vez que, depois de muitos ajustes, consegue ter uma ponderação no modelo híbrido de 85% face aos 15% do modelo estatístico.

Ainda assim, a utilização deste modelo fica condicionada pelos factores descritos. No entanto, a sua utilização torna-se viável se forem considerados os erros associados.

6. LIMITAÇÕES E RECOMENDAÇÕES PARA TRABALHOS FUTUROS

Relativamente às limitações encontradas no decorrer do projecto, a primeira a tratou-se duma limitação de tempo, uma vez que o projecto inicial seria para decorrer numa empresa com *software* e dados próprios. Este foi adiado até ser cancelado e, desse modo, condicionou o tempo disponível para a realização daquele que viria a ser o tema final.

Além disso, as restantes limitações foram amplamente apresentadas no capítulo anterior, uma vez que condicionaram os resultados obtidos. Estas decorreram de problemas associados à maturidade do *software* utilizado. Uma vez que estas condicionaram os resultados dos modelos, são também apresentados como recomendações para trabalhos futuros, depois das situações estarem todas resolvidas no *software*. Estas situações poderão ser repartidas em:

I. Categoria de Neutras - a não consideração das notícias neutras na aprendizagem e classificação dos modelos faz aumentar o erro consideravelmente para as categorias positivas e negativas, uma vez que todas as notícias neutras são classificadas nessas duas categorias. Assim sendo, a categoria de neutras torna-se tão importante quanto as outras, uma vez que é necessário distingui-las. Neste sentido, tendo em conta que o *software* não as considera actualmente, seria importante actualizá-lo no sentido de evitar o que foi dito, uma vez que a consideração destas se torna muito relevante para a maioria das situações.

II. Identificação de Entidades Nomeadas - trata-se de uma tarefa bastante importante e que poderá ser resolvida com a construção de um modelo BeR. No entanto, seria muito relevante considerar a inclusão de um algoritmo que identificasse entidades no modelo estatístico. Esta identificação é importante porque faz com que o modelo considere apenas relevante, para a classificação, as palavras que caracterizam ou descrevem as entidades. Ou seja, uma entidade não deverá servir para classificar uma notícia mas sim as características que lhes são atribuídas no seguimento do texto analisado.

III. Normalização de palavras - Seria muito relevante trabalhar a este nível. Exemplos de palavras que deverão ser fruto de uma normalização são as palavras que contêm um ou mais hífenes. O algoritmo de *lemmatization* deverá também ser melhorado de forma a conseguir identificar determinadas formas das palavras em português, como, por exemplo, "mantém-se", uma vez que este não considera esta palavra como uma

derivação do verbo "manter". Outro exemplo é o verbo "subir" que em determinadas situações poderá ver a letra u substituída pela letra o, sendo que este não consegue identificá-las como pertencentes ao verbo em causa. Relativamente a outros casos, foram detectadas situações em que a não remoção de acentos e caracteres especiais dificultaram a criação de regras. Estas situações aqui explicadas provocam o aumento da criação de regras ou, no caso de não serem detectadas, a não classificação de situações que o analista pensaria estarem a ser consideradas através do algoritmo de expansão de palavras. Neste sentido, qualquer trabalho a desenvolver na língua portuguesa deverá ter sempre em atenção que a normalização de palavras é uma importante tarefa a desempenhar.

IV. Indicadores de avaliação dos modelos - a não consideração dos documentos da classe neutra tem efeitos nos indicadores que avaliam os modelos. Esta situação faz com que os indicadores acabem por ser sempre mais favoráveis, uma vez que não consideram os falsos positivos e falsos negativos que derivam da classe neutra, fazendo com que os resultados apresentados para o modelo estatístico não contemplem esta situação.

A resposta obtida pelo *SAS* para estas questões remete para o facto do *software* se tratar de uma versão muito prematura e que a prioridade do desenvolvimento foi para as categorias positivas e negativas, deixando as restantes questões para futuros desenvolvimentos.

Referir ainda que existe um importante trabalho a fazer no sentido de melhorar as situações de negação, ou seja, quando existe alguma palavra (ex. "não") que inverta o sentido da frase. Apesar de terem sido elaboradas regras que possibilitam a incorporação desta informação no modelo BeR, será necessário explorar melhor esta vertente, no sentido de contemplar mais situações.

Quanto ao trabalho futuro, para além do melhoramento deste modelo de modo a considerar aquilo que foi dito anteriormente, será possível construir novos modelos que utilizem a informação gerada pelo aqui desenvolvido. A título de exemplo, seria interessante construir um modelo que, tendo em conta o número de notícias negativas e positivas, verificasse qual a relação deste com o efectivo comportamento da economia ou, também, com a evolução da bolsa nacional. Se esta verificação fosse comprovada, poderemos obter um modelo que preveja a tendência da bolsa ao longo do dia, o que seria uma importante ajuda para pequenos investidores.

Este projecto possibilita também, como era pretendido inicialmente, uma ajuda na elaboração de novos projectos para organizações que pretendam iniciar a extracção

de conhecimento nas bases de dados de documentos textuais, uma vez que a metodologia aqui implementada poderá servir de exemplo.

7. BIBLIOGRAFIA

Aranha, C. N. (2007). *Uma abordagem de pré-processamento automático para mineração de textos em português: sob o enfoque da inteligência computacional*. PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO - PUC-RIO.

Berry, M. W., & Kogan, J. (2010). *Text Mining: Applications and Theory*. Chichester: John Wiley & Sons Ltd.

Chowdhury, G. G. (2003). Natural language processing. *Annual Review of Information Science and Technology*, *37*(1), 51–89. doi:10.1002/aris.1440370103

Dörre, J., Gerstl, P., & Seiffert, R. (1999). Text mining: finding nuggets in mountains of textual data. *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 398–401). New York, NY, USA: ACM. doi:10.1145/312129.312299

Fan, W., Wallace, L., Rich, S., & Zhang, Z. (2006). Tapping the power of text mining. *Commun. ACM*, *49*(9), 76–82. doi:10.1145/1151030.1151032

Feldman, R, Fresko, M., Hirsh, H., Aumann, Y., Liphsta, O., Schler, Y., & Rajman, M. (1998). Knowledge Management: A Text Mining Approach. *Proc. the 2nd Int. Conf. on Practical Aspects of Knowledge Management (PAKM98)* (pp. 9.1–9.10).

Feldman, Ronen. (2006). *Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge: Cambridge Univ. Press.

Feldman, S. (1999). NLP meets the jabberwocky: Natural language processing in information retrieval : Search Engine Section. (Online, Ed.)*Online (Weston, CT)*, *23*(3), 62–72. Retrieved from http://www.refdoc.fr/Detailnotice?idarticle=11558626

Gharehchopogh, F. S. (2010). Approach and review of user oriented interactive data mining. *Application of Information and Communication Technologies (AICT), 2010 4th International Conference on* (pp. 1–4). doi:10.1109/ICAICT.2010.5611792

Gupta, V., & Lehal, G. (2009). A Survey of Text Mining Techniques and Applications. *Journal of Emerging Technologies in Web Intelligence*, *1*(1). Retrieved from http://ojs.academypublisher.com/index.php/jetwi/article/view/01016076

Hearst, M. A. (1999). Untangling text data mining. *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* (pp. 3–10). Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.3115/1034678.1034679

Hotho, A., Nürnberger, A., & Paaß, G. (2005). A brief survey of text mining. *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology*.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 168–177). New York, NY, USA: ACM. doi:10.1145/1014052.1014073

Indurkhya, N., & Damerau, F. J. (2010). *Handbook of Natural Language Processing, Second Edition*. Taylor & Francis. Retrieved from http://books.google.pt/books?id=nK-QYHZ0-_gC

Inniss, T. R., Lee, J. R., Light, M., Grassi, M. A., Thomas, G., & Williams, A. B. (2006). Towards applying text mining and natural language processing for biomedical ontology acquisition. *Proceedings of the 1st international workshop on Text mining in bioinformatics* (pp. 7–14). New York, NY, USA: ACM. doi:10.1145/1183535.1183539

Jackson, P., & Moulinier, I. (2002). *Natural Language Processing for Online Applications: Text Retrieval, Extraction, and Categorization*. John Benjamins Pub. Retrieved from http://books.google.pt/books?id=jkkoj7U5g4kC

Jiang, E. (2006). Learning to Semantically Classify Email Messages. In D.-S. Huang, K. Li, & G. Irwin (Eds.), *Intelligent Control and Automation* (Vol. 344, pp. 700–711). Springer Berlin / Heidelberg. Retrieved from http://dx.doi.org/10.1007/978-3-540-37256-1_86

Junior, J. R. C. (2008). *Desenvolvimento de uma metodologia para Mineração de Textos*. PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO - PUC-RIO.

Kao, A., & Poteet, S. R. (2010). *Natural Language Processing and Text Mining* (1st ed.). Springer Publishing Company, Incorporated.

Kim, S.-B., Han, K.-S., Rim, H.-C., & Myaeng, S. H. (2006). Some Effective Techniques for Naive Bayes Text Classification. *Knowledge and Data Engineering, IEEE Transactions on*, *18*(11), 1457–1466. doi:10.1109/TKDE.2006.180

Konchady, M. (2006). *Text Mining Application Programming (Programming Series)*. Rockland, MA, USA: Charles River Media, Inc.

Li, G., & Liu, F. (2010). A clustering-based approach on sentiment analysis. *Intelligent Systems and Knowledge Engineering (ISKE), 2010 International Conference on* (pp. 331–337). doi:10.1109/ISKE.2010.5680859

Liddy, E. (2003). Natural Language Processing. *Encyclopedia of Library and Information Science*. New York: Marcel Decker, Inc.

Liu, B. (2010). Sentiment Analysis and Subjectivity. In Nitin Indurkhya & F. J. Damerau (Eds.), *Handbook of Natural Language Processing, Second Edition*. Boca Raton, FL: CRC Press, Taylor and Francis Group.

Liu, B., Hu, M., & Cheng, J. (2005). Opinion observer: analyzing and comparing opinions on the Web. *Proceedings of the 14th international conference on World Wide Web* (pp. 342–351). New York, NY, USA: ACM. doi:10.1145/1060745.1060797

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press. Retrieved from http://books.google.pt/books?id=t1PoSh4uwVcC

PSE - Produtos e serviços de Estatística, L. (2011). *Feedback Management em Portugal*.

Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Found. Trends Inf. Retr.*, *2*(1-2), 1–135. doi:10.1561/1500000011

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10* (pp. 79–86). Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.3115/1118693.1118704

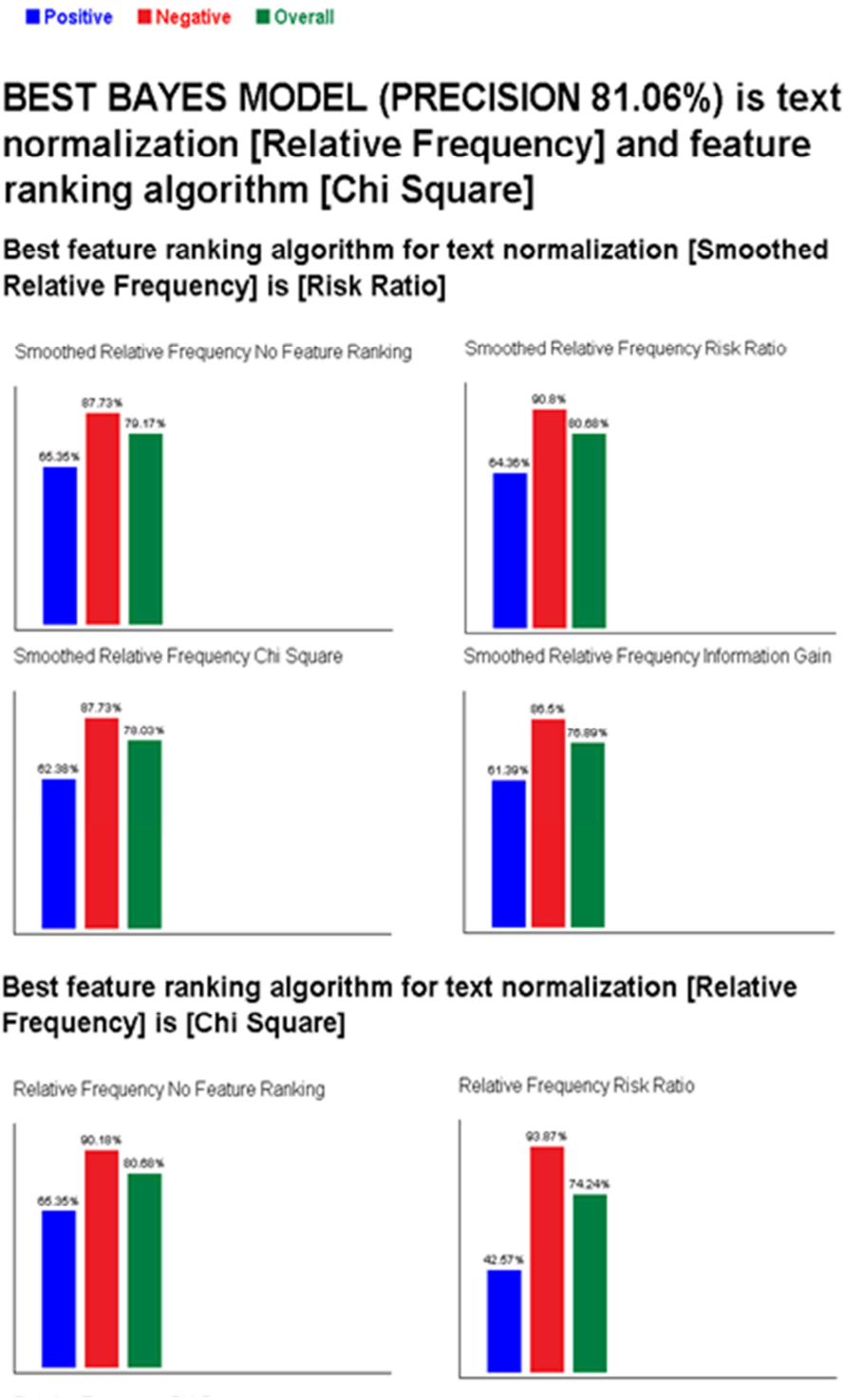
Prabowo, R., & Thelwall, M. (2009). Sentiment analysis: A combined approach. *Journal of Informetrics*, *3*(2), 143–157. doi:10.1016/j.joi.2009.01.003

Tan, A. (1999). Text Mining: The state of the art and the challenges. *Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases* (pp. 65–70).

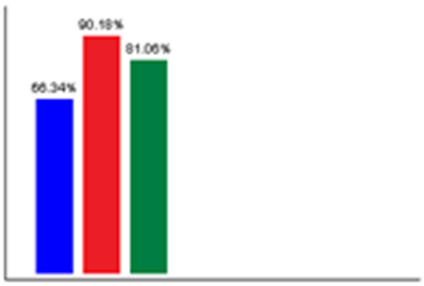
Wiebe, J., & Riloff, E. (2005). Creating Subjective and Objective Sentence Classifiers from Unannotated Texts. In A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing* (Vol. 3406, pp. 486–497). Springer Berlin / Heidelberg. Retrieved from http://dx.doi.org/10.1007/978-3-540-30586-6_53

8. ANEXOS

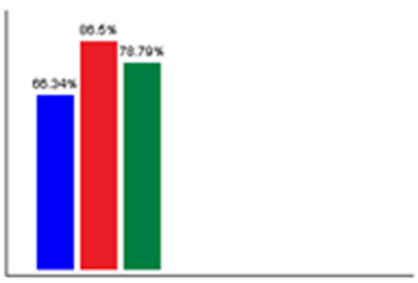
8.1. RESULTADOS OBTIDOS PARA O MODELO ESTATÍSTICO



Relative Frequency Chi Square

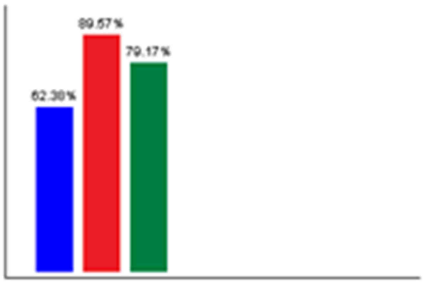


Relative Frequency Information Gain

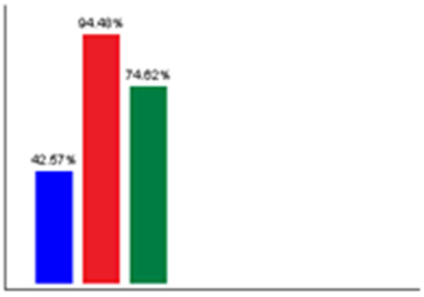


Best feature ranking algorithm for text normalization [Relative Frequency] is [Chi Square]

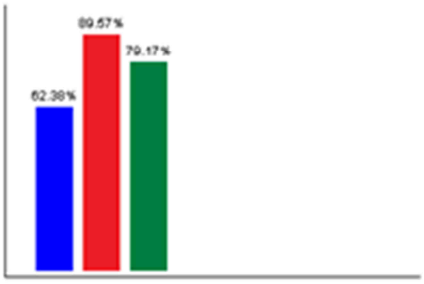
Okapi BM25 No Feature Ranking



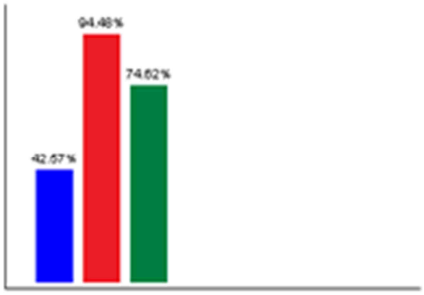
Okapi BM25 Risk Ratio



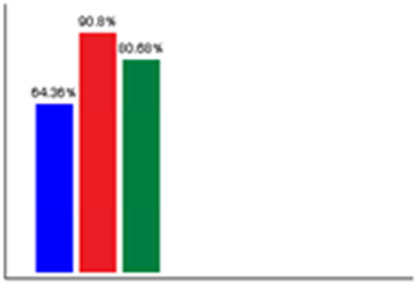
Okapi BM25 Chi Square



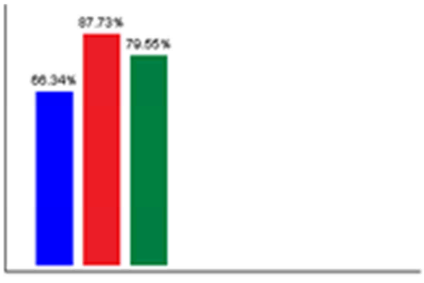
Okapi BM25 Information Gain

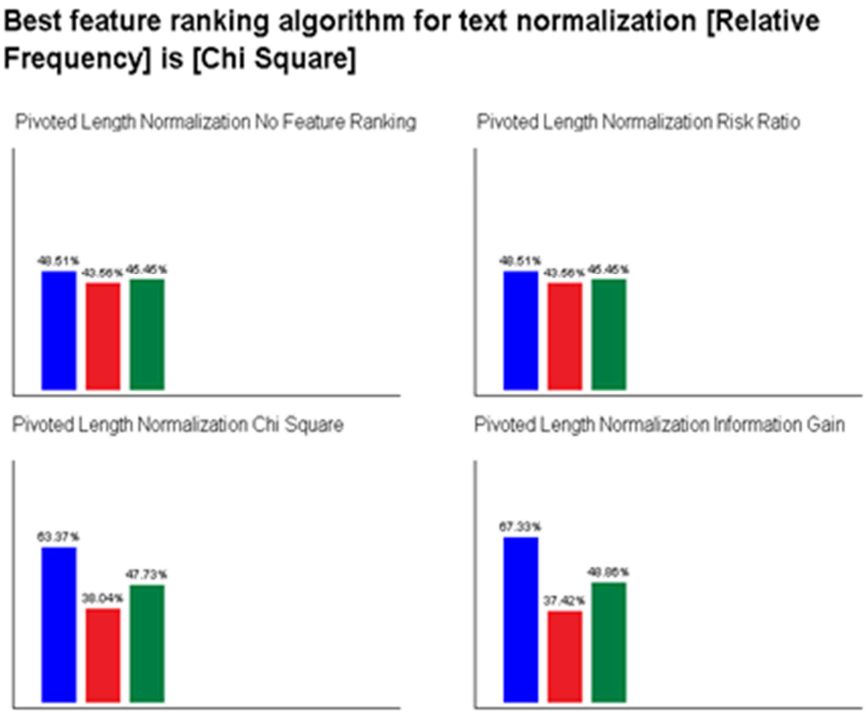


Okapi BM25 Chi Square

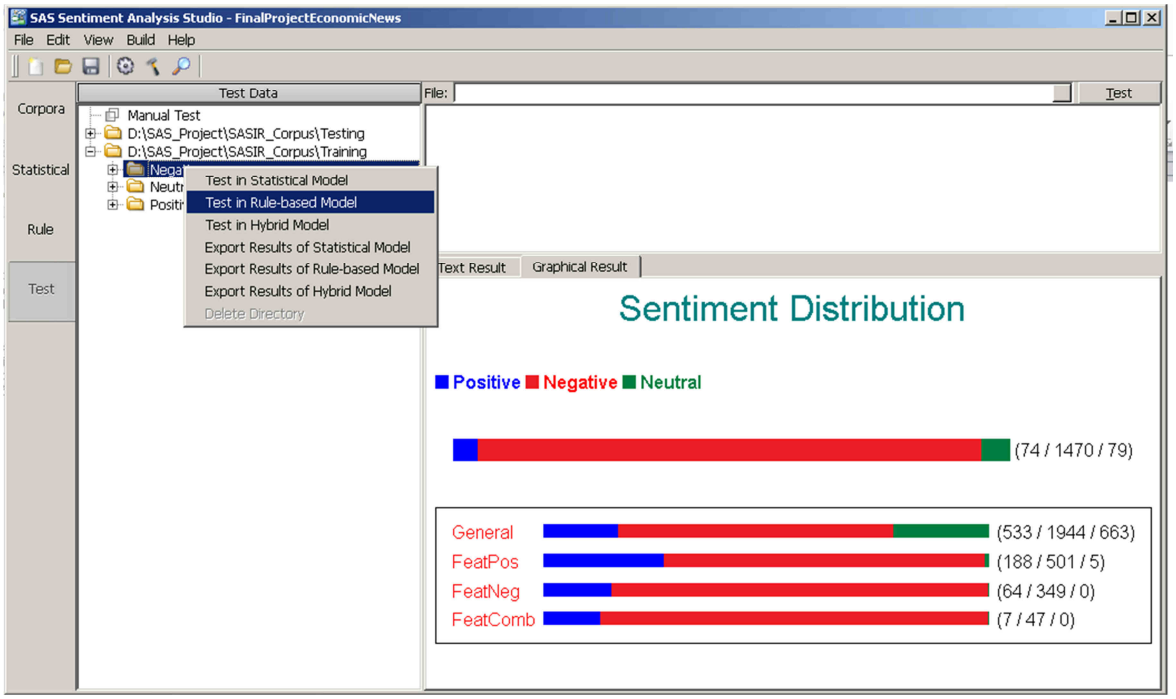


Okapi BM25 Information Gain

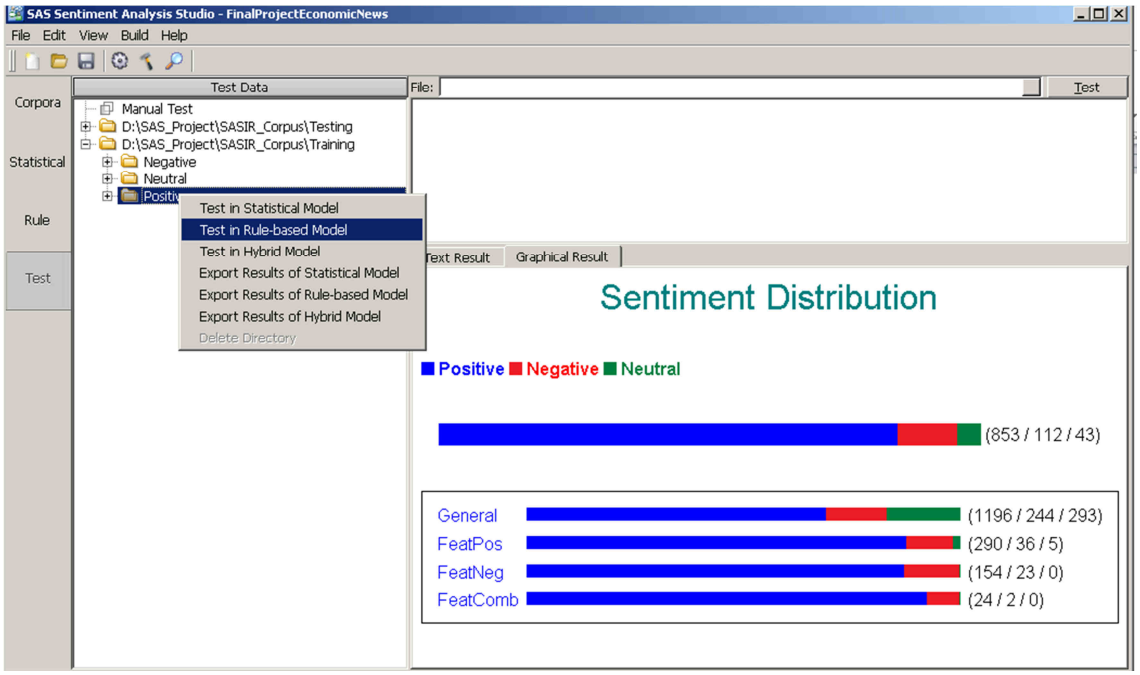




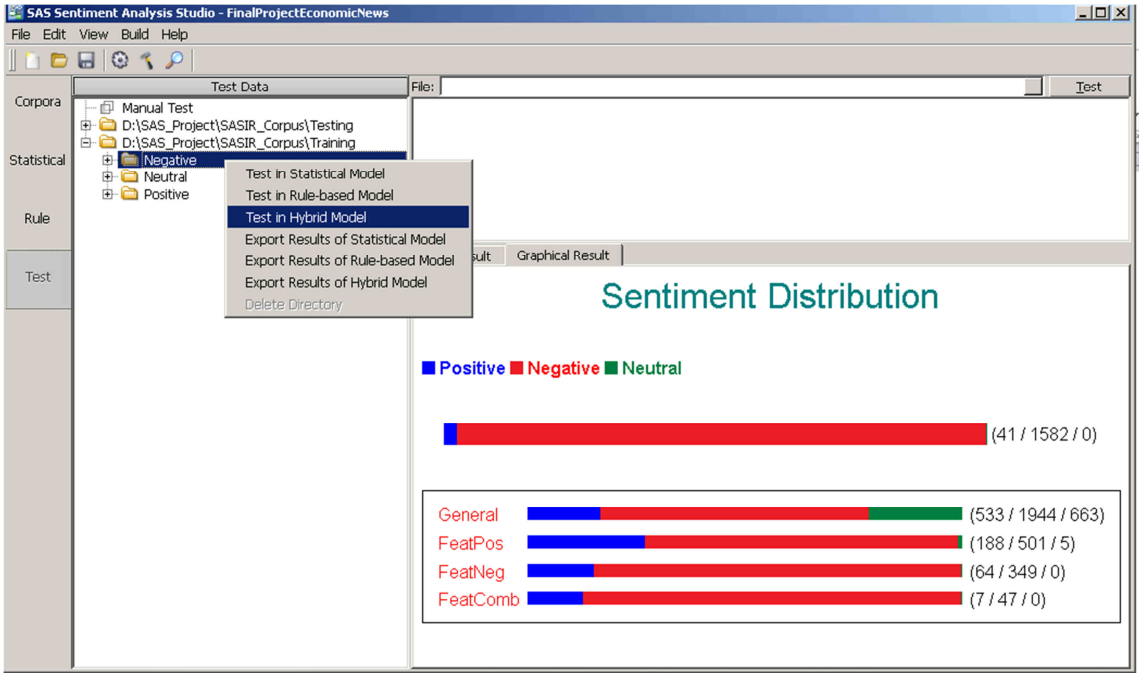
8.2. RESULTADOS PARA A FASE DE DESENVOLVIMENTO DO MODELO BER (NEGATIVAS)



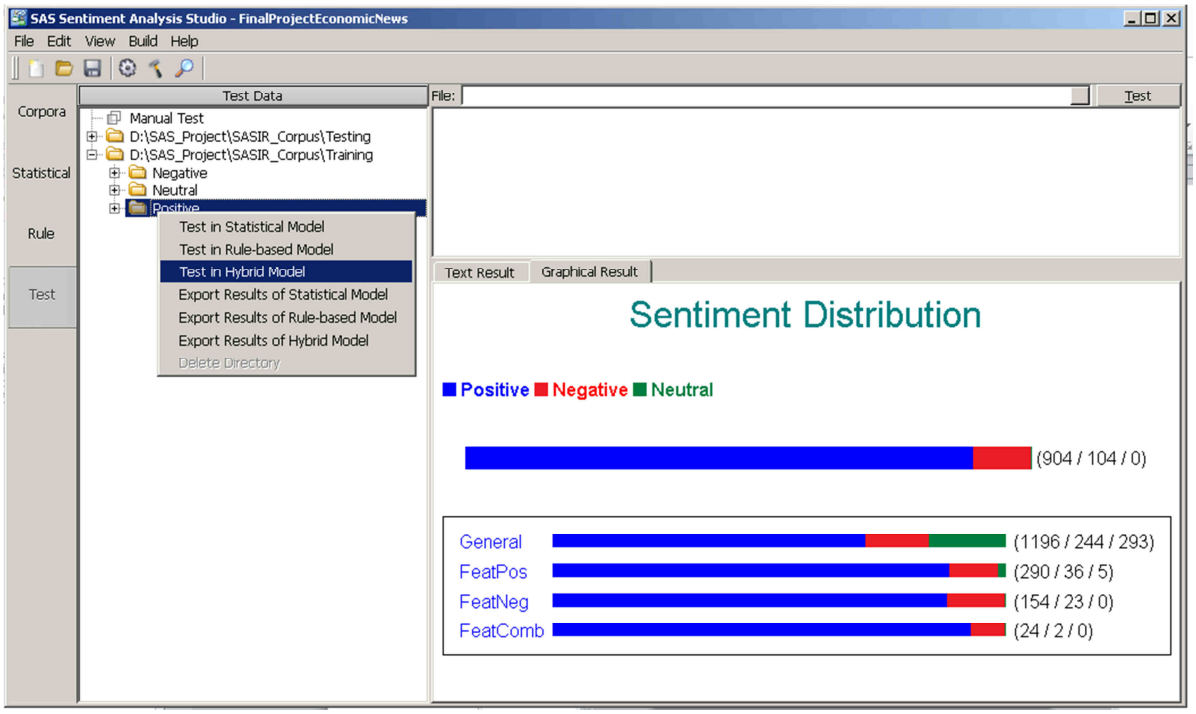
8.3. RESULTADOS PARA A FASE DE DESENVOLVIMENTO DO MODELO BER (POSITIVAS)



8.4. RESULTADOS PARA A FASE DE DESENVOLVIMENTO DO MODELO HÍBRIDO (NEGATIVAS)



8.5. RESULTADOS PARA A FASE DE DESENVOLVIMENTO DO MODELO HÍBRIDO (POSITIVAS)



8.6. ESTRUTURA DE REGRAS IMPLEMENTADA NO MODELO BER