Capítulo 1

Métodos para Análise de Sentimentos em mídias sociais

Fabrício Benevenuto, Filipe Ribeiro, Matheus Araújo

Resumo

Análise de sentimentos tem se tornado um importante tópico na Web, especialmente em redes sociais, com o desenvolvimento de aplicações para monitoramento de produtos e marcas, assim como a análise da repercussão de eventos importantes. Vários métodos e técnicas vêm sendo propostos de forma independente na literatura. Este minicurso oferece uma introdução ao pesquisador que pretende explorar esse tema. Inicialmente, é apresentada uma visão geral sobre análise de sentimentos e suas aplicações mais populares. Em seguida, discute-se os principais métodos e técnicas existentes na literatura, suas características e formas de execução. Finalmente, é feita uma comparação entre estes métodos e apresentando vantagens, desvantagens e possíveis limitações de cada um.

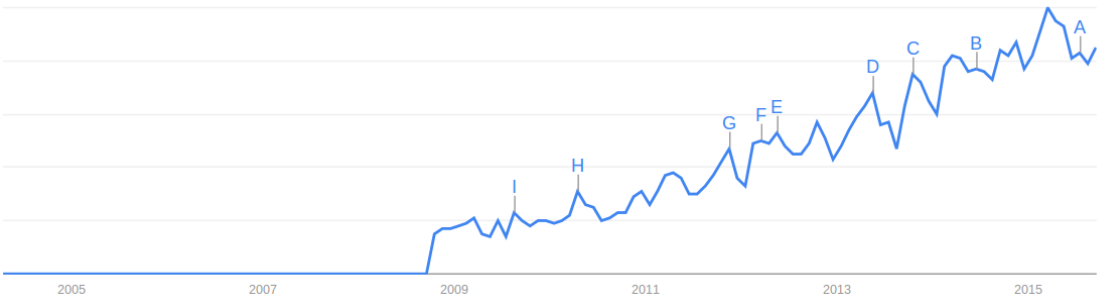

Figura 1.1. Pesquisa do termo "Sentiment Analysis" no Google Trends.

1.1. Introdução

O principal objetivo da análise de sentimentos é definir técnicas automáticas capazes de extrair informações subjetivas de textos em linguagem natural, como opiniões e sentimentos, a fim de criar conhecimento estruturado que possa ser utilizado por um sistema de apoio ou tomador de decisão. A identificação de sentimentos em textos é uma das áreas de pesquisa mais destacadas em Processamento de Linguagem Natural desde o início dos anos 2000, quando se tornou uma área de pesquisa muito ativa [Liu, 2010]. Devido a sua importância para as empresas e para a sociedade, esse tema não tem ficado restrito apenas a uma área da computação e tem atingido outras áreas tais como psicologia e ciências sociais. Nos últimos anos, as atividades industriais que envolvem análise de sentimentos têm apresentado significativo crescimento sendo objeto de desenvolvimento em grandes empresas e ocasionando o surgimento de numerosas start-ups.

A partir da explosão das redes sociais de uso global como o Twitter em 2006, a análise de sentimentos começou a ter um valor social muito importante. A Figura 1.1 endossa tal situação ao exibir o crescimento na quantidade de buscas pelo termo "*Sentiment Analysis*" no buscador Google à partir de 2006. A facilidade para difusão de informações oferecidas pelas redes sociais e seu papel na sociedade moderna representam uma das novidades mais interessantes desses últimos anos, captando o interesse de pesquisadores, jornalistas, empresas e governos. A interligação densa que muitas vezes surge entre os usuários ativos gera um espaço de discussão que é capaz de motivar e envolver indivíduos ilustres ou influentes em discussões, ligando pessoas com objetivos comuns e facilitando diversas formas de ação coletiva. As redes sociais são, portanto, a criação de uma revolução digital, permitindo a expressão e difusão das emoções e opiniões através da rede. De fato, redes sociais são locais onde as pessoas discutem sobre tudo expressando opiniões políticas, religiosas ou mesmo sobre marcas, produtos e serviços.

Opiniões nas redes sociais, se devidamente recolhidas e analisadas, permitem não só compreender e explicar diversos fenômenos sociais complexos, mas também prevê-los. Considerando-se que hoje em dia os atuais avanços tecnológicos permitem o armazenamento e recuperação de enorme quantidade de dados eficientemente, o foco atual está em desenvolver metodologias para extração de informações e criação de conhecimento a partir de fontes de dados distintas. Redes sociais representam um emergente e desafiador setor no contexto da Web. A dificuldade está em extrair opiniões e informações úteis expressas através de mensagens curtas de texto, e assim gerar rapidamente dados que devem

ser analisados de forma eficiente, para finalmente serem utilizados em tomadas de decisões e em processos de tempo real. Esta dificuldade gera um problema multidisciplinar na computação, envolvendo a mineração de dados, o processamento de linguagem natural e a aprendizagem de máquina. A fim de tornar os dados textuais qualitativos efetivamente, a quantificação de "o que as pessoas pensam" torna-se um passo obrigatório, desafiador e de interesse de diversas comunidades científicas.

Neste minicurso pretende-se proporcionar uma ampla visão sobre as técnicas e estratégias existentes para análise de sentimentos, suas principais aplicações no contexto de redes sociais online, além de demonstrar o uso de ferramentas práticas e atuais. Nosso objetivo é estabelecer uma sólida base para alunos e pesquisadores das diversas áreas da computação, em particular, os que possuem interessem em realizar análise e mineração de dados de redes sociais e fazer o uso de técnicas de análise de sentimento. Finalmente, pretende-se identificar as principais ferramentas existentes, apontar seus códigos e também listar as principais bases de dados rotuladas para análise de sentimentos existentes. Acreditamos que isso é fundamental para o pesquisador que pretende começar a explorar o tema.

1.1.1. Terminologias e Conceitos

Diante da recente popularidade desse tema, vários termos e conceitos vêm sendo descritos para tarefas associadas a detecção de sentimentos. A seguir é apresentada cada uma delas:

Polaridade: Representa o grau de positividade e negatividade de um texto. Normalmente esta é a saída para os métodos de análise de sentimentos que serão estudados neste trabalho. Alguns métodos tratam a polaridade como um resultado discreto binário (positivo ou negativo) ou ternário (positivo, negativo ou neutro). Por exemplo, a frase "Como você está bonita hoje" é **positiva** e a frase "Hoje é um péssimo dia" é negativa, já a frase "Hoje é 21 de Outubro" não possui polaridade e normalmente é classificada como **neutra**.

Força do sentimento: Representa a intensidade de um sentimento ou da polaridade sendo também uma forma de saída de alguns métodos. Normalmente é um ponto flutuante entre (-1 e 1) ou até entre $-\infty$ e $+\infty$, muitas vezes tornando necessário o uso de um *threshold* para identificar a neutralidade de uma sentença. Há trabalhos que por exemplo medem a força de sentimentos nos títulos das notícias como o *Magnetic News*[Reis et al., 2014] [Reis et al., 2015b], capaz de separar eficientemente para o usuário notícias boas de notícias ruins .

Sentimento/Emoção: Indica um sentimento específico presente em uma mensagem (ex.: raiva, surpresa, felicidade, etc.). Alguns métodos apresentam abordagens capazes de identificar qual sentimento em específico uma sentença representa. Como exemplo a abordagem léxica Emolex [Mohammad and Turney, 2013], a qual é baseada a partir da avaliação de milhares de sentenças em inglês para 9 sentimentos diferentes: *joy, sadness, anger, fear, trust, disgust, surprise, anticipation, positive, negative*.

Subjetividade vs. Objetividade: Uma sentença objetiva possui normalmente um fato ou uma informação, enquanto sentenças subjetivas expressam sentimentos pessoais e opiniões. Algumas técnicas utilizam a análise da objetividade para estimar se com-

pensa realizar a análise de sentimentos como apresentado em [Feldman, 2013]. Portanto entender se um conjunto de dados possui mais sentenças objetivas ou subjetivas pode influenciar diretamente os resultados. Cabe ressaltar que textos informais (ex.: coletados de redes sociais) tendem a ser mais subjetivos que textos formais (ex.: coletados de notícias).

1.1.2. Frentes de Pesquisa

As frentes de pesquisa nessa área são divididas em diferentes níveis de granularidade conforme a tarefa de detecção de sentimentos nos textos. Quanto menor a granularidade, mais específica é a classificação.

Estado emocional: A identificação do estado emocional a partir de informações contidas no texto é uma das principais frentes de pesquisa. Ferramentas com este foco permitem a empresas acompanharem a satisfação pós-venda de seus usuários sendo um recurso valioso. Um outro exemplo no qual entender o estado emocional pode ser essencial, são pesquisas capazes de caracterizar e prever experiências de depressão pós-parto em mães de recém-nascidos [De Choudhury et al., 2014] através de dados compartilhados no Facebook .

Análise de Sentimentos para comparação ou *Comparative Sentiment Analysis:* Em diversos casos usuários não informam a opinião direta sobre um produto ou pessoa, no entanto, eles fornecem opiniões comparativas em sentenças como "Este computador Apple aparenta ser bem melhor do que aquele Asus","Eu dirijo um carro X, mas a mudança de marcha é bem pior que a do carro Y". O objetivo da análise de sentimentos neste caso é identificar as sentenças que contém as opiniões para ser comparadas (utilizando, por exemplo, advérbios como *pior que*, *melhor que*) e assim extrair a entidade referida daquela opinião [Feldman, 2013].

***Nível de Documento*:** Neste nível de granularidade, a classificação de sentimentos ocorre com a análise de um texto como um todo. Ou seja, nesse nível, assume-se que todo o texto está relacionado a um único assunto que possui certa polaridade. Na prática, se no documento possuir várias entidades com opiniões diferentes, então seus sentimentos podem ser diferentes. Desta forma é difícil assimilar um sentimento ao documento todo, mas um caso interessante em que a análise em nível de documento pode ser utilizado é em *reviews* de produtos ou filmes por exemplo [Liu, 2010].

***Nível de Sentença*:** É neste nível de análise que este trabalho se dedica, pois um único documento pode conter múltiplas opiniões ou mesmo entidades. Neste caso é assumido que o texto foi dividido em frases ou sentenças que possam conter uma opinião individualmente. Cabe ressaltar que, em geral, postagens e comentários em mídias sociais seguem um padrão de sentenças curtas. Quando se pode monitorar as redes sociais, abre-se uma variedade de oportunidades de estudo, um caso interessante é o monitoramento do Twitter para previsão de bolsa de valores [Bollen et al., 2010].

***Nível de Palavra ou Dicionário*:** Nesta frente de pesquisa os trabalhos focam em otimizar os Léxicos de sentimentos existentes na literatura. Não é claro a melhor maneira de se construir um dicionário de sentimentos. No entanto, existem diversos dicionários e suas principais diferenças são constituídas pelas palavras que os formam e às vezes na adição de gírias e acrônimos vindas das redes sociais, como "vc", "blz", "tb". A inclusão

de diferentes termos é importante para alcançar melhor desempenho quando se trabalha com o foco em mídias sociais. Existem outras diferenças entre tais dicionários como a forma que é avaliada a palavra, binária (positivo/negativa) ou proporcional à força do sentimento (-1 a 1) [Nielsen, 2011a].

***Nível de Aspecto*:** Nesse nível de granularidade, uma sentença pode ser julgada por várias entidades e pode conter múltiplos sentimentos associados a ela. Por exemplo, a sentença "Esse hotel, apesar de possuir um ótimo quarto, tem um atendimento péssimo!" possui duas diferentes polaridades associadas a "quarto" e "atendimento" para o mesmo hotel. Enquanto "quarto" é considerado positivo, "atendimento" pode ser analisado de forma negativa. Esta necessidade de avaliar a opinião para cada entidade é comum em reviews de produtos ou em fórums de discussões. O principal do foco deste minicurso se dá na detecção de polaridade no nível de sentença por dois principais motivos. O primeiro deles é a aplicabilidade no contexto de redes sociais em que grande parte dos textos produzidos são sentenças ou textos curtos. Em casos como o do twitter, por exemplo, existe a limitação no número de caracteres postados. Além disso, a análise de sentimentos em sentença é, geralmente, a base para os demais níveis. A análise de sentimento em documentos emprega, frequentemente, a análise de sentimento de menor granularidade destacada neste minicurso para avaliar trechos menores e depois contabilizar o sentimento global do documento.

Os métodos atuais de detecção de sentimentos em sentenças podem ser divididos em duas classes: os baseados em aprendizado de máquina e os métodos léxicos. Métodos baseados em aprendizado de máquina geralmente dependem de bases de dados rotuladas para treinar classificadores [Pang et al., 2002], o que pode ser considerado uma desvantagem, devido ao alto custo na obtenção de dados rotulados. Por outro lado, métodos léxicos utilizam listas e dicionários de palavras associadas a sentimentos específicos. Apesar de não dependerem de dados rotulados para treinamento, a eficiência dos métodos léxicos está diretamente relacionada a generalização do vocabulário utilizado, para os diversos contextos existentes. A seção seguinte apresentará em detalhes as abordagens supervisionadas e léxicas que vêm sendo utilizadas nesse contexto.

1.2. Técnicas Supervisionadas versus Não Supervisionadas

Nesta seção é destacada como diferentes técnicas lidam com os principais desafios oriundos da análise de sentimentos textuais. Existem duas principais abordagens para o problema de extração de sentimentos em textos. A primeira delas é embasada nos conceitos de aprendizagem de máquina partindo da definição de características que permitam distinguir entre sentenças com diferentes sentimentos, treinamento de um modelo com sentenças previamente rotuladas e utilização do modelo de forma que ele seja capaz de identificar o sentimento em sentenças até então desconhecidas. A segunda abordagem não conta com treinamento de modelos de aprendizado de máquina e, em geral, são baseadas em tratamentos léxicos de sentimentos que envolvem o cálculo da polaridade de um texto a partir de orientação semântica das palavras contidas neste texto.

1.2.1. Técnicas Supervisionadas

A primeira abordagem composta por técnicas supervisionadas emprega o termo supervisionado justamente pelo fato de exigir uma etapa de treinamento de um modelo com amostras previamente classificadas. O procedimento para realizar a aprendizagem de máquina compreende quatro etapas principais: 1 - obtenção de dados rotulados que serão utilizados para treino e para teste; 2 - definição das *features* ou características que permitam a distinção entre os dados; 3 - treinamento de um modelo computacional com um algoritmo de aprendizagem; 4 - aplicação do modelo.

1.2.1.1. Dados rotulados

O dado rotulado necessário na etapa 1 descrita acima nada mais é do que uma entrada com seu respectivo rótulo ou classificação. No caso da análise de sentimentos seria uma sentença acompanhada de sua polaridade. Para desenvolver um bom modelo supervisionado é necessário, dentre outras coisas, que uma amostragem substancial do domínio do problema esteja disponível e previamente rotulada, seja para gerar o conjunto de dados para treino ou para testes.

Esta é uma das grandes dificuldades do aprendizado de máquina especialmente no que diz respeito à análise de sentimentos por dois motivos principais: a alta subjetividade envolvida na tarefa e a demanda de tempo necessária para que especialistas definam a polaridade de muitas sentenças. Muitas sentenças são altamente ligadas a algum situação ou evento específico e a definição da polaridade pode ser extremamente difícil por quem não esteja inserido no contexto em questão. Imagine uma situação em que avaliadores desejam definir a polaridade de *tweets* relativos ao debate presencial de um país ou estado do qual não possuem nenhum conhecimento da situação política. Certamente haverá grande dificuldade e possibilidade de rotulações incorretas. Situações com utilização de sarcasmo e ironia também tornam a avaliação uma tarefa complexa. Soma-se a isso a necessidade de que muitas sentenças sejam rotuladas, o que poderia demandar semanas de trabalho de um especialista.

Algumas alternativas para geração de dados rotulados têm sido adotadas e são descritas a seguir:

- Distant Supervisor: É uma técnica que aborda a utilização de características pre-existentes no texto para rotulação automática. Uma das formas mais utilizadas é definir a polaridade das frases de acordo com emoticons existentes associando o sentimento representado pelo emoticon à frase como um todo [Hannak et al., 2012]. Por exemplo, uma frase contendo =) seria classificada como positiva. Pesquisadores de stanford disponibilizaram uma base de dados com mais de um milhão de *tweets* classificados desta maneira [Go et al., 2009a]. Note que esta abordagem de classificação não é 100% confiável e pode introduzir ruídos no modelo treinado.

- *Amazon Mechanical Turk(AMT)* [1]: Este é um sistema criado pela *Amazon* que permite pagar para que pessoas realizem pequenas tarefas remotamente. Muitos pesquisadores têm usado tal sistema para disponibilizar diversas sentenças e criar tarefas de classificação de polaridade que serão executadas pelos *Turkers* por valores razoavelmente baixos. Os experimentos de classificação que utilizam o "Conhecimento das multidões" a partir do AMT tem sido utilizado em diversos trabalhos para a geração de bases de dados rotuladas.

Mesmo com tal dificuldade na geração deste tipo de dados, diversos esforços foram realizados e encontram-se disponíveis atualmente muitos datasets rotulados de contextos variados. Na seção 1.3 de comparação eles serão detalhados e também estão disponibilizados [2].

1.2.1.2. Definição de features

A tarefa de definição das características que permitirão classificar os dados é também de suma importância para a criação de um classificador eficiente. Denominadas *features*, essas características devem ser atributos que permitam uma boa distinção entre o conjunto de dados a serem classificados. Imagine um sistema que deseja identificar uma fruta dentre três possíveis: limão, goiaba e melancia. As *features* seriam quaisquer características inerentes às frutas que permitam distingui-las tais como: peso, diâmetro, cor da casca, sabor (doce, azedo, etc.). Note que a feature cor da casca não seria boa para o exemplo proposto já que as três frutas possuem a mesma cor, que é verde. Obviamente este é um exemplo hipotético e simples, no entanto, o princípio da qualidade de escolha das *features* deve ser aplicado em exemplos mais complexos.

No que se refere à categorização de textos, o conjunto de features é composto, em geral, pelas próprias palavras presentes no texto. Uma representação extremamente simples é chamada de bag of words, segundo a qual cada sentença é representada por grande array de 0's e 1's sendo que cada coluna indica a presença (1) ou não (0) de uma palavra. Note que o array representa um mapeamento para um dicionário de palavras onde cada posição no array representa uma palavra.

A abordagem bag-of-words mais simples é chamada de unigrama o que indica que cada palavra representa uma *feature*. Algumas variações porém, permitem que cada

[1] http://www.mturk.com

[2] `http://homepages.dcc.ufmg.br/~fabricio/benchmark_sentiment_analysis.html`

feature possa ser representada por um grupo de palavras (bigrama, trigrama, etc. Veja os três exemplos a seguir:*1 - The restaurant is good!*; *2 - The restaurant is not good!*; *3 - The restaurant is very good!*

Imagine que um classificador irá utilizar cada palavra como uma *feature*. A *feature 'good'* possivelmente será relacionada às frases positivas já que, em geral, *'good'* é uma palavra com conotação positiva. No entanto, *good* ocorre na segunda frase *'The restaurant is not good'* e não tem uma conotação positiva pois é precedida pela palavra *'not'*. Por outro lado, a terceira frase *'The restaurant is very good'* possui *'good'* precedida pela palavra *'very'* o que enfatiza ainda mais o adjetivo *'good'*. Em uma representação unigrama da segunda frase as features seriam: *'the', 'restaurant', 'is', 'not', 'good'*. Já em uma representação bigrama as features seriam grupos de duas palavras: *'the restaurant','restaurant is', 'is not', 'not good'*. Note que com a representação Bigrama, a dupla *'not good'* seria uma feature que possivelmente estaria associada a sentenças negativas.

Outra abordagem utilizada para representação de *features* no contexto de análise de sentimentos é chamada *TF-IDF*, que é uma evolução a partir do IDF que é proposto pelo trabalho de Karen Jones [Jones, 2004], cuja intuição básica é que um termo que ocorre em muitos documentos não é um bom discriminador, e deve ser dado menos peso do que aquela que ocorre em alguns documentos. A fórmula para aplicação dessa metodologia é a seguinte:

$$w_{ij} = tf_{ij} \cdot log(\frac{N}{df_i}) \tag{1}$$

Onde w_{ij} é o peso para o termo i na sentença j, N é o número de sentenças na coleção, tf_{ij} é a frequência do termo i na sentença j e df_i é a frequência do termo i na coleção de sentenças.

É importante ressaltar que faz-se necessária a realização de pré-processamento nos dados que serão utilizados pela técnica de Aprendizado de Máquina. Especialmente no que diz respeito ao tratamento de texto o passo de pré-processamento é extremamente importante pois irá eliminar *features* que não agregam muita informação. Em geral, trabalhos com processamento textual realizam a remoção de *stop words* do texto, que é uma lista conhecida e que não agrega muito no que diz respeito ao sentimento de uma sentença. Dentre as *stop words* da língua inglesa estão palavras como: *the*, *a*, *about*, , etc.

Outras ocorrências textuais também podem ser tratadas como *features* e podem eventualmente apresentarem melhoras nos classificadores. Palavras alongadas (*'goooood'*), pontuação repetida ('!!!!') e ocorrência de emoticons positivos ou negativos são alguns dos exemplos.

Dentre as diversas abordagens supervisionadas para a análise de sentimento, o SVM tem sido umas das mais utilizadas devido, principalmente, aos resultados promissores que tem apresentado. A seguir, esta abordagem será apresentada em um maior nível de detalhe.

1.2.1.3. Exemplo de uso de um classificador supervisionado - SVM

SVM, que em tradução literal quer dizer Máquinas de Vetores de Suporte Vetores (*SVM - Support Vectors Machine*) é uma abordagem de aprendizagem introduzida por Vapnik em 1995 [Cortes and Vapnik, 1995]. O método difere dos de outros classificadores por ser uma classe de classificadores probabilísticos cujo problema é encontrar uma superfície de decisão que "melhor" separa os dados em duas classes. Para espaço linearmente separáveis, a superfície de decisão é um hiperplano que pode ser escrito como:

$$wx + b = 0$$

onde X é um objeto arbitrário para ser classificado, o vetor w e a constante b são aprendidas a partir de um conjunto de treino de objetos linearmente separáveis.

O método SVM está entre os principais na área de aprendizado de máquina e conta com aplicações em diversos domínios tais como: análise de imagens, categorização de textos, bioinformática e outros [Lorena and De Carvalho, 2008]. Como diversos problemas comuns de aprendizado de máquina não possuem uma separação linear entre os dados foi proposta a utilização de kernels, que são, de maneira simples, funções que multiplicadas ao conjunto de dados aumentam a dimensionalidade e permitem melhor separação entre as entradas. Os Kernels existentes não serão detalhados uma vez que tal assunto foge do escopo proposto. Um importante parâmetro na construção de classificadores SVM é o parâmetro C que indica de maneira simplista o quanto se deseja evitar classificações erradas no treino.

A seguir iremos detalhar o desenvolvimento de um método de aprendizado de máquina que utiliza SVM proposto por [Mohammad et al., 2013] para análise de sentimentos. A ideia principal com este exemplo é destacar as principais features utilizadas para a construção de um modelo com alta precisão treinado e testado em datasets de *twitter*.

Este método utiliza um kernel linear com parâmetro C=0,005 de forma que a margem não seja muito grande. Como os dados utilizados provêm do Twitter, é realizado um préprocessamento a fim evitar ruídos. Este pré-processamento normaliza todas as URLs para http://someurl e todas as ID'S de usuários para @algumusuario. Uma das features mais importantes são as partes do discurso contidos na sentença, para extrair essa e outras características inerentes do texto é necessário utilizar uma ferramenta para realizar um processamento de linguagem natural nos textos e assim identificar partes do discurso de cada sentença [3]

Dentre as diversas *features* utilizadas pelo autor no treinamento de seu modelo destacam-se as seguintes. Ocorrência de termos (unigramas e multigramas) presentes em dicionários léxicos. Soma do pesos dos *tokens* presentes na sentença de acordo com 5 dicionários léxicos. Número de tokens com exclamação ou pontos de interrogação contínuos (ex: *"cool!!!!!"*). Presença de emoticons positivos e emoticons negativos. Número de palavras alongadas (ex: *"looooove"*). Número de termos precedidos por uma negação. Número de ocorrências de palavras com todos os caracteres em maiúsculo.

[3] Existem diversas bibliotecas disponibilizadas, como o NLTK para a linguagem Python.

Este SVM foi treinado em 9.912 tweets rotulados por humanos, sendo que 8.258 foram utilizado patra treinamento e 1.654 para a criação dos dicionários léxicos. Outros 3.813 tweets não vistos pelo algoritmo previamente foram utilizados para testes. Os autores revelam que as *features* relacionadas à presença de multigramas e da pontuação segundo o dicionário léxico apresentaram o maior ganho de informação, ou seja, foram as características que mais permitiram separar os tweets positivos dos neutros e negativos. A métrica *F-Score* alcançada pelo modelo foi de 0,69 segundo os autores.

1.2.1.4. Considerações sobre Aprendizado de Máquina

Uma importante consideração a ser destacada a respeito de técnicas de aprendizado de máquina é que neste tipo de estratégia o modelo gerado pode ir muito bem nos conjuntos de dados para o qual ele foi treinado fazendo com que resultados sejam razoáveis no treino mas no momento de testes os resultados apresentem resultados bem diferentes. Tal situação, conhecida como *overfitting*, deve ser evitada e existem metodologias que devem ser seguidas para que isto não aconteça. No entanto, isto foge do escopo do minicurso.

As técnicas de aprendizado apresentam algumas dificuldades que serão descritas a seguir. A primeira delas diz respeito à aplicabilidade do modelo que, em geral, são bem restritos ao contexto para o qual foram criados. Outro ponto é a necessidade de boa quantidade dados para treinamento, previamente validados, muitas vezes de difícil obtenção. A escolha dos dados para treinamento deve ainda ser cuidadosa pois caso sejam mal escolhidos podem criar um viés muito grande no modelo tornando-o tendencioso a dar como saída uma classe específica. Além disso, a abordagem supervisionada pode ser computacionalmente caro em termos de processamento da CPU e memória gerar o modelo da aprendizagem. Esta característica pode restringir a capacidade de avaliar um sentimento em dados de streaming por exemplo. Por fim, algumas características utilizadas para alimentar a aprendizagem de máquina são derivadas de algoritmos que geram um modelo dificilmente interpretável por seres humanos. Isto torna os modelos difíceis de generalizar, modificar ou estender (para outros domínios por exemplo) [Hutto and Gilbert, 2014a].

É importante ressaltar que outros detalhes característicos do aprendizado de máquina não foram enfocados neste minicurso como divisão de dados rotulados para treino e teste, detalhes de parâmetros, etc. Recomenda-se o estudo mais detalhado a respeito da metodologia de desenvolvimento de soluções com Aprendizado de Máquina para o desenvolvimento de um novo classificador.

1.2.1.5. Outras Abordagens Supervisionadas

Deep Learning é uma abordagem da área de aprendizado de máquina responsável pela criação de modelos abstratos complexos criados a partir de diversas transformações não lineares. Por outro lado as redes neurais artificiais são modelos computacionais inspirados em nosso cerébro sendo capazes de realizar aprendizado de máquina assim como tarefas para reconhecimentos de padrões. Embasadas nestas duas áreas, pesquisas recentes utilizam *Deep neural networks* para criar modelos bem precisos para identificação de polaridades nos textos, seus modelos ensinam as máquinas hierarquicamente e contex-

tualmente, permitindo que o conhecimento seja dividido em varias e então processados em camadas. Um exemplo é o trabalho gerado por [Severyn and Moschitti, 2015], suas abordagens se sobressaíram ao comparar suas propostas com a de outros participantes no SemEval 2015 [4].

1.2.2. Técnicas não supervisionadas

As técnicas não supervisionadas, diferentemente das supervisionadas, não carecem de sentenças previamente rotuladas e treinos para a criação de um modelo. Esta é uma das suas principais vantagens uma vez que desta forma não mantem aplicação restrita ao contexto para o qual foram treinados. Dentre as técnicas não supervisionadas destacam-se aquelas com abordagens léxicas, baseadas em um dicionário léxico de sentimento, uma espécie de dicionário de palavras que ao invés de possuir como contéudo o significado de cada palavra, possui em seu lugar um significado quantitativo (i.e. pode ser um número entre -1 a 1, onde -1 é o valor sentimental mais negativo e 1 o valor mais positivo) ou mesmo valor qualitativo (i. e. positivo/negativo, feliz/triste). Abordagens léxicas assumem que palavras individuais possuem o que é chamado de polaridade prévia, que é, uma orientação semântica idependente de contexto e que pode ser expressada com um valor numérico ou classe [Taboada et al., 2011].

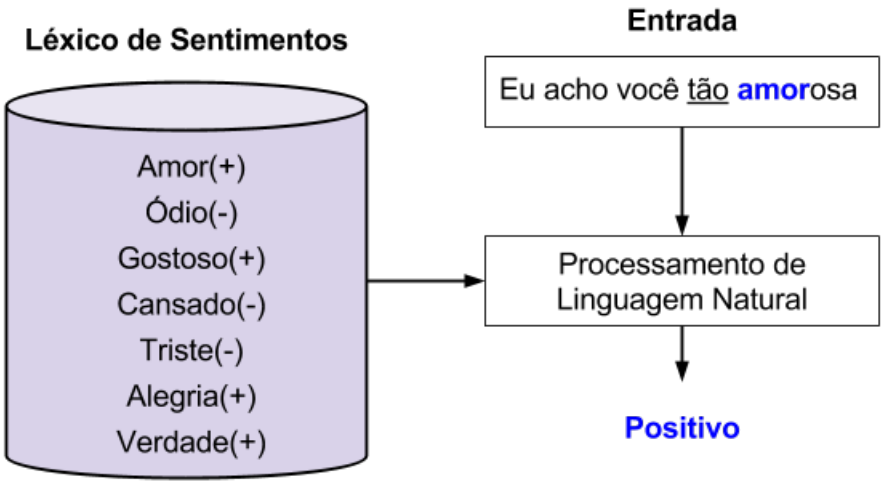

Figura 1.2. Léxico de sentimentos

A análise de sentimentos baseadas em abordagens léxicas é atualmente uma das estratégias mais eficientes, seja na utilização de recursos computacionais, seja em capacidade de predição. Ela se baseia na utilização de um grande dicionário de termos, onde cada termo está associado a um sentimento. A Figura 1.2 generaliza o funcionamento de um método de análise de sentimentos léxico. O processo de classificação inicia quando o método recebe uma sentença de entrada, em seguida é realizado um processamento de linguagem natural, assim como uma pesquisa no léxico dos termos que formam esta mensagem. Ao final do processo o método é capaz de inferir qual é a polaridade ou sentimento implícito na sentença de entrada.

[4] http://alt.qcri.org/semeval2015/

Tabela 1.1. Dicionários Léxicos de Sentimentos na Literatura.[Hutto and Gilbert, 2014a]

Binário Positivo/Negativo	Intensidade do Sentimento
LIWC	ANEW
GI (General Inquirer)	SentiWordNet
Opinion Lexicon	SenticNet

Os dicionários apresentados na tabela 1.1 são clássicos na literatura, sendo que muitos trabalhos recentemente publicados tem o objetivo de combinar, estender ou melhorá-los. Eles estão divididos na forma como foram classificados originalmente, sendo de forma binária (positivo vs. negativo) ou a partir da intensidade do sentimento (entre -1 e 1). Esta diferença ocorre pois muitas aplicações se beneficiam caso forem capazes de não somente determinar a polaridade binária ou ternária (postiva, negativa ou neutral), mas também a força de um sentimento expressado no texto. A possibilidade de definir a intensidade da polaridade permite, por exemplo, detectar mudanças na intensidade do sentimento ao longo do tempo, o aquecimento ou esfriamento de um tema além de outras aplicações. [Hutto and Gilbert, 2014a].

1.2.2.1. Construção de dicionários

Um dos principais desafios deste tipo de técnicas é a construção de um dicionário léxico abrangente. O incontável volume de conteúdo textual produzido na Web diariamente vai desde publicações jornalísticas e artigos acadêmicos com linguagem extremamente formal até reviews sobre produtos e postagens em redes sociais com palavreado informal, muitas vezes contendo gírias, conteúdo jocoso e palavras de baixo calão. Esta dicionário quase infinito, utilizado todos os dias na Internet dificulta a criação de um dicionário léxico amplo. Soma-se a isto o surgimento diário de neologismos e *hashtags* criadas nas redes sociais gerando um volume infindável de palavras.

É importante ressaltar que, em se tratando de abordagens léxicas é essencial uma etapa de pré-processamento sobre o conteúdo textual tratado. Existe uma linha de pesquisa cujo estudo foca exatamente neste aspecto, chamada de Processamento de Linguagem Natural (PLN) e envolve basicamente o estudo e a compressão por computadores de como humanos naturalmente falam, escrevem e se comunicam. Obviamente os computadores não possuem a capacidade de interpretação que os humanos e necessitam de algoritmos precisos e não ambíguos para serem capazes de realizar tal tarefa.

Com métodos de PLN, computadores quebram peças de texto em elementos gramaticais. Como exemplo considere a seguinte sentença: “The amazing Cloud delivers data to me ASAP”. O processamento computacional divide as palavras em elementos gramaticais (“amazing” = adjetivo; “cloud” = substantivo; “delivers” = verbo), compreende que “cloud” referencia “cloud computing” e reconhece “ASAP” como um acrônimo para “As Soon As Possible”. A tarefa de quebra o texto em elementos gramaticais é chamada de *POS (Part-of-Speech)* e pode ser realizada por uma série de bibliotecas em diversas linguagens de programação tais como: NLTK para Python [nlt, 2015] e a bibli-

oteca Java criada por pesquisadores de Stanford [std, 2015]. Estas informações sobre o texto são fundamentais para análises texuais posteriores e para a análise de sentimento especialmente pelo fato de que a mudança na característica gramatical de uma palavra pode mudar o significado e a intensidade do sentimento envolvido na mesma.

Outros pré-processamentos também são de extrema importância nas abordagens léxicas tais como o *stemming* e a remoção de *stopwords*. O *stemming* consiste em obter a raiz de uma palavra. Por exemplo, a palavra "working", "work" e "worked" não terão diferenças de polaridade e como forma de reduzir a quantidade de palavras do dicionário podem ser reduzidas à raiz "work". Muitas palavras presentes nas sentenças não agregam informações importantes no que diz respeito ao sentimento de uma sentença. Tais palavras chamadas de *stop words* são facilmente encontradas na Internet entre elas estão palavras como como: *the*, *a*, *about*, *some*, etc.

A criação e validação manual de "significados sentimentais" para palavras está entre as formas mais robustas e confiáveis de construção de dicionários léxicos no entanto é também uma das maneiras que mais demanda tempo [Hutto and Gilbert, 2014a]. Assim sendo, muitas abordagens utilizam geração automática ou semi-automática e, na maioria dos casos utilizam como ponto de partida dicionários já criados anteriormente. Outra forma de criação de dicionários léxicos que vem se popularizando conta com o auxílio do "conhecimento das multidões" como já mencionado anteriormente e usa ferramentas como o *Amazon Mechanical Turk*.

- LIWC [Tausczik and Pennebaker, 2010] - O LIWC é uma ferramenta bem estabelecida e utilizada em diversas áreas, e contou com o aval de psicólogos, sociólogos e linguistas durante seu desenvolvimento. Ela possui um dicionário léxico de aproximadamente 4500 palavras e raízes de palavras, fazendo parte de oitenta categorias. Todas as palavras de raiva por exemplo, estão incluídas em pelo menos duas categorias que são emoções negativas e emoção geral. As categorias foram definidas e populadas com base em buscas em diversos dicionários como Thesaurus, questionários e listas feitas por pesquisadores. Para a correta definição, três juízes independentes julgaram cada uma das palavras e definiram em qual categoria ela deveria estar. O percentual de acordo entre os juízes ficou entre 93 e 100%. A construção do dicionário começou entre 1992 e 1994 e ainda sofre alterações e incrementos nos dias de hoje.

- *General Inquirer (GI)* [Stone et al., 1966] - O General Inquirer é um dos dicionários léxicos manualmente criados mais antigos que ainda é usado amplamente. O GI foi desenvolvido e refinado na década de 60, e foi desenvolvido como uma ferramenta para a análise de conteúdo, técnica usada à época por cientistas sociais, cientistas políticos e psicólogos para objetivamente identificar características específicas de mensagens. O dicionário contém mais de 11000 palavras classificadas em uma ou mais das 183 categorias. Dentre o total de palavras, 1915 foram rotuladas como positiva e 2291 como negativas.

- *Opinion Lexicon* - [Hu and Liu, 2004] diponibilizaram publicamente um léxico de aproximadamente 6800 palavras, sendo 2006 com orientação semântica positiva e 4783 com negativa. O dicionário foi inicialmente construído a partir do Wordnet,

uma base de dados de palavras em inglês em que elas são agrupadas em conjuntos de sinônimos (*synsets*). O *Opinion Lexicon* e, diferentemente, dos dois anteriores está mais atualizado para lidar com expressões em redes sociais e reviews de produtos.

- *ANEW* [Bradley and Lang, 1999] - as Normas Afetivas para Palavras em inglês, do acrônimo em inglês *ANEW - Affective Norms for English Words*, provê uma pontação para 1034 palavras em inglês. Diferentemente do LIWC, as palavras no ANEW foram *ranqueadas* em termos de *prazer (pleasure)*, excitação *(arousal)* e *dominância (dominance)* sendo que cada palavra possui um valor de 1 a 9 para cada categoria. Uma extensão deste dicionário é o AFFIN criado a partir do ANEW mas incrementando elementos léxicos provenientes de redes sociais.

- *Sentiwordnet* [Esuli and Sebastiani, 2006] - o *Sentiwordnet* é uma extensão do Wordnet no qual 147307 conjuntos de sinônimos estão anotados com três pontuações numéricas para positividade, negatividade e objetividade (neutralidade). Cada pontuação varia de 0.0 a 1.0 e a soma é 1.0 para cada conjunto de sinônimos. A pontuação foi calculada utilizando-se vários algoritmos semi-supervisionados. Este dicionário não 100% validado por humanos como os anteriores mas é extremamente útil em uma gama de tarefas.

- *SenticNet* [Cambria et al., 2010] - *SenticNet* é um dicionário semântico e afetivo para opinião em nível de conceito e análise de sentimento. Ele foi construído através do que é denominado pelos autores de *sentic computing*, um pardigma que explora Inteligência Artificial e técnicas de Web semântica para processar opiniões via mineração de grafos e redução de dimensionalidade. O dicionário consiste de 14244 conceitos de senso comum como *adoration* e *admiration* com informações associadas à polaridade de sentimento em uma escala contínua de -1 até 1.

- *Sentiment140* [Mohammad et al., 2013] e *NRC Hashtag* [Mohammad, 2012] - estes dois dicionários foram gerados automaticamente de maneira similar à técnica descrita anteriormente chamada de *Distant Supervised*. Automaticamente rotulou-se milhares tweets com base na ocorrência de emoticons positivos ou negativos (*Emolex*) ou na ocorrência de hashtags com palavras positivas como e negativas como (*NRC Hashtag*). À partir da rotulação automática verificou-se, com contagens relativamente simples, quais palavras ocorriam com maior frequência em tweets positivos ou tweets negativos. Essa metodologia proporcionou a criação de dois dicionários, o Sentiment140 com mais de 1 milhão de termos e o NRC Hashtag com mais de 800 mil termos, em ambos os casos somandos-se unigramas, bigramas e termos que ocorrem em pares separados.

- *VADER* [Hutto and Gilbert, 2014a] - para a construção de um dicionário léxico os autores do VADER criaram uma lista de palavras com base em dicionários já bem estebelecidos como LIWC, ANEW e GI. Em seguida, foram adicionadas numerosas construções léxicas presentes em microblogs tais como emoticons, acrônimos e gírias que expressam sentimentos, resultando em 9000 novos candidatos a serem incluídos no dicionário. Em seguida, verificou-se por meio do "conhecimento das multidões" (*AMT*) quais destas possuíam realmente aplicabilidade através de

pontuação atribuída por *Turkers* variando de -4 (extremamente negativa) a 4 (extremamente positiva). Por fim, restaram 7000 construções léxicas no dicionário sendo que para a inclusão a média entre as avaliações obtidas com o AMT deveria ser diferente de 0 (neutro) e com desvio padrão entre as pontuações abaixo de um limiar estabelecido pelos autores. A palavra “okay” por exemplo possui uma pontuação de 0.9, enquanto “great” tem o valor 3.1 e “horrible” é -2.5.

Faz-se importante ressaltar que um dicionário léxico por si só não é capaz de prover classificação de sentenças de maneira eficaz e o simples somatório da pontuação de cada uma das palavras pode apresentar resultados pouco satisfatórios. Vários métodos, no entanto, se baseiam nas polaridades prévias definidas por dicionários léxicos em conjunto com outras heurísticas e processamentos que proporcionam eficiência na predição do sentimento global das sentenças.

Em geral, as heurísticas baseiam-se em pistas gramaticais e sintáticas que mudam a intensidade do sentimento e vão além da simples soma de pontuações. Dentre elas destacam-se: 1) Pontuação (número de exclamações ao final de uma frase por exemplo), 2) Capitalização (“I HATE YOU”, por exemplo, é mais intenso do que “i hate you”), 3) Negações e Intensificadores (“The service here is not good” é negativo enquanto "The service here is very good"é muito positivo), 4) Conjunção "but"que muda polaridade, 5) Trigrama para identificar negações mais complexas (por exemplo “The food here isn’t really all that great.”)

Alguns métodos que utilizam esta abordagem são o Umigon, o VADER e o SO-CAL que serão apresentados com mais detalhes na próxima seção. Com o intuito de aplicar os demais dicionários na análise de sentimentos foram desenvolvidos métodos que aplicam as mesmas heurísticas empregadas pelo método VADER. Os resultados dos experimentos são detalhados na seção 1.4

1.3. Apresentação de Métodos para Análise de Sentimentos no nível de sentença

Esta seção apresenta vinte um dos principais métodos para a análise de sentimentos disponibilizados na literatura. Foi realizada uma extensa busca na literatura a fim de identificar quais métodos eram capazes de detectar polaridades em nível de sentença e especialmente aqueles que tinham código-fonte disponível ou possibilidades de implementação à partir do artigo. Durante a realização desta tarefa diversos autores foram contactados e os códigos dos métodos requisitados a fim de abranger as diferentes técnicas já discutidas. Os códigos desenvolvidos e obtidos dos autores foram disponibilizados na página `http://homepages.dcc.ufmg.br/~fabricio/benchmark_sentiment_analysis.html`.

É importante ressaltar que tais abordagens têm sido utilizadas como métodos de prateleira (*off-the shelf*), isto é, pesquisadores e demais usuários interessados em aplicar a análise de sentimento para algum propósito específico escolhem alguma das soluções disponíveis e aceitas na literatura e aplicam para o fim desejado. Mesmo no caso dos métodos supervisionados, verifica-se a utilização do modelo previamente treinado com a base de dados original e não após realização de uma nova etapa de treinamento.

A tabela 1.2 apresenta uma descrição a respeito de cada um dos métodos e se eles atuam de maneira Supervisionada (S) ou não supervisionada (NS). Outras iformações como as saídas do método (por exemplo, -1,0,1, significando negativo, positivo e neutro, respectivamente), os datasets utilizado para validá-lo e outros métodos com os quais foram comparados estão na tabela 1.3. Os métodos foram apresentados em ordem cronológica para visão geral dos esforços existentes ao longo dos anos. Algums métodos foram ligeiramente modificados para se adequar ao padrão de saída para detecção de polaridade. Particularmente a implementação do OpinionFinder gera saída (-1,0,1) para cada pista de sentimento encontrada na sentença sendo que cada sentença tem mais de uma pista. Para o cálculo da polaridade global foram somadas as polaridades das pistas. Também foram atribuídos valores de polaridade para métodos cuja saída consistia de um conjunto de sentimentos como o PANAS. Para o Happiness Index, os valores 1 a quatro foram considerados negativos, o valor 5 neutro, e os valores 6 a 9 positivos. Diversos outros métodos também tiveram adaptações similares. Métodos baseados em aprendizado de máquina foram utilizados de maneira não supervisionada, isto é, utilizou-se um modelo previamente treinado para decodificar as sentenças desejadas.

Pode-se perceber pela tabela 1.3 que muitos dos métodos criados não seguem um padrão para avaliação de seus resultados deixando uma lacuna quanto à real eficiência da abordagem recém criada em comparação com abordagens mais antigas.

Tabela 1.2. Métodos para análise de sentimentos em sentença - Descrição

Nome	Descrição	NS	S
Emoticons	Possui uma lista de emoticons dividida em positivos(":)") e negativos(":("). O texto é classificado de acordo com a classe que tiver mais emoticons. Apesar de possuir uma alta taxa de acertos este método depende muito da presença do emoticon no texto.	✓	
Opinion Lexicon [Hu and Liu, 2004]	Também conhecido como Sentiment Lexicon, consiste de uma lista com cerca de 6.800 palavras rotuladas como positivas e 6.800 palavras rotuladas como negativas, incluindo gírias e abreviações no idioma Inglês. Este é um método léxico criado a partir de textos coletados em reviews de produtos em sites de compra.	✓	
Opinion Finder (MPQA) [Wilson et al., 2005a] [Wilson et al., 2005b]	É uma ferramenta considerada híbrida pois utiliza um léxico de sentimentos mas utiliza Naive Bayes para distinguir se uma sentença é subjetiva ou objetiva.	✓	✓
Happiness Index [Dodds and Danforth, 2009]	É uma escala de sentimentos que utiliza o popular ANEW (um conjunto de palavras ligadas a emoções do Inglês). Este método foi contruido para avaliar textos entre 1 a 0, indicando a quantidade de felicidade existente. Em particular os autores utilizaram este método para mostrar que a "quantidade de felicidade" nas letras das músicas diminui entre 1961 e 2007.	✓	
SentiWordNet [Esuli and Sebastiani, 2006] [Baccianella et al., 2010]	É um léxico contruído a partir de outro léxico já conhecido chamado WordNet [Miller, 1995]. No WordNet os autores agruparam adjetivos, substantivos, verbos em conjuntos de palavras que fossem similares formando uma rede de palavras. Já os autores do SentiWordNet associaram uma polaridade entre algumas palavras-semestes do WordNet e propagaram essa polaridade nas palavras similares da WordNet criando um amplo léxico de sentimentos.	✓	✓
LIWC [Tausczik and Pennebaker, 2010]	O LIWC é uma ferramenta bem estabelecida e utilizada em diversas áreas, e contou com o aval de psicólogos, sociólogos e linguistas durante seu desenvolvimento. Ela possui um dicionário léxico de aproximadamente 4500 palavras e raízes de palavras, fazendo parte de oitenta categorias das mais variadas(ansiedade, saúde, lazer, etc).	✓	
SenticNet [Cambria et al., 2010]	SenticNet é um dicionário semântico e afetivo para opinião em nível de conceito e análise de sentimento. Ele foi construído através do que é denominado pelos autores de sentic computing, um pardigma que explora Inteligência Artificial e técnicas de Web semântica para processar opiniões via mineração de grafos e redução de dimensionalidade. Ele é público e provê um bom material para mineração de opniões em nível semântico e não apenas sintático.	✓	
AFINN [Nielsen, 2011b]	É um léxico contruído a partir do ANEW mas com o foco em redes sociais, contendo gírias e acrônimos e palavrões da lingua Inglesa. Ele possui uma lista de 2.477 termos classificados entre -5(mais negativo) e +5(mais positivo).	✓	
SO-CAL [Taboada et al., 2011]	É um método léxico que leva em conta a orientação semântica das palavras(SO). Criado contendo unigramas (verbos, advérbios, substanticos e adjetivos) e multi-gramas (intensificadores e frases) numa escala entre + 5 e -5. Os autores também incluíram analisador de partes do discurso e negação.	✓	
Emoticons DS (Distant Supervision)[Hannak et al., 2012]	É um léxico que possui termos gerados a partir de uma extensa base de dados do Twitter. Estes termos foram classificados automáticamente baseando-se na frequência de emoticons positivos ou negativos nas sentenças.	✓	
NRC Hashtag [Mohammad, 2012]	É um léxico que utilizam a técnica de supervisionamento distante para classificar seus termos. De forma geral ele classifica os termos provenientes do Twitter considerando as hashtags que o contém(i.e #joy, #sadness, etc).	✓	
Pattern.en [De Smedt and Daelemans, 2012]	É um pacote da linguagem python para lidar com processamento de linguagem natural. Um de seus módulos é responsável para inferir o sentimento no texto. Criado para ser rápido ele é baseado em polaridades associadas ao WordNet.	✓	
SASA [Wang et al., 2012]	Foi criado para detectar sentimentos no Twitter durante as eleições presidênciais de 2012 nos Estados Unidos. Ele foi contruído apartir de modelos estatístico do classificador Naïve Bayes em cima de unigramas classificados. Ele também explora emoções em emoticons e exclamações.		✓
PANAS-t [Gonçalves et al., 2013]	Tem como objetivo inicial detectar as flutuações de humor dos usuários no Twitter. O método é um léxico adaptado a partir de uma versão adaptada do PANAS Positive Affect Negative Affect Scale [Watson and Clark, 1985]. O PANAS é uma conhecida escala psicométrica que possue um grande conjunto de palavras associadas a 11 diferentes tipos de humor (surpresa, medo, etc).	✓	
EmoLex [Mohammad and Turney, 2013]	É um léxico criado a partir do Amazon Mechanical Turk, no qual pessoas foram pagas para classificar os termos. Cada entrada esta associada a 8 sentimentos básicos em inglês: joy, sadness, anger, etc definidos por [Plutchik, 1980]. A base do Emolex foi contruída utilizando termos do Macquarie Thesaurus e palavras do General Inquirer e do Wordnet.	✓	
SANN [Pappas and Popescu-Belis, 2013]	Foi construído para inferir a nota de avaliação de comentários dos usuários de produtos utilizando análise de sentimentos. Os comentários foram integrados em um classificador (kNN) ou K-Vizinhos mais próximos.	✓	✓
Sentiment140 Lexicon [Mohammad et al., 2013]	É um léxico criado de maneira similar ao NRC Hashtag [Mohammad, 2012]. Foi utilizado um classificador SVM que utilizava features como: número e categoria de emoticons	✓	
SentiStrength [Thelwall, 2013]	Builds a lexicon dictionary annotated by humans and improved with the use of Machine Learning.	✓	✓
Stanford Recursive Deep Model [Socher et al., 2013]	Tem como proposta uma variação do modelo de redes neurais chamadas Redes Neurais Recursivas que processa todas as sentenças procurando identificar sua estrutura e computar suas interações. É uma abordagem interes pois a técnica leva em consideração a ordem das palavras na sentença por exemplo, que é ignorada por vários métodos.	✓	✓
Umigon [Levallois, 2013]	Pertence a fámilia de léxicos e foi proposto para detectar sentimentos no Twitter, alé de subjetividade. O método utiliza diversos recursos linguísitivos como onomatopéias, exclamações, emoticons, etc. Ele possui heurísticas responsáveis para disambiguar o texto baseada em negações palavras alongadas e hashtags.	✓	
Vader [Hutto and Gilbert, 2014b]	Possui como base um dicionário léxico criado a partir de uma lista de palavras com base em dicionários já bem estebelecidos como LIWC, ANEW e GI. Em seguida, foram adicionadas construções léxicas presentes em microblogs tais como emoticons, acrônimos e gírias que expressam sentimentos.	✓	

Tabela 1.3. Métodos para análise de sentimentos em sentença - Saída e Comparações.

Nome	Saída	Validação	Comparado a
Emoticons	**-1, 1**	-	-
Opinion Lexicon	**-1, 0, 1**	Reviews de produtos da Amazon e CNET	-
Opinion Finder (MPQA)	**Negative, Neutral, Positive**	MPQA [Wiebe et al., 2005]	Comparado com versões diferentes do próprio Opinion Finder.
Happiness Index	**1, 2, 3, 4, 5, 6, 7, 8, 9**	Letras de músicas, Blogs, Mensagens oficiais do governo,	-
SentiWordNet	**-1, 0, 1**	-	General Inquirer (GI)[Stone et al., 1966]
LIWC	**negEmo, posEmo**	-	-
SenticNet	**Negative, Positive**	Opiniões de pacientes (Indisponível)	SentiStrength [Thelwall, 2013]
AFINN	**-1, 0, 1**	Twiter [Biever, 2010]	OpinonFinder [Wilson et al., 2005a], ANEW [Bradley and Lang, 1999], GI [Stone et al., 1966] e Sentistrength [Thelwall, 2013]
SO-CAL	**[<0), 0, (>0]**	Epinion [Taboada et al., 2006a], MPQA[Wiebe et al., 2005], Myspace[Thelwall, 2013],	MPQA[Wiebe et al., 2005], GI[Stone et al., 1966], SentiWordNet [Esuli and Sebastiani, 2006], Dicionário "Maryland"[Mohammad et al., 2009], Dicionário gerado pelo Google [Taboada et al., 2006b]
Emoticons DS (Distant Supervision)	**-1, 0, 1**	Validação com dataset não rotulado do twitter [Cha et al., 2010]	-
NRC Hashtag	**-1, 0, 1**	Twitter (SemEval-2007) [Strapparava and Mihalcea, 2007]	-
Pattern.en	**<0.1, ≥0.1]**	Reviews de produtos sem especificação da fonte	-
SASA [Wang et al., 2012]	**Negative, Neutral, Unsure, Positive**	Tweets "Políticos" rotulados por "turkers" (AMT) (indisponível)	-
PANAS-t	**-1, 0, 1**	Validação com dataset não rotulado do twitter [Cha et al., 2010]	-
EmoLex	**-1, 0, 1**	-	Comparado com dados padrão ouro porém não foram especificados
SANN	**neg, neu, pos**	Seu próprio dataset - Ted Talks	Comparação com outras abordagens de recomendação multimídia.
Sentiment140	**Negative, Neutral, Positive**	Twitter e SMS (Semeval 2013, tarefa 2) [Nakov et al., 2013].	Outras ferramentas apresentadas no Semeval 2013
SentiStrength	**-1,0,1**	Seus próprios datasets - Twitter, Youtube, Digg, Myspace, BBC Forums and Runners World.	Com as 9 melhoras técnicas de Aprendizado de Máquina para cada teste.
Stanford Recursive Deep Model	**very negative, negative, neutral, positive, very positive**	Movie Reviews [Pang and Lee, 2004]	Naïve Bayes e SVM com features unigrama e bigrama.
Umigon	**Negative, Neutral, Positive**	Twitter e SMS (Semeval 2013, tarefa 2) [Nakov et al., 2013].	[Mohammad et al., 2013]
Vader	**-1, 0, 1**	Seus próprios datasets - Twitter, Reviews de Filmes, Reviews Técnicos de Produtos, Opiniões de usuários do NYT.	(GI)[Stone et al., 1966], LIWC, [Tausczik and Pennebaker, 2010], SentiWordNet [Esuli and Sebastiani, 2006], ANEW [Bradley and Lang, 1999], SenticNet [Cambria et al., 2010] e outras abordagens de Aprendizado de Máquina.

1.4. Comparação entre Métodos

A grande aplicabilidade da análise de sentimentos em diversos segmentos tem levado uma série de empresas e pesquisadores de áreas distintas a investirem tempo e dinheiro em soluções que fazem interface com esta linha de pesquisa. É comum encontrar trabalhos que utilizam algum dos métodos para análise de sentimentos disponíveis atualmente como ferramenta para a produção de artefatos posteriormente aplicados a uma situação específica.

Os pesquisadores do Facebook, em um experimento polêmico [Kramer et al., 2014], utilizaram o LIWC [Tausczik and Pennebaker, 2010] para definir a polaridade das postagens em sua rede social e limitar a exibição de postagens com conteúdo apenas negativo na linha do tempo de um grupo de usuários e apenas positivo na linha do tempo de outro grupo. O objetivo, um tanto quanto controverso, era verificar a ocorrência de contágio emocional em redes sociais, ou seja, definir o quanto postagens negativas ou positivas influenciam no humor e nas futuras postagens de quem as lê. O Sentistrength [Thelwall, 2013] foi utilizado para o desenvolvimento de uma ferramenta chamada Magnet News, que permite aos leitores de jornais escolher se desejam ler notícias boas ou notícias ruins [Reis et al., 2014] enquanto o OpinionFinder foi utilizado, por exemplo, para definir a polaridade de postagens em blogs [Chenlo and Losada, 2011].

Assim como nos exemplos citados acima, muitos dos métodos existentes vem sendo empregados no desenvolvimento de aplicações sem um entedimento concreto da sua aplicabilidade em diferentes contextos, suas vantagens, limitações e eficiência comparado aos demais métodos. Alguns autores executaram alguns experimentos prévios para definir o melhor método a ser usado, como no caso do Magnet News, no entanto, a utilização caixa-preta sem preocupação explícita com os aspectos mencionados anteriormente é ocorre na maioria dos casos. Além disso, em uma análise minuciosa dos trabalhos em que são apresentados os métodos percebe-se que não existe um esforço no sentido de comparar o novo método proposto com métodos apresentados anteriormente.

Um esforço prévio no sentido de comparar métodos foi conduzido recentemente [Goncalves et al., 2013], no entanto, diversos métodos com novas abordagens e bons resultados tem sido apresentados recentemente e não encontra-se disponível na literatura um *benchmark* padronizado para compará-los. Com o objetivo de preencher esta lacuna, foi conduzida a construção de um *benchmark* de comparação entre os principais métodos disponíveis atualmente. Uma primeira etapa consistiu de um extensa busca na literatura relacionada por conjuntos de dados (*datasets* rotulados também chamados de dados padrão ouro (*Golden Standard Data*). Tais dados são compostos por sentenças cuja polaridade foi definida previamente de maneira precisa, em geral, realizada por humanos. Para medir-se a qualidade de um método de maneira abrangente é preciso de uma quantidade razoável de sentenças previamente rotuladas. A tabela 1.4 apresenta detalhes de vinte datasets rotulados com sentenças provenientes de diversos contextos como comentários em sites de notícias (Comments_BBC e Commens_NYT) e vídeos (Comments_TED e Comments_YTB), reviews de produtos e filmes (Amazon, Reviews_I e Reviews_II), postagens em redes sociais e micoblogs (Myspace, Tweets_RND_I, etc) além de dois pequenos datasets construídos pelos próprios autores contendo sentenças rotuladas para tweets com hashtags #sarcasm (sarcasmo) e #irony (ironia) obtidos de uma amostra ale-

Tabela 1.4. Datasets Rotulados.

Dataset	Nomeclatura	# Msgs	# Pos	# Neg	# Neu	# Médio de frases	# Médio de palavras	# de Aval.
Comments (BBC) [Thelwall, 2013]	**Comments_BBC**	1.000	99	653	248	3,98	64,39	3
Comments (Digg) [Thelwall, 2013]	**Comments_Digg**	1.077	210	572	295	2,50	33,97	3
Comments (NYT) [Hutto and Gilbert, 2014b]	**Comments_NYT**	5.190	2.204	2.742	244	1,01	17,76	20
Comments (TED) [Pappas and Popescu-Belis, 2013]	**Comments_TED**	839	318	409	112	1	16,95	6
Comments (Youtube) [Thelwall, 2013]	**Comments_YTB**	3.407	1.665	767	975	1,78	17,68	3
Reviews-Filmes [Pang and Lee, 2004]	**Reviews_I**	10.662	5.331	5.331	-	1,15	18,99	-
Reviews-Filmes [Hutto and Gilbert, 2014b]	**Reviews_II**	10.605	5.242	5.326	37	1,12	19,33	20
Posts Myspace [Thelwall, 2013]	**Myspace**	1.041	702	132	207	2,22	21,12	3
Reviews-Produtos [Hutto and Gilbert, 2014b]	**Amazon**	3.708	2.128	1.482	98	1,03	16,59	20
Tweets (Debate) [Diakopoulos and Shamma, 2010]	**Tweets_DBT**	3.238	730	1.249	1.259	1,86	14,86	Indef.
Tweets (Irony) (Rotulado pelos autores)	**Irony**	100	38	43	19	1,01	17,44	3
Tweets (Sarcasm) (Rotulado pelos autores)	**Sarcasm**	100	38	38	24	1	15,55	3
Tweets (Random) [Thelwall, 2013]	**Tweets_RND_I**	4.242	1.340	949	1.953	1,77	15,81	3
Tweets (Random) [Hutto and Gilbert, 2014b]	**Tweets_RND_II**	4.200	2.897	1.299	4	1,87	14,10	20
Tweets (Random) [Narr et al., 2012]	**Tweets_RND_III**	3.771	739	488	2.536	1,54	14,32	3
Tweets (Random) [Aisopos, 2014]	**Tweets_RND_IV**	500	139	119	222	1,90	15,44	Indef
Tweets (Specific domains w/ emot.) [Go et al., 2009b]	**Tweets_STF**	359	182	177	-	1,0	15,1	Indef.
Tweets (Specific topics) [Sanders, 2011]	**Tweets_SAN**	3737	580	654	2503	1,60	15,03	1
Tweets (Semeval2013 Task2) [Nakov et al., 2013]	**Tweets_Semeval**	6.087	2.223	837	3.027	1,86	20,05	5
Runners World forum [Thelwall, 2013]	**RW**	1.046	484	221	341	4,79	66,12	3

atória. A tabela detalha o número de sentenças positivas, negativas e neutras para cada dataset além da média de palavras e frases presentes na sentenças. Além disso, a tabela destaca uma nomenclatura que é utilizada no texto e o número de avaliadores responsáveis por determinar a polaridade de cada sentença. Note que em alguns casos o número de avaliadores não é disponível (Indef.) e no caso do dataset Reviews_I, a polaridade da sentença foi definida com base na nota dada ao filme pelo usuário.

1.4.1. Detalhes dos experimentos

Pelo menos três diferentes abordagens são encontradas para descobrir a polaridade das sentenças. A primeira delas divide a tarefa em dois passos: 1 - identificar sentenças que não expressam sentimento também chamadas de sentenças objetivas ou neutras e 2 - detectar a polaridade (positiva ou negativa) para as sentenças restantes, as sentenças subjetivas. Outra maneira comum de se detectar a polaridade é alcançada com classificação direta em uma das três classes. Por fim, alguns métodos classificam sentenças apenas como positivas ou negativas, assumindo que apenas sentenças polarizadas serão tratadas. As duas primeiras abordagens são chamadas abordagens 3-classes uma vez que detectam as sentenças neutras, positivas e negativas enquanto a terceira abordagem, chamada 2-classes detectam apenas as duas últimas polaridades.

Confrontar os resultados de métodos com diferentes abordagens é uma tarefa complicada. O principal desafio diz respeito à comparação de métodos 2-classes com métodos 3-classes. Esta comparação é importante pois métodos 3-classes podem ser utilizados em

um contexto com apenas duas classes e vice-versa, especialmente em um cenário de utilização caixa preta que tem sido comum.

Com o intuito de realizar uma comparação completa e justa entre as diferentes abordagens foram propostas duas rodadas de experimentos. Os primeiros experimentos focaram na abordagem 3-classes e foram executados apenas com os datasets que continham sentenças neutras porém também incluíram os métodos 2-classes. A decisão por incluir os métodos 2-classes se deu pelo fato de haver a expectativa que alguns métodos 2-classes poderiam eventualmente obter melhores resultados do que métodos 3-classes. Além disso, em alguns casos, os métodos 2-classes não são capazes de definir se uma sentença é positiva ou negativa, caso em que as sentenças foram consideradas neutras para os experimentos.

Os experimentos 2-classes por sua vez foram executados com todos os datasets padrão ouro excluindo-se as sentenças neutras. Esta segunda etapa de experimentos incluiu também os métodos 3-classes. Da mesma forma que no experimento anterior, desejava-se verificar se métodos 3-classes apresentariam resultados superiores a métodos 2-classes em contextos de dupla polaridade apenas. No caso de um método 3-classes detectar sentenças neutras asumiu-se que o método não foi capaz de decodificar a sentença e ela se encaixa no conjunto de sentenças com polaridade indefinida. Dessa forma introduziu-se o conceito de cobertura, que indica o percentual de sentenças em que se pôde detectar a polaridade com a seguinte fórmula: $Cobertura = \frac{\#Sent. - \#Indef.}{\#Sent.}$. Note que mesmo métodos 2-classes possuem o valor de cobertura já que não são capazes de detectar a polaridade de algumas sentenças e esta suposição não representa uma falha metodológica.

1.4.2. Métricas

Um aspecto chave na avaliação dos métodos para a análise de sentimentos diz respeito às métricas utilizadas. Neste contexto, três métricas principais são comumente empregadas para validar a eficiência de um método: acurácia, precisão e revocação.

A acurácia indica o percentual de sentenças corretamente classificadas, isto é, a soma acertos de todas as classes dividido pelo número total sentenças classificadas. Note que a acurácia, por si só, pode não ser uma métrica eficaz uma vez que, conhecendo-se previamente a prevalência de classe em determinado contexto, basta atribuir a cada sentença decodificada a classe de maior ocorrência e obter-se-á boa acurácia.

Já a precisão é calculada para cada classe individualmente e evidencia o percentual de sentenças corretamente classificadas para aquela classe. Ou seja, basta dividir os acertos da classe pela quantidade de elementos classificados como pertencendo àquela classe. Um alto valor de precisão também pode ser ilusório no caso em que muitos elementos da classe não são classificados como pertencendo à classe. Imagine uma situação em que existam 100 sentenças positivas e 100 sentenças negativas no conjunto de dados, e o método em questão tenha atribuído como sendo positivas apenas 10 sentenças das quais 9 eram corretas. A precisão será 90%, contudo do total de 100 sentenças positivas apenas 9% foram corretamente definidas. A revocação tem como funcionalidade dar indícios da situação explicitada anteriormente e é calculada justamente pelo total de sentenças corretamente classificadas para uma classe sobre o total de sentenças desta classe na base de dados.

De fato, precisão e revocação em conjunto dão boas indicações da eficiência de um método em predizer a polaridade de sentenças, sendo assim, utiliza-se também a F1-Score que nada mais é a média harmônica das duas métricas anteriores. Formalmente, as métricas para o experimento 3-classes são calculadas como pode ser apresentado na tabela a seguir. O cálculo das métricas para o experimento 2-classes segue o mesmo princípio, eliminando-se a classe neutra.

		Predição		
		Positiva	Neutra	Negativa
Correto	Positiva	a	b	c
	Neutra	d	e	f
	Negativa	g	h	i

Cada letra na tabela acima representa o número de instâncias de texto cuja classe correta é X e cuja predição é a classe Y, onde X:Y $\in$ positive; neutral; negative. A revocação (R) da classe X é a taxa de número de elementos corretamente classificados pelo total de elementos na classe X. Já a precisão (P) de uma classe X é taxa de número de elementos classificados corretamente pelo total de elementos classificadosc omo sendo da X. Por exemplo, a precisão da classe negativa é computada como: $P(neg) = i/(c+f+i)$; enquanto a revocação é: $R(neg) = i/(g+h+i)$; e o $F1$ é a média harmônica entre ambos precisão e revocação. Neste caso $F1(neg) = \frac{2P(neg)\cdot R(neg)}{P(neg)+R(neg)}$.

A acurácia global é calculada pela seguinte fórmula: $A = \frac{a+e+i}{a+b+c+d+e+f+g+h+i}$. Ela considera igualmente importante a correta classificação de cada sentença, independente da classe, ou seja, ela mede basicamente a capacidade de um método predizer uma entrada corretamente. Por fim, utilizou-se também a Macro-F1, utilizada para medir a efetividade global de classificação já que a F1 aplica-se a cada classe individualmente. A Macro-F1 é calculada com base na média das medidas F1 de cada classe separadamente, independente do tamanho relativo de cada classe. Desta forma, a acurácia global e a Macro-F1 fornecem parâmetros complementares para a verificação da efetividade de classificação de um método. A Macro-F1 é especialmente importante quando a distribuição entre classes é enviesada permitindo verificar a capacidade do método de obter bons resultados em classes com pequenas quantidades de sentença.

Como forma de permitir uma comparação global entre os métodos foi utilizado um critério de comparação simples mas que permite ter uma ideia interessante da performance. A métrica é basicamente o rank médio em que um método ficou em cada dataset. Por exemplo, se um método ficou em primeiro lugar, ou seja no rank 1, em todos os datasets, seu rank médio será, obviamente 1. Para realizar este cálculo bastou somar o rank do dataset em cada dataset e divir pela quantidade de datasets utilizados no experimento.

1.4.3. Resultados

Por questões de espaço apenas alguns dos resultados serão exibidos, porém os resultados completos estão disponíveis na Web [5].

[5] `http://homepages.dcc.ufmg.br/~fabricio/benchmark_sentiment_analysis.html`

Tabela 1.5. Resultados do experimento com 3-classes para 4 datasets

Dataset	Method	Acur.	Sentimento Posit.			Sentimento Negat.			Sentimento Neut.			MacroF1
			P	R	F1	P	R	F1	P	R	F1	
Tweets _SemEval	AFINN	62,36	61,10	70,09	65,28	44,08	31,91	37,02	71,43	58,57	64,37	55,56
	Emolex	48,74	48,15	62,71	54,47	31,27	17,71	22,61	57,90	41,30	48,21	41,76
	Emoticons DS	36,59	36,55	100,00	53,53	75,00	0,08	0,16	100,00	0,03	0,07	17,92
	Happiness Index	48,81	43,61	65,27	52,29	36,96	7,54	12,53	36,82	45,16	40,56	35,13
	NRC Hashtag	36,95	42,04	75,03	53,88	24,57	16,94	20,05	53,33	3,70	6,92	26,95
	Opinion Finder	57,63	67,57	27,94	39,53	40,75	48,62	44,34	58,20	86,06	69,44	51,10
	Opinion Lexicon	60,37	62,09	62,71	62,40	41,19	34,18	37,36	66,41	60,75	63,46	54,41
	PANAS-t	53,08	90,95	9,04	16,45	51,56	62,26	56,41	51,65	99,01	67,89	46,92
	Pattern.en	50,19	58,07	68,47	62,84	24,68	29,82	27,01	67,73	35,22	46,34	45,40
	SANN	54,77	52,72	47,59	50,02	38,91	20,92	27,21	58,95	66,90	62,67	46,64
	SASA	50,63	46,34	47,77	47,04	33,07	12,14	17,76	56,39	61,12	58,66	41,15
	SO-CAL	58,83	58,89	59,02	58,95	40,39	33,14	36,41	39,89	59,96	47,91	47,76
	SentWordNet	48,33	55,54	53,44	54,47	19,67	24,82	21,95	61,22	47,57	53,54	43,32
	SentiStrength	57,83	78,01	27,13	40,25	47,80	53,55	50,52	55,49	89,89	68,62	53,13
	SenticNet	39,90	39,81	86,55	54,54	31,85	8,98	14,01	38,18	7,20	12,12	26,89
	Sentiment140	42,25	0,00	0,00	0,00	26,79	100,00	42,25	50,57	66,14	57,31	33,19
	Stanford DM	0,23	0,72	0,18	0,29	0,15	0,91	0,26	0,47	0,07	0,12	0,22
	Umigon	65,88	75,18	56,14	64,28	39,66	53,18	45,44	70,65	75,78	73,13	60,95
	Vader	60,05	56,08	79,26	65,68	44,13	26,60	33,19	76,88	46,02	57,57	52,15
Tweets _RND_III	AFINN	64,41	40,81	72,12	52,13	49,67	28,29	36,05	85,95	62,54	72,40	53,53
	Emolex	54,76	31,67	59,95	41,44	40,14	19,53	26,27	77,48	54,64	64,08	43,93
	Emoticons DS	20,34	19,78	99,46	33,00	62,07	0,60	1,19	53,85	0,55	1,09	11,76
	Happiness Index	55,16	29,13	61,98	39,64	50,65	9,50	16,01	43,35	59,16	50,03	35,23
	NRC Hashtag	30,47	28,25	77,40	41,39	24,18	19,59	21,64	79,08	8,77	15,78	26,27
	Opinion Finder	71,55	57,48	32,75	41,72	49,85	48,56	49,20	75,95	89,90	82,34	57,75
	Opinion Lexicon	63,86	40,65	66,17	50,36	48,84	27,73	35,38	81,96	64,66	72,29	52,68
	PANAS-t	68,79	79,49	8,39	15,18	48,57	51,52	50,00	68,75	98,86	81,10	48,76
	Pattern.en	53,57	36,25	76,86	49,26	35,19	22,50	27,45	84,20	45,68	59,23	45,31
	SANN	66,88	42,70	48,71	45,51	46,35	26,93	34,07	77,99	77,99	77,99	52,52
	SASA	55,37	29,42	54,53	38,22	42,46	19,28	26,52	78,30	57,15	66,08	43,60
	SO-CAL	66,51	43,06	68,88	52,99	51,84	30,55	38,44	45,77	66,94	54,37	48,60
	SentWordNet	55,85	37,42	58,19	45,55	24,04	19,57	21,58	79,25	59,00	67,64	44,92
	SentiStrength	73,80	70,94	41,95	52,72	57,53	49,80	53,39	75,35	92,26	82,95	63,02
	SenticNet	33,47	23,66	86,60	37,17	41,47	10,06	16,19	43,44	15,37	22,71	25,36
	Sentiment140	55,05	0,00	0,00	0,00	28,14	100,00	43,92	71,14	66,00	68,47	37,46
	Stanford DM	31,90	64,48	38,57	48,26	15,58	72,55	25,65	75,64	19,77	31,35	35,09
	Umigon	74,12	57,67	70,23	63,33	48,83	46,71	47,75	88,80	76,34	82,10	64,39
	Vader	59,82	37,52	81,73	51,43	47,99	24,25	32,22	89,26	52,28	65,94	49,86
Comments _BBC	AFINN	50,10	16,22	60,61	25,59	82,62	54,14	65,42	40,11	30,24	34,48	41,83
	Emolex	44,10	15,51	65,66	25,10	83,19	45,62	58,93	35,27	31,85	33,47	39,17
	Emoticons DS	10,00	9,85	98,99	17,92	66,67	0,22	0,44	0,00	0,00	0,00	9,18
	Happiness Index	33,60	11,83	64,65	20,00	84,93	28,05	42,18	26,46	34,68	30,02	30,73
	NRC Hashtag	64,00	20,72	23,23	21,90	70,20	87,13	77,76	52,50	8,47	14,58	38,08
	Opinion Finder	51,80	14,96	35,35	21,02	78,76	66,39	72,04	33,71	36,29	34,95	42,67
	Opinion Lexicon	55,00	20,67	62,63	31,08	85,27	61,98	71,79	40,82	40,32	40,57	47,81
	PANAS-t	27,10	16,67	6,06	8,89	75,61	50,82	60,78	25,35	94,35	39,97	36,55
	Pattern.en	46,00	14,39	58,59	23,11	77,30	49,93	60,67	38,16	23,39	29,00	37,59
	SANN	40,10	14,50	59,60	23,32	79,49	41,61	54,63	33,45	37,90	35,54	37,83
	SASA	38,20	17,03	47,47	25,07	70,75	50,86	59,18	25,19	39,52	30,77	38,34
	SO-CAL	55,50	20,88	57,58	30,65	80,47	65,61	72,28	28,57	34,68	31,33	44,75
	SentWordNet	42,40	14,90	59,60	23,84	81,63	44,57	57,66	34,56	37,90	36,15	39,22
	SentiStrength	44,20	47,37	18,18	26,28	86,64	91,45	88,98	29,37	84,68	43,61	52,96
	SenticNet	27,90	11,91	88,89	21,00	82,69	20,90	33,37	26,39	7,66	11,88	22,08
	Sentiment140	50,60	0,00	0,00	0,00	73,23	100,00	84,54	28,60	58,47	38,41	40,98
	Stanford DM	65,50	43,37	36,36	39,56	71,01	92,54	80,36	37,50	14,52	20,93	46,95
	Umigon	45,70	28,35	36,36	31,86	76,35	74,65	75,49	29,31	61,69	39,74	49,03
	Vader	49,10	15,96	71,72	26,10	82,57	49,05	61,54	50,42	24,19	32,70	40,11
Comments _NYT	AFINN	42,45	64,81	41,79	50,81	80,29	68,59	73,98	7,89	77,87	14,32	46,37
	Emolex	42,97	55,12	53,72	54,41	75,35	48,67	59,14	7,22	54,10	12,74	42,10
	Emoticons DS	42,58	42,55	99,77	59,66	78,57	0,37	0,73	0,00	0,00	0,00	30,20
	Happiness Index	31,81	48,42	50,18	49,29	71,70	25,96	38,12	5,36	54,10	9,76	32,39
	NRC Hashtag	54,84	55,38	45,74	50,10	61,55	68,92	65,03	8,33	15,16	10,76	41,96
	Opinion Finder	29,38	68,77	18,78	29,51	76,52	82,66	79,47	6,29	88,11	11,75	40,24
	Opinion Lexicon	44,57	65,95	43,15	52,17	79,81	70,65	74,95	7,94	73,77	14,34	47,15
	PANAS-t	5,88	69,23	1,23	2,41	62,07	75,00	67,92	4,75	99,18	9,07	26,47
	Pattern.en	45,39	55,15	44,69	49,37	63,65	61,12	62,36	7,85	45,90	13,41	41,71
	SANN	27,92	56,74	29,40	38,73	78,02	55,13	64,61	5,93	79,51	11,04	38,13
	SASA	30,04	49,92	30,13	37,58	59,11	52,83	55,80	5,74	61,07	10,49	34,62
	SO-CAL	50,79	64,36	51,13	56,99	77,25	68,36	72,53	8,68	65,98	15,34	48,29
	SentWordNet	32,20	57,35	34,53	43,10	70,31	56,63	62,73	6,08	70,08	11,19	39,01
	SentiStrength	33,61	80,45	33,65	47,45	75,25	74,51	74,88	16,36	88,39	27,62	49,98
	SenticNet	46,48	42,99	87,74	57,70	70,13	22,59	34,18	7,69	2,68	3,97	31,95
	Sentiment140	34,66	0,00	0,00	0,00	65,76	100,00	79,34	5,83	64,34	10,69	30,01
	Stanford DM	51,93	73,39	21,14	32,83	59,48	92,67	72,46	9,65	38,11	15,40	40,23
	Umigon	24,08	68,76	16,38	26,46	68,78	80,38	74,13	5,88	88,93	11,04	37,21
	Vader	48,84	61,96	52,40	56,78	80,09	63,00	70,52	9,51	70,90	16,77	48,03

A tabela 1.5 apresenta os resultados do experimento 3-classes para 4 datasets distintos: Tweets_Semeval, Tweets_RND_III, Comments_BBC e Comments_NYT. Pode-se perceber que não existe um método que sempre obtem a melhor predição em diferentes datasets. O Umigon, apesar de ter ganhado em todos os datasets de twitter no experimento 3-classes, ficou em quarto lugar geral para os mesmos experimentos conforme table 1.6 e no dataset Comments_NYT acabou ficando na décima terceira posição, com macro F1 bem inferior à dos primeiros colocados. Isto sugere que uma investigação preliminar deve ser conduzida quando se utilizar um novo dataset. Outra constatação interessante é a permanência dos mesmos cinco métodos melhores classificados em ambos experimentos: Sentistrength, AFINN, OpinionLexion, Umigon e Vader.

Tabela 1.6. Tabela de rank médio

2 Classes		3 Classes	
Método	**Rank médio**	**Método**	**Rank médio**
SentiStrength	3,4	SentiStrength	1,5
AFINN	3,5	Opinion Lexicon	4,2
Opinion Lexicon	3,7	SO-CAL	4,4
Umigon	4,2	AFINN	5,1
Vader	5,4	Vader	6,0
SO-CAL	5,5	Umigon	6,3
Opinion Finder	8,1	PANAS-t	8,15
Pattern.en	8,4	Pattern.en	8,2
SANN	9,6	Opinion Finder	9,65
SentWordNet	10	Emolex	10,1
Emolex	10,2	SANN	10,5
Stanford DM	11,8	Stanford DM	11,6
SASA	13,1	NRC Hashtag	11,8
NRC Hashtag	13,75	SentWordNet	12,4
PANAS-t	14,2	SASA	13
Sentiment140	15,1	Happinness Index	15,2
Happinness Index	15.2	SenticNet	15,2
SenticNet	16,2	Sentiment140	18
Emoticons DS	18,7	Emoticons DS	18,5

1.4.4. Abordagem Multilíngue

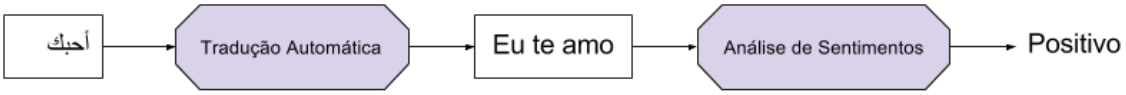

Figura 1.3. Simples técnica para realizar a análise em outros idiomas

Há um enorme número de pesquisas relacionadas a criação de métodos e vários acabam até se tornando populares. No entanto, poucos esforços vêm sendo feitos no desenvolvimento de métodos para detecção de sentimentos em mensagens em idiomas diferentes do inglês. Neste caso há tentativas de recriar uma técnica supervisionada, a partir de novos dados rotulados ou mesmo traduzindo dicionários léxicos de métodos já existentes. Mas estas abordagens não são tão efetivas pois é caro obter dados rotulados e desenvolver um novo método, assim como muitas palavras e gírias especificas de uma língua não estão contidas no léxicos ou serão traduzidos erroneamente.

Uma forma simples e eficiente para realizar a análise de sentimentos em diferentes idiomas é combinar o poder dos métodos já existentes em inglês com a eficiência dos tradutores automáticos como apresentado na Figura 1.3. Uma simples tradução do texto de entrada em ferramentas como o Google Tradutor no idioma que o método análise pode trazer bons resultados como os apresentados por [Reis et al., 2015a].

1.5. iFeel - Uma ferramenta online para análise de sentimentos

Como uma extensão dos esforços de nossa equipe em buscar e agrupar estes diversos métodos na literatura. Foi disponibilizado em `www.ifeel.dcc.ufmg.br` o serviço iFeel [Araújo et al., 2014]. Este serviço facilita o acesso aos diversos métodos de análise de sentimentos discutidos neste trabalho. Sendo assim uma ferramenta bem útil para aqueles que gostariam de ter contato com a análise de sentimento mesmo sem conhecimento algum da área. O iFeel também pode ajudar pesquisadores que pretendem avaliar novos métodos em uma única plataforma.

Referências

[nlt, 2015] (2015). Natural language toolkit. `http://www.nltk.org/`. Accessed September 23, 2015.

[std, 2015] (2015). The stanford natural language processing group. `http://nlp.stanford.edu/software/tagger.shtml`. Accessed September 23, 2015.

[Aisopos, 2014] Aisopos, F. (2014). Manually annotated sentiment analysis twitter dataset ntua. `www.grid.ece.ntua.gr`.

[Araújo et al., 2014] Araújo, M., Gonçalves, P., Cha, M., and Benevenuto, F. (2014). ifeel: A system that compares and combines sentiment analysis methods. In *Proceedings of the companion publication of the 23rd international conference on World wide web companion*, pages 75–78. International World Wide Web Conferences Steering Committee.

[Baccianella et al., 2010] Baccianella, S., Esuli, A., and Sebastiani, F. (2010). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *LREC*. European Language Resources Association.

[Biever, 2010] Biever, C. (2010). Twitter mood maps reveal emotional states of america. *The New Scientist*, 207.

[Bollen et al., 2010] Bollen, J., Mao, H., and Zeng, X. (2010). Twitter mood predicts the stock market. *CoRR*, abs/1010.3003.

[Bradley and Lang, 1999] Bradley, M. M. and Lang, P. J. (1999). Affective norms for English words (ANEW): Stimuli, instruction manual, and affective ratings. Technical report, Center for Research in Psychophysiology, University of Florida.

[Cambria et al., 2010] Cambria, E., Speer, R., Havasi, C., and Hussain, A. (2010). Senticnet: A publicly available semantic resource for opinion mining. In *AAAI Fall Symposium Series*.

[Cha et al., 2010] Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, K. P. (2010). Measuring user influence in twitter: The million follower fallacy. In *International AAAI Conference on Weblogs and Social Media (ICWSM)*.

[Chenlo and Losada, 2011] Chenlo, J. M. and Losada, D. E. (2011). Effective and efficient polarity estimation in blogs based on sentence-level evidence. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 365–374, New York, NY, USA. ACM.

[Cortes and Vapnik, 1995] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.*, 20(3):273–297.

[De Choudhury et al., 2014] De Choudhury, M., Counts, S., Horvitz, E. J., and Hoff, A. (2014). Characterizing and predicting postpartum depression from shared facebook

data. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, pages 626–638, New York, NY, USA. ACM.

[De Smedt and Daelemans, 2012] De Smedt, T. and Daelemans, W. (2012). Pattern for python. *The Journal of Machine Learning Research*, 13(1):2063–2067.

[Diakopoulos and Shamma, 2010] Diakopoulos, N. and Shamma, D. (2010). Characterizing debate performance via aggregated twitter sentiment. In *Proc. CHI*.

[Dodds and Danforth, 2009] Dodds, P. S. and Danforth, C. M. (2009). Measuring the happiness of large-scale written expression: songs, blogs, and presidents. *J. of Happiness Studies*, 11.

[Esuli and Sebastiani, 2006] Esuli and Sebastiani (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proc. LREC*.

[Feldman, 2013] Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4):82–89.

[Go et al., 2009a] Go, A., Bhayani, R., and Huang, L. (2009a). Twitter sentiment classification using distant supervision. *Processing*.

[Go et al., 2009b] Go, A., Bhayani, R., and Huang, L. (2009b). Twitter sentiment classification using distant supervision. *Processing*.

[Goncalves et al., 2013] Goncalves, P., Araujo, M., Benevenuto, F., and Cha, M. (2013). Comparing and combining sentiment analysis methods. In *Proc. COSN*.

[Gonçalves et al., 2013] Gonçalves, P., Benevenuto, F., and Cha, M. (2013). PANAS-t: A Pychometric Scale for Measuring Sentiments on Twitter. abs/1308.1857v1.

[Hannak et al., 2012] Hannak, A., Anderson, E., Barrett, L. F., Lehmann, S., Mislove, A., and Riedewald, M. (2012). Tweetin' in the rain: Exploring societal-scale effects of weather on mood. In *ICWSM*.

[Hu and Liu, 2004] Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. Proc. KDD'04, pages 168–177.

[Hutto and Gilbert, 2014a] Hutto, C. and Gilbert, E. (2014a). Vader: A parsimonious rule-based model for sentiment analysis of social media text.

[Hutto and Gilbert, 2014b] Hutto, C. J. and Gilbert, E. (2014b). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *ICWSM*.

[Jones, 2004] Jones, K. S. (2004). Idf term weighting and ir research lessons. *Journal of Documentation*, 60(5):521–523.

[Kramer et al., 2014] Kramer, A. D. I., Guillory, J. E., and Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 111(24):8788–90.

[Levallois, 2013] Levallois, C. (2013). Umigon: sentiment analysis for tweets based on terms lists and heuristics. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 414–417, Atlanta, Georgia, USA. Association for Computational Linguistics.

[Liu, 2010] Liu, B. (2010). *Sentiment analysis and subjectivity*.

[Lorena and De Carvalho, 2008] Lorena, A. C. and De Carvalho, A. C. P. L. F. (2008). Evolutionary tuning of SVM parameter values in multiclass problems. *Neurocomputing*, 71(16-18):3326–3334.

[Miller, 1995] Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38.

[Mohammad et al., 2009] Mohammad, S., Dunne, C., and Dorr, B. (2009). Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, EMNLP '09, pages 599–608, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Mohammad and Turney, 2013] Mohammad, S. and Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29.

[Mohammad, 2012] Mohammad, S. M. (2012). # emotional tweets. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 246–255. Association for Computational Linguistics.

[Mohammad et al., 2013] Mohammad, S. M., Kiritchenko, S., and Zhu, X. (2013). Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proc. SemEval-2013*.

[Nakov et al., 2013] Nakov, P., Kozareva, Z., Ritter, A., Rosenthal, S., Stoyanov, V., and Wilson, T. (2013). Semeval-2013 task 2: Sentiment analysis in twitter.

[Narr et al., 2012] Narr, S., Hülfenhaus, M., and Albayrak, S. (2012). Language-independent twitter sentiment analysis. *Knowledge Discovery and Machine Learning (KDML)*, pages 12–14.

[Nielsen, 2011a] Nielsen, F. Å. (2011a). A new anew: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*.

[Nielsen, 2011b] Nielsen, F. Å. (2011b). A new anew: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*.

[Pang and Lee, 2004] Pang, B. and Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proc. Annual meeting of ACL Conference*.

[Pang et al., 2002] Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of EMNLP*, pages 79–86.

[Pappas and Popescu-Belis, 2013] Pappas, N. and Popescu-Belis, A. (2013). Sentiment analysis of user comments for one-class collaborative filtering over ted talks. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 773–776. ACM.

[Plutchik, 1980] Plutchik, R. (1980). *A general psychoevolutionary theory of emotion*, pages 3–33. Academic press, New York.

[Reis et al., 2015a] Reis, J., Araújo, M., Pereira, A. C., Benevenuto, F., and Gonçalves, P. (2015a). Uma abordagem multilíngue para análise de sentimentos. In *CSBC 2015 - BraSNAM ()*.

[Reis et al., 2015b] Reis, J., Benevenuto, F., Vaz de Melo, P., Prates, R., Kwak, H., and An, J. (2015b). Breaking the news: First impressions matter on online news. In *Proceedings of the 9th International AAAI Conference on Web-Blogs and Social Media*, Oxford, UK.

[Reis et al., 2014] Reis, J., Goncalves, P., Vaz de Melo, P., Prates, R., and Benevenuto, F. (2014). Magnet news: You choose the polarity of what you read. In *International AAAI Conference on Web-Blogs and Social Media*.

[Sanders, 2011] Sanders, N. (2011). Twitter sentiment corpus by niek sanders. `http://www.sananalytics.com/lab/twitter-sentiment/`.

[Severyn and Moschitti, 2015] Severyn, A. and Moschitti, A. (2015). Twitter sentiment analysis with deep convolutional neural networks. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 959–962, New York, NY, USA. ACM.

[Socher et al., 2013] Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Conf. on Empirical Methods in NLP*.

[Stone et al., 1966] Stone, P. J., Dunphy, D. C., Smith, M. S., and Ogilvie, D. M. (1966). *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press.

[Strapparava and Mihalcea, 2007] Strapparava, C. and Mihalcea, R. (2007). Semeval-2007 task 14: Affective text. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 70–74, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Taboada et al., 2006a] Taboada, M., Anthony, C., and Voll, K. (2006a). Methods for creating semantic orientation dictionaries. In *Conference on Language Resources and Evaluation (LREC)*, pages 427–432.

[Taboada et al., 2006b] Taboada, M., Anthony, C., and Voll, K. (2006b). Methods for creating semantic orientation dictionaries. In *Conference on Language Resources and Evaluation (LREC)*, pages 427–432.

[Taboada et al., 2011] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Comput. Linguist.*, 37(2):267–307.

[Tausczik and Pennebaker, 2010] Tausczik, Y. R. and Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *J. of Lang. and Soc. Psych.*, 29.

[Thelwall, 2013] Thelwall, M. (2013). Heart and soul: Sentiment strength detection in the social web with sentistrength. `http://sentistrength.wlv.ac.uk/documentation/SentiStrengthChapter.pdf`.

[Wang et al., 2012] Wang, H., Can, D., Kazemzadeh, A., Bar, F., and Narayanan, S. (2012). A system for real-time twitter sentiment analysis of 2012 u.s. presidential election cycle. In *ACL System Demonstrations*.

[Watson and Clark, 1985] Watson, D. and Clark, L. (1985). Development and validation of brief measures of positive and negative affect: the panas scales. *J. of Pers. and So. Psych.*, 54.

[Wiebe et al., 2005] Wiebe, J., Wilson, T., and Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 1(2):0.

[Wilson et al., 2005a] Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., Riloff, E., and Patwardhan, S. (2005a). Opinionfinder: a system for subjectivity analysis. In *HLT/EMNLP on Interactive Demonstrations*.

[Wilson et al., 2005b] Wilson, T., Wiebe, J., and Hoffmann, P. (2005b). Recognizing contextual polarity in phrase-level sentiment analysis. In *ACL Conference on Empirical Methods in Natural Language Processing*.