UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA

ALINE GRACIELA LERMEN DOS SANTOS

Experimentos Comparativos Combinando Aprendizado Supervisionado e Tradução Automática para Mineração de Emoções em Textos Multilíngues

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação

Orientadora: Prof[a]. Dr[a]. Karin Becker
Co-orientadora: Prof[a]. Dr[a]. Viviane Moreira

Porto Alegre
2016

CIP – CATALOGAÇÃO NA PUBLICAÇÃO

Lermen dos Santos, Aline Graciela

Experimentos Comparativos Combinando Aprendizado Supervisionado e Tradução Automática para Mineração de Emoções em Textos Multilíngues / Aline Graciela Lermen dos Santos. – Porto Alegre: da UFRGS, 2016.

78 f.: il.

Dissertação (mestrado) – Universidade Federal do Rio Grande do Sul. , Porto Alegre, BR–RS, 2016. Orientadora: Karin Becker; Co-orientadora: Viviane Moreira.

1. Análise de sentimentos. 2. Análise de sentimentos multilíngues. 3. Emoção. 4. Mineração de emoções. I. Becker, Karin. II. Moreira, Viviane. III. Título.

AGRADECIMENTOS

Eu não poderia começar os agradecimentos de outra forma senão agradecendo aos meus pais. Obrigada pelo apoio, confiança e incentivo que me deram durante toda a carreira acadêmica. Em especial à minha mãe, eu não teria chegado até aqui sem ela. Agradeço também aos outros membros da minha família pelo incentivo que sempre me deram para continuar estudando.

Agradeço especialmente ao meu marido, pelo apoio, carinho, companheirismo e ajuda que sempre me prestou durante nossa jornada pela vida pessoal e acadêmica. Certamente não teria conseguido sem ele.

Agradeço a meus colegas do Instituto de Informática, pela colaboração em diversos trabalhos e pelas trocas de ideias. Agradeço também aos professores do Instituto de Informática pela ajuda prestada nos momentos de dúvidas.

Agradecimentos a minhas orientadoras pela ajuda constante durante todo o caminho até a conclusão deste trabalho. E agradeço também ao CNPq pela bolsa de estudos durante o primeiro ano do curso de mestrado.

“All our knowledge has its origin in our perceptions”

— LEONARDO DA VINCI

RESUMO

Com o avanço da Internet pelo mundo, as pessoas passaram a interagir cada vez mais com a Web, principalmente após o surgimento das redes sociais, criando conteúdo que pode ser explorado de diversas formas. Esse aumento de usuários tem sido global, ou seja, pessoas de diversos países passaram a produzir textos de diversos idiomas. Esses textos compõem um rico conteúdo para Análise de Sentimentos Multilíngue. A maior parte dos trabalhos da área se foca em Mineração de Opinião, analisando o sentimento através da polaridade. Outro tipo de sentimento que tem atraído atenção é a emoção, embora não seja amplamente explorada a Análise de Sentimentos Multilíngue usando emoção.

Este trabalho utiliza técnicas geralmente usadas para Mineração de Opinião e polaridade para Análise de Sentimentos Multilíngues usando emoção. O objetivo deste trabalho é comparar diferentes combinações de aprendizado de máquina supervisionado e tradução automática para criar *corpora* em diferentes idiomas a partir de *corpora* anotados já existentes. As duas formas de utilizar as traduções comparadas são: criando classificadores de emoção separados por idiomas, chamados monolíngues, e criando um classificador composto do idioma original e das traduções, chamado multilíngue. É feito ainda um experimento cruzando dois *corpora*, visando avaliar o uso da tradução de um *corpus* com os textos originais do outro.

Os resultados dos experimentos mostram não apenas o sucesso de analisar emoção usando aprendizado supervisionado e tradução automática, mas que o classificador multilíngue supera os classificadores monolíngues. O experimento cruzando os *corpora* mostra que para algumas emoções os *corpora* estão alinhados, mas que para outras é preciso que haja maior similaridade nos textos.

Palavras-chave: Análise de sentimentos. Análise de sentimentos multilíngues. Emoção. Mineração de emoções.

Comparative Experiments Combining Supervised Learning and Machine Translation for Multilingual Emotion Mining

ABSTRACT

With the growth of the Internet around the world, people began to interact more and more with the Web, especially after the emergence of social networks, creating content that can be exploited in several ways. This increase in the number of users has been global, that is, people from different countries started producing texts in several languages. These texts comprise a rich content for Multilingual Sentiment Analysis. Most of the work in the area focus in Opinion Mining, analyzing the feeling through polarity. Another type of feeling that has attracted attention is emotion, although not extensively explored in Multilingual Sentiment Analysis.

This work uses techniques commonly used for Opinion Mining and polarity for Multilingual Sentiment Analysis using emotion. The objective of this study is to compare different combinations of supervised machine learning and automatic translation to create corpora in different languages from existing annotated corpora. The two ways to use the translations compared are: creating emotion classifiers separated by languages, called monolingual, and creating a composed classifier, with the original language and it's translations, called multilingual. An experiment crossing the two corpora used is made, to evaluate the use of the translation of one corpus with the original texts of the other.

The results of the experiments show not only the success of analysing emotion using supervised machine learning and automatic translation, but that the multilingual classifier exceeds the monolingual classifiers. The experiment crossing the corpora shows that to some emotions the corpora are aligned, but for others there needs to be greater similarity in the texts.

Keywords: Sentiment Analysis, Multilingual Sentiment Analysis, Emotion, Emotion Mining.

LISTA DE ABREVIATURAS E SIGLAS

AS	Análise de Sentimentos
ASM	Análise de Sentimentos Multilíngues
LSA	*Latent Semantic Analysis*
MO	Mineração de Opinião
MOM	Mineração de Opinião Multilíngue
PLN	Processamento de Linguagem Natural
RBF	*Radial Basis Function*
SMO	*Sequential Minimal Optimization*
SVD	*Single Value Decomposition*
SVM	*Support Vector Machine*
TF-IDF	*Term Frequency–Inverse Document Frequency*

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O crescimento da Internet nos últimos anos resultou em um aumento gigantesco na quantidade de usuários da rede, que passaram a interagir com ela de outras formas, não só consumindo informações, mas também as produzindo em grande escala. Hoje as pessoas estão conectadas 24 horas por dia através de seus *smartphones*, tablets, computadores, videogames, *smart watches*, entre outros dispositivos, interagindo constantemente entre si, em especial através de redes sociais como Twitter, Facebook, Instagram, além de blogs, gerando conteúdo que pode ser explorado de diferentes formas.

Análise de Sentimentos (AS) é a área de estudo que analisa as emoções, sentimentos, opiniões e posição/postura de alguém expressos em textos (LIU, 2015). Pode-se avaliar o sentimento através de valência, polaridade, emoção ou outros estados afetivos (MOHAMMAD, 2015b). Em geral, valência e polaridade classificam o sentimento em positivo, negativo ou neutro; já emoção e outros estados afetivos classificam a emoção em categorias básicas (como alegria, frustração e raiva) ou outras formas de mensuração.

A área de AS ganhou força nas pesquisas com trabalhos de Mineração de Opinião (MO), com destaque para os trabalhos que tratam de revisões de produtos (PANG; LEE, 2008), além de outras aplicações, como popularidade de marcas e pessoas (MOSTAFA, 2013), sistemas de recomendações (CHEN; WANG, 2014), previsão de indicadores a partir de sentimentos como resultados de eleições (TSAKALIDIS et al., 2015; TUMITAN; BECKER, 2014) ou movimentos da bolsa de valores (NJOLSTAD et al., 2014). As abordagens mais usadas para AS são aprendizado de máquina e uso de dicionários de sentimentos (LIU, 2012; TSYTSARAU; PALPANAS, 2012), sendo que ambas dependem de recursos: o aprendizado de máquina necessita de *corpora* anotados para treinamento, enquanto a segunda abordagem requer dicionários de sentimentos. Um dos fatores mais importantes para a AS produzir bons resultados é a qualidade dos recursos disponíveis, como por exemplo, *corpora* manualmente anotados.

Dentre os desafios de AS encontra-se a análise de textos em diferentes idiomas - por exemplo, sites de lojas com revisões de produtos em diferentes idiomas, reações a notícias em diferentes países, entre outros. Considerando as abordagens para AS citadas anteriormente, trabalhar com diferentes idiomas traz um desafio extra: a grande maioria dos recursos se concentra no idioma inglês, incluindo *corpora* anotados e dicionários de sentimento. Assim surgiu a Análise de Sentimentos Multilíngues (ASM), que propõe técnicas para classificação de sentimentos expressados em textos independente do idioma. Trabalhos de ASM focam-se na geração de recursos para a AS em idiomas diferentes de inglês (BANEA et al., 2008; SCHULZ;

WOMSER-HACKER; MANDL, 2010; MOLINA-GONZÁLEZ et al., 2013; STEINBERGER et al., 2012), técnicas para a AS que não dependam de características do idioma (LIN et al., 2014; SOLAKIDIS; VAVLIAKIS; MITKAS, 2014; GO; BHAYANI; HUANG, 2009), uso de tradução automática (BAUTIN; VIJAYARENU; SKIENA, 2008; BANEA; MIHALCEA; WIEBE, 2010; BALAHUR et al., 2014), uso de *corpora* paralelos (ALMEIDA et al., 2015) e AS *cross-lingual*[1] (WAN, 2011; HAJMOHAMMADI; IBRAHIM; SELAMAT, 2014).

A vasta maioria dos trabalhos de AS e de ASM trabalha com sentimento através de polaridade. Entretanto, outro tipo de sentimento tem atraído bastante atenção e se tornando foco de estudos: emoção. A partir de 2015 os japoneses contam com Pepper[2], o primeiro robô humanoide desenvolvido para convívio com humanos, capaz de reconhecer e reagir às emoções expressas por quem está falando com ele – este é um exemplo de como aplicações envolvendo emoções estão se alastrando para o nosso cotidiano.

Emoção foi reconhecida como um tipo de sentimento mais discriminatório do que a polaridade em aplicações envolvendo previsões com base em sentimentos (ASUR; HUBERMAN, 2010; TUMASJAN et al., 2010; BOLLEN; MAO; ZENG, 2011), identificação de condições psicológicas e comparação de obras literárias e estilos de escrita (MOHAMMAD, 2012). Uma opinião geralmente expressa um sentimento direcionado a um alvo específico, que pode ser avaliada por polaridade; já emoção nem sempre é associada a um alvo ou representa uma postura (TSYTSARAU; PALPANAS, 2012). Ainda, identificar emoções em um texto é uma tarefa mais complicada do que identificar a polaridade de uma opinião por não haver relação de oposição entre as emoções, ou seja, a presença de uma emoção não exclui a presença de outras, tonando-as independentes. Além desses fatores, não há um consenso quanto a quais seriam as emoções consideradas básicas (ORTONY; TURNER, 1990), tornando-se uma desvantagem quando se considera que os recursos disponíveis voltados a emoções são dispersos em diferentes formas de classificar e mensurar emoções.

Todas essas áreas de estudo possuem suas particularidades e não se cruzam com frequência. Este trabalho visa preencher a lacuna que outros trabalhos deixaram em aberto, que é a Mineração de Emoções em Textos Multilíngues. Os recursos para classificação de emoções já são escassos para o idioma inglês, quando comparados com os recursos voltados à polaridade, sendo mais escassos ainda para outros idiomas. Considerando os recursos disponíveis para AS envolvendo emoções, o presente trabalho adapta abordagens de ASM que usam polaridade para

[1] A tradução literal para *cross-lingual* é cruzamento de línguas/idiomas, entretanto, essa tradução não é utilizada com frequência no idioma português, sendo utilizado o termo multilíngue. Como em inglês *cross-lingual* e *multilingual* não são sinônimos, este trabalho vai utilizar o termo original *cross-lingual*.

[2] https://www.aldebaran.com/en/a-robots/who-is-pepper

classificar emoções.

Este trabalho classifica emoções em textos de diferentes idiomas usando aprendizado de máquina supervisionado em conjunto com tradução automática para criar *corpora* anotados em idiomas com escassez de recursos. Entre os objetivos deste trabalho estão:

1. Comparar as formas de usar os *corpora* gerados com tradução automática, entre treinar um classificador para cada idioma (classificadores monolíngues para os idiomas originais e traduzidos) ou combinando todos os idiomas para treinar um único classificador (classificador multilíngue).
2. Comparar os efeitos das diferentes opções de pré-processamento e algoritmos, verificando se há diferença significativa entre elas, considerando tanto os classificadores monolíngues quanto o multilíngue.
3. Comparar o desempenho entre duas direções de tradução: a tradução das instâncias de treinamento do classificador para o idioma das instâncias de teste contra a tradução das instâncias de teste para o idioma do classificador.

As hipóteses trabalhadas nesta dissertação são: a) tradução automática produz *corpora* satisfatórios para ASM utilizando emoção como tipo de sentimento; b) a combinação de textos de diferentes idiomas, mesmo que provenientes de tradução automática, produz resultados melhores para ASM do que o uso individual de cada idioma; e c) combinações de pré-processamento de texto e algoritmos de classificação que melhoram os resultados da ASM usando polaridade como tipo de sentimento também melhoram os resultados quando o tipo é emoção; e d) os *corpora* gerados a partir de tradução automática podem ser utilizados para ASM com outros *corpora* em idiomas originais sem grandes perdas.

Para realizar as comparações, adaptou-se a função *scp* (*sentiment classification performance*) proposta por Balahur e Turchi (2014), que visa analisar o desempenho da classificação baseada em tradução mediante uma combinação de parâmetros relativos a pré-processamento, algoritmos de classificação e desenho experimental. A função adaptada é denominada *ecp* (*emotion classification performance*), e seus parâmetros são constituídos das diferentes opções de pré-processamentos, algoritmos e estrutura avaliados nos experimentos, ampliando as opções que são utilizadas na função *scp*.

São usados dois *corpora* anotados em idiomas originais distintos, que são traduzidos para outros dois idiomas, além do idioma do outro *corpus*, totalizando quatro idiomas analisados. Os resultados obtidos sobre os *corpora* no idioma original são comparados com os resultados obtidos com as traduções geradas a partir dele, visando verificar se há diferença significativa

entre o uso dos textos originais e o uso das traduções. A abordagem usada é adaptável para mais idiomas, bastando que haja suporte do tradutor automático aos idiomas escolhidos, visto que não faz uso de recursos de Processamento de Linguagem Natural (PLN)[3] específicos para os idiomas (por exemplo, etiquetadores morfossintáticos).

Para atingir os objetivos desta dissertação, são apresentados dois experimentos: no primeiro, é comparado o uso de um classificador de emoção para cada idioma com um classificador de emoção multilíngue treinado com os textos em todos os idiomas considerados, usando dois *corpora* separadamente; o segundo experimento faz cruzamento dos *corpora*. No segundo experimento, utiliza-se a tradução de um *corpus* para o idioma do outro, comparando a tradução de um classificador para o idioma das instâncias de teste e a tradução das instâncias de teste para o idioma do classificador pronto. Os resultados apresentados são obtidos de dois *corpora* anotados originais em idiomas distintos, mostrando que a abordagem utilizada independe de idioma. Os resultados do primeiro experimento mostram o sucesso de classificar emoções em diferentes idiomas usando tradução automática e aprendizado supervisionado, permitindo a criação de recursos em outros idiomas; considerando as diferentes formas de usar as traduções geradas, o classificador multilíngue obtém melhores resultados que os classificadores monolíngues, mostrando que a combinação dos idiomas é benéfica para a classificação de emoções. Assim, o primeiro experimento confirma as três primeiras hipóteses consideradas. Os resultados do segundo experimento apresentam resultados bem heterogêneos, dependendo da emoção analisada, consequência das diferenças entre os *corpora*. Devido aos resultados inconclusivos, não é possível confirmar ou rejeitar a hipótese (d).

As contribuições deste trabalho são:

a) Mostrar que é possível produzir recursos para AS voltados a emoções em diferentes idiomas, através da aplicação de abordagens de ASM amplamente usadas com polaridade para análise de emoções.
b) Mostrar que a combinação de idiomas é benéfica para classificação de emoções em ASM.
c) Mostrar que combinações específicas de pré-processamento do texto e algoritmos produzem melhores resultados na classificação de emoções dos textos.

O restante do trabalho se divide da seguinte forma: inicialmente é apresentada uma introdução de AS, com conceitos básicos da área no Capítulo 2; o Capítulo 3 apresenta a ASM e seus principais desafios, com trabalhos relacionados; o Capítulo 4 apresenta a proposta para Mineração de Emoção em Textos Multilíngues e os experimentos realizados. O Capítulo 5 encerra

[3] Em inglês, NLP (*Natural Language Processing*)

este trabalho com a Conclusão e possibilidades de trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

A Análise de Sentimentos é a área de estudo que se propõe a analisar opiniões, sentimentos, avaliações, posturas e emoções das pessoas na linguagem escrita (LIU, 2012). Neste capítulo serão descritos os conceitos necessários para melhor compreensão deste trabalho.

2.1 Sentimento

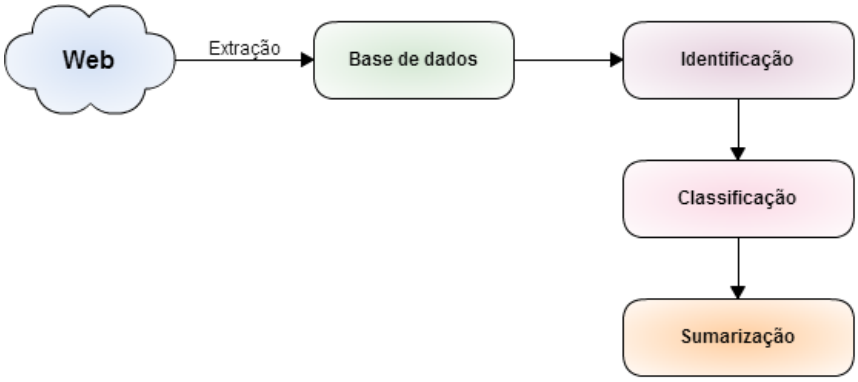
Sentimentos e os conceitos relativos a eles, como avaliação, postura, emoção e humor, são relacionados a nossas crenças e sensações (LIU, 2015). O conceito de sentimento pode ser usado de diferentes formas, não havendo consenso entre os autores na literatura estudada sobre uma forma padrão. Neste trabalho, é considerado que existem diferentes tipos de sentimento, que podem ser avaliados de formas distintas. Algumas das formas de avaliação amplamente utilizadas na literatura para determinar o sentimento são valência (ou polaridade), emoção e força (BRAVO-MARQUEZ; MENDOZA; POBLETE, 2013). O uso de polaridade é o mais simples, visto que posiciona o sentimento em até três categorias: positiva, negativa ou neutra. Quando se usa emoção, é possível classificar o sentimento em categorias como tristeza, alegria, surpresa, entre outras, que são definidas no escopo da análise. A força define níveis de intensidade para o tipo de sentimento, podendo ela ser emoção ou polaridade.

A AS pode ser feita em diferentes níveis, segundo Liu (2012): a) nível de documento, no qual se avalia o sentimento geral expresso em um texto; b) nível de sentença, no qual se avalia o sentimento presente em cada sentença do texto, permitindo tratar diversos assuntos em um texto de entrada; e c) nível de entidade e aspecto, no qual procura-se identificar especificamente a quem ou o que se referem os sentimentos expressados no texto.

O processo de AS pode ser descrito em três etapas, segundo Tsytsarau e Palpanas (2012): identificação, classificação e sumarização. A Figura 2.1 apresenta um resumo sobre as etapas do processo, onde a entrada são documentos de textos.

A etapa de identificação tem por objetivo identificar os tópicos mencionados no texto (incluindo aspectos, caso existam) e tentar associá-los com o respectivo conteúdo subjetivo (BECKER; TUMITAN, 2013). A dificuldade desta etapa depende de fatores como o nível de análise escolhido e a fonte dos dados. Na etapa de classificação ocorre a definição da orientação do sentimento (por exemplo, entre positiva e negativa quando se usa polaridade). A classificação do sentimento pode ser feita por técnicas supervisionadas (com auxílio de exemplos de treinamento) ou não supervisionada (por exemplo, uso de dicionários de sentimentos ou abordagens estatísticas). A etapa de sumarização agrega os sentimentos analisados em um sumário, que

Figura 2.1: Etapas do processo de Análise de Sentimentos



Fonte: a autora

pode ser usado em recursos de visualização ou entrada para outras aplicações. O foco deste trabalho está na etapa de classificação do sentimento.

2.2 Emoções

Recentemente, houve um aumento nas pesquisas envolvendo emoções, especialmente na área de inteligência artificial. Uma das aplicações comerciais é o Pepper, citado anteriormente, projetado para ser companhia para pessoas, capaz de reconhecer emoções e responder de acordo com elas. O 6º *Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*[1] (WASSA 2015) contou com trabalhos para criação de recursos para análise de emoções em textos do Twitter (MOHAMMAD, 2015a; VILARES; ALONSO; GÓMEZ-RODRIGUEZ, 2015).

Trabalhos na área de AS também já usaram emoções para melhorar resultados de classificação. É o caso do trabalho de Bollen, Mao e Zeng (2011), que mostra que a predição do valor da bolsa de valores Dow Jones pode ser melhorada com a análise de emoções baseada em sentimento de *posts* do Twitter. O trabalho de Tumasjan et al. (2010) usa as análises de emoções presentes no LIWC (TAUSCZIK; PENNEBAKER, 2010) para analisar *tweets* referentes à eleição na Alemanha. Vilares, Alonso e Gómez-Rodrıguez (2015) utilizam emoções como parte do conjunto de atributos para melhorar a performance de classificação de polaridade em *tweets*. Em Mohammad (2015a), o tesauro de Roget é apresentado associado a um léxico de emoções.

Não há um consenso quanto a quais seriam as emoções básicas, e cada autor pode escolher o grupo de emoções que melhor se encaixe em seus objetivos (ORTONY; TURNER, 1990).

[1] http://optima.jrc.it/wassa2015/

Muitos trabalhos utilizam a classificação de Ekman (EKMAN, 1992a; EKMAN, 1992b), que considera como básicas as emoções *anger*, *disgust*, *fear*, *joy*, *sadness* e *surprise*[2], dentre eles (LIU; LIEBERMAN; SELKER, 2003; ALM; ROTH; SPROAT, 2005; NEVIAROUSKAYA; PRENDINGER; ISHIZUKA, 2007; MARTINAZZO; PARAISO, 2010; CHERRY; MOHAMMAD; BRUIJN, 2012; GHAZI; INKPEN; SZPAKOWICZ, 2014). Já o conjunto de emoções básicas de Plutchik é composto pelas seis emoções de Ekman e adiciona a esse conjunto duas emoções: *anticipation* e *trust*[3], sendo utilizado no NRC (MOHAMMAD; TURNEY, 2010) e em Brooks et al. (2013). O trabalho de Tumasjan et al. (2010) trabalha com diferentes dimensões, incluindo emoções positivas, emoções negativas, *sadness*, *anxiety*[4] e *anger*. Ainda, Vo e Collier (2013) utilizam para analisar *tweets* relacionados a terremotos as emoções *calm*, *upleasantness*, *sadness*, *anxiety*, *fear* e *relief*[5].

Emoção é considerada uma forma de avaliação mais rica e pode expressar melhor o sentimento contido em textos subjetivos. Outra vantagem de usar emoção como tipo de sentimento é a possibilidade de usar diferentes emoções, conforme o objetivo da análise: por exemplo, ao analisar comentários de um filme de terror, pode-se procurar saber se medo foi a emoção predominante nos comentários, assim como verificar em comentários de notícias de economia se as pessoas expressam confiança com o cenário econômico atual. Este trabalho utiliza o conjunto de emoções básicas de Ekman, devido à disponibilidade de *corpora* anotados e por ser um dos conjuntos de emoções mais utilizados na literatura pesquisada.

2.3 Recursos para Análise de Sentimentos

Existem diversos recursos para auxiliar as tarefas de AS. Entre eles estão *corpora* anotados, léxicos ou dicionários de sentimentos e recursos específicos para PLN. Nesta seção serão abordados os recursos mais utilizados para AS.

Um *corpus* anotado é um conjunto de textos que contêm informações adicionais - essas informações adicionais são os rótulos, usados para a tarefa de classificação. Os *corpora* anotados podem ser usados para treinamento de classificadores (isto é, os textos que compõem o *corpus* contêm a informação da classe que representam) e para *gold standard*, que é considerado um conjunto com os melhores resultados possíveis para o classificador, sendo utilizado para comparação dos resultados do sistema testado. *Corpora* anotados são recursos difíceis de se obter,

[2] Traduções para o português, respectivamente: raiva, nojo, medo, alegria, tristeza e surpresa
[3] Traduções para o português, respectivamente: expectativa e confiança
[4] Tradução para o português: ansiedade
[5] Traduções para o português, respectivamente: calma, desagrado, tristeza, ansiedade, medo e alívio

devido ao grande esforço que deve ser empreendido na sua criação: é preciso que um conjunto de instâncias seja manualmente analisado por pessoas, geralmente removendo as instâncias que não atingiram um consenso entre os anotadores, ou que seja usado um sistema de anotação automática. Diversos *corpora* são disponibilizados para pesquisa, estando anotados ou não, para diferentes tarefas, como por exemplo, para o SemEval[6]. A maioria dos *corpora* anotados disponíveis estão no idioma inglês, além de serem anotados para polaridade. A Amazon oferece, através do Mechanical Turk[7], anotação de *corpora*, dentre outros serviços que necessitem esforço humano. Há ainda os *corpora* paralelos, nos quais um texto é acompanhado de suas traduções alinhadas. A escassez de *corpora* anotados para emoções é alta: devido ao uso de polaridade ser mais simples, a vasta maioria dos recursos para AS se concentra nela; ainda, os *corpora* voltados a emoções se encontram majoritariamente em inglês, havendo pouca variedade disponível em outros idiomas.

Dicionários de sentimentos reúnem uma extensa variedade de termos e rótulos, ou pesos que representam o sentimento associado, conforme o objetivo do dicionário (polaridade ou emoções, por exemplo). Dentre os dicionários mais utilizados para a AS se encontram SentiWordNet (ESULI; SEBASTIANI, 2006), WordNet Affect (STRAPPARAVA; VALITUTTI et al., 2004) e LIWC, no idioma inglês, e SentiLex-PT (SILVA; CARVALHO; SARMENTO, 2012), no idioma português, sendo o foco destes a polaridade, e o NRC e a expansão do ANEW disponibilizada por Warriner, Kuperman e Brysbaert (2013), em inglês e voltados para emoção. A Figura 2.2 mostra um trecho do dicionário NRC contendo alguns termos e os pesos a eles associados (1 = presença e 0 = ausência da emoção no termo para o conjunto de emoções de Plutchik). Considerando a variedade de emoções que podem ser usadas para AS, os poucos dicionários disponíveis são dispersos entre diferentes conjuntos de emoções, e concentram-se predominantemente no idioma inglês.

Figura 2.2: Trecho do dicionário NRC

rattlesnake	positive:0	negative:0	anger:0	anticipation:0	disgust:0	fear:1	joy:0	sadness:0	surprise:0	trust:0
raucous	positive:0	negative:1	anger:0	anticipation:0	disgust:0	fear:0	joy:0	sadness:0	surprise:0	trust:0
rave	positive:1	negative:1	anger:1	anticipation:0	disgust:1	fear:0	joy:1	sadness:0	surprise:1	trust:1
raven	positive:0	negative:0	anger:0	anticipation:0	disgust:0	fear:0	joy:0	sadness:0	surprise:0	trust:0
ravenous	positive:0	negative:1	anger:1	anticipation:0	disgust:0	fear:1	joy:0	sadness:1	surprise:0	trust:0
ravine	positive:0	negative:0	anger:0	anticipation:0	disgust:0	fear:1	joy:0	sadness:0	surprise:0	trust:0
raving	positive:0	negative:1	anger:1	anticipation:1	disgust:0	fear:1	joy:1	sadness:0	surprise:1	trust:0
rawhide	positive:0	negative:0	anger:0	anticipation:0	disgust:0	fear:0	joy:0	sadness:0	surprise:0	trust:0

Fonte: dicionário NRC (MOHAMMAD; TURNEY, 2010)

[6] https://en.wikipedia.org/wiki/SemEval
[7] https://www.mturk.com/mturk/welcome

Como os objetos de análise são textos, é necessária uma série de recursos para prepará-los para a etapa de classificação, geralmente da área de PLN. Entre os recursos específicos para PLN mais comuns se encontram, além de outros:

- Tokenizadores: separam uma sequência de textos em partes menores, como palavras, em unidades chamadas tokens para representação dos atributos. Podem ser de diferentes tipos, como alfabéticos ou n-gramas.
- Etiquetador morfossintático[8]: determina a classe gramatical de cada palavra.
- *Stemming*: reduz termos flexionados ou de diferentes formas a um mesmo radical.
- Remoção de *stopwords*: retira do texto termos comuns que não agregam significado ao texto, como artigos e preposições.

Também são amplamente usados em análise de textos *weighting schemes*, que atribuem pesos aos termos nos textos, de acordo com sua relevância. Entre as opções mais comuns estão booleano, que considera apenas a presença (peso 1) ou ausência (peso 0) dos termos nas instâncias, e TF-IDF, que reflete quão importante um termo é em um *corpus*, atribuindo peso maior a termos que podem determinar uma instância e que não sejam tão frequentes no *corpus*. Outro processo utilizado é a seleção de atributos, visando selecionar apenas um subconjunto de atributos que seja mais relevante para a tarefa de classificação, removendo atributos irrelevantes ou redundantes. Um dos tipos de seleção de atributos mais utilizados é *Information Gain* (InfoGain), que tem por objetivo manter os termos que possuem correlação positiva com a classe.

Este trabalho utiliza dois *corpora* anotados em diferentes idiomas, além de usar tokenizadores, remoção de *stopwords*, *weighting schemes* e seleção de atributos.

2.4 Abordagens para Análise de Sentimentos

As abordagens de AS podem ser classificadas em quatro categorias: léxica (ou uso de dicionários de sentimentos), aprendizado de máquina, estatística e semântica (TSYTSARAU; PALPANAS, 2012), sendo que a maioria dos trabalhos se encontra nas duas primeiras categorias.

A abordagem léxica é também conhecida por baseada em dicionários. Essa abordagem requer que, para cada palavra ou expressão do texto, verifique-se a orientação da mesma no dicionário. Logo, o principal ponto desta abordagem depende da criação dos dicionários ou

[8] Em inglês, *part of speech tagging*, ou *POS tagging*

que já estejam prontos - a seção 2.3 apresentou alguns exemplos de dicionários existentes e amplamente usados.

Tsytsarau e Palpanas (2012) afirmam que a abordagem usando aprendizado de máquina é uma solução sofisticada para o problema de classificação, podendo ser descrita como um processo em dois passos: (1) aprender um modelo a partir de um *corpus* de treinamento e (2) classificar os dados não vistos de acordo com o modelo treinado. Assim, a abordagem de aprendizado de máquina envolve inicialmente um treinamento com um *corpus*, sendo ele anotado (para o aprendizado supervisionado) ou não (aprendizado não-supervisionado), que irá gerar um modelo de classificação - alguns trabalhos utilizam também aprendizado semi-supervisionado, onde apenas uma parte do *corpus* é anotada.

Em Santos, Becker e Moreira (2014b) a abordagem léxica foi comparada com abordagem de aprendizado de máquina, sendo esta última efetivamente superior. O dicionário de sentimento é limitado aos termos presentes no dicionário escolhido, o que impacta no desempenho da classificação de emoções dependendo do *corpus*. Outro fator limitante é a escassez de uma maior variedade de dicionários contendo termos rotulados para emoções mesmo em inglês, sendo a escassez da disponibilidade para outros idiomas ainda maior. Considerando as limitações da abordagem, o uso de dicionário de sentimentos foi descartado.

2.4.1 Aprendizado supervisionado

Nos trabalhos de AS, a abordagem mais utilizada é a do tipo supervisionado, utilizando classificação. Dentre os algoritmos mais usados na área de AS para criar o modelo do classificador estão Naïve Bayes e *Support Vector Machine* (SVM).

Naïve Bayes, da área dos algoritmos probabilísticos, é um dos algoritmos mais simples de classificação e frequentemente usado na área de MO por atingir bons resultados (PANG; LEE; VAITHYANATHAN, 2002; DEMIRTAS, 2013). Sua base é a aplicação do Teorema de Bayes, no qual as variáveis são independentes entre si.

Support Vector Machine (SVM) é um algoritmo linear binário, usado para classificação e regressão. O modelo SVM representa as instâncias como pontos no espaço e busca encontrar um hiperplano (ou conjunto de hiperplanos) que separe as instâncias por classes. Dentre as opções, encontra-se a variação otimizada SMO (*Sequential Minimal Optimization*). É utilizado nos trabalhos (BALAHUR et al., 2013; HAJMOHAMMADI; IBRAHIM; SELAMAT, 2014).

Outra opção de algoritmo no aprendizado supervisionado, que não é tão utilizada quanto Naïve Bayes ou SVM, é o *Radial Basis Function* (RBF), que pertence à família dos algoritmos

de redes neurais. O modelo é representado por um conjunto de nodos interconectados através de pesos. Os pesos são definidos através de um processo adaptativo durante o treinamento, e representam a força da conexão entre os neurônios. As redes RBF geralmente possuem três camadas: a camada de entrada, a camada oculta com uma função de ativação RBF não-linear e a camada de saída.

O trabalho de Balahur e Turchi (2014) utiliza em conjunto com SVM os meta algoritmos *Adaptative Boosting* (ou AdaBoost) e Bagging, que combinam decisões de múltiplos classificadores chamados de fracos ("*weak learners*"), mas que trabalham com estratégias diferentes. No AdaBoost, as saídas dos classificadores fracos são combinadas em uma soma através de pesos que representa a saída final do classificador Boosting. O AdaBoost é adaptativo no sentido de que os classificadores fracos são ajustados a favor das instâncias classificadas negativamente por classificações anteriores, diminuindo as chances de haver sobreajuste (*overfitting*), que é o ajuste excessivo ao conjunto de treinamento. Já no Bagging, cada modelo na combinação tem voto com o mesmo peso. Para que haja variância no modelo, o Bagging treina cada modelo usando uma parte do conjunto de treinamento sorteada aleatoriamente.

Esta dissertação faz uso de aprendizado de máquina do tipo supervisionado, devido aos bons resultados que tem obtido em trabalhos de AS. Neste trabalho, são utilizados os algoritmos Naïve Bayes e SVM, pelo sucesso demonstrado em trabalhos da área, e RBF, visando utilizar um algoritmo de família diferente dos anteriores, além dos meta algoritmos AdaBoost e Bagging, cada um com as opções de algoritmos base RBF e SVM.

2.4.2 Medidas de avaliação

Em tarefas de classificação, os termos Verdadeiro Positivo, Verdadeiro Negativo, Falso Positivo e Falso Negativo comparam os resultados do classificador testado com resultados considerados verdadeiros. Os possíveis resultados se encontram na Tabela 2.1, que representa a matriz de confusão:

Tabela 2.1: Matriz de confusão

		Classe real	
		Positivo	**Negativo**
Classe Prevista	**Positivo**	Verdadeiro Positivo (VP)	Falso Positivo (FP)
	Negativo	Falso Negativo (FN)	Verdadeiro Negativo (VN)

Fonte: a autora

Os resultados dos experimentos reportados utilizam as medidas de avaliação Precisão, Revocação e Medida-F (em inglês, respectivamente, *Precision*, *Recall* e *F-Measure*).

Precisão indica a quantidade de instâncias que o sistema classificou corretamente de todas as instâncias classificadas para determinada classe, calculada pela fórmula 2.1; Revocação indica a quantidade de instâncias corretamente classificadas de todas as instâncias pertencentes à determinada classe, calculada pela fórmula 2.2; Medida-F considera tanto a precisão quanto a revocação para o seu cálculo, geralmente sendo a média harmônica entre elas, sendo calculada pela fórmula 2.3.

$$\text{Precisão} = \frac{VP}{VP + FP} \tag{2.1}$$

$$\text{Revocação} = \frac{VP}{VP + FN} \tag{2.2}$$

$$\text{Medida-F} = 2 \times \frac{\text{Precisão } \times \text{ Revocação}}{\text{Precisão } + \text{ Revocação}} \tag{2.3}$$

Para comparar os resultados obtidos pode ser aplicado o teste-t bicaudal, usado geralmente com confiança de 0,95, para verificar se as diferenças obtidas são estatisticamente significativas ou não.

3 ANÁLISE DE SENTIMENTOS MULTILÍNGUES

Neste capítulo são introduzidos a motivação da ASM e seus desafios. São também discutidos os trabalhos relacionados pesquisados para a proposta deste trabalho.

3.1 Motivação para Análise de Sentimentos Multilíngues

A Web tem sido muito explorada como fonte de sentimentos, pela quantidade e riqueza do conteúdo que disponibiliza. Entretanto, o inglês ainda é o idioma dominante no conteúdo disponibilizado, criando duas consequências: a) os recursos e técnicas disponíveis para AS são voltados prioritariamente ao idioma inglês; e b) a oportunidade de processar indistintamente sentimentos expressos em diferentes idiomas, seja pela ausência de conteúdo em um idioma específico, ou como forma de comparar sentimentos expressos em diferentes contextos culturais.

A ASM propõe técnicas para classificação de sentimentos sem um idioma alvo específico (BANEA et al., 2008). A maioria dos trabalhos em ASM trata da classificação de sentimento de acordo com o conceito de polaridade (BANEA et al., 2008; NARR; HULFENHAUS; ALBAYRAK, 2012; BADER et al., 2011; LIN; TAN; CHENG, 2011), enquanto a classificação de emoção em textos multilíngue é uma área pouco explorada (SANTOS; BECKER; MOREIRA, 2014a; SANTOS; BECKER; MOREIRA, 2014b).

A maioria dos trabalhos nesta área propõem-se a: a) desenvolver técnicas para fazer a AS sem a necessidade de recursos voltados a um idioma específico; ou b) criar recursos para análise em idiomas alvo a partir de recursos existentes em outros idiomas. Outras opções envolvem o uso de *corpora* paralelos e AS *cross-lingual*.

3.2 ASM sem o uso de recursos linguísticos específicos

Para a ASM ser feita sem o uso de recursos linguísticos específicos, como dicionários de sentimentos ou *corpora* anotados, pode-se fazer uso de características que expressem os mesmos sentimentos em diferentes idiomas. Um típico exemplo de tais características são *emoticons*, que se tornaram extremamente populares e são usados de maneira uniforme por usuários de todo mundo para expressar sentimentos com quantidade limitada de caracteres. Por exemplo, *tweets* são rotulados para polaridade automaticamente como positivos ou negativos por Go, Bhayani e Huang (2009) para o idioma inglês, Alves et al. (2014) para o idioma português e Narr, Hulfenhaus e Albayrak (2012) para os idiomas inglês, francês, alemão e português.

Narr, Hulfenhaus e Albayrak (2012) utilizam *emoticons* na anotação de *tweets* para criar um

corpus de treinamento para um classificador de polaridade. É criado um classificador para cada idioma (inglês, francês, alemão e português), além de um classificador multilíngue combinando todos os idiomas, treinados com o algoritmo Naïve Bayes. Os resultados mostram que a abordagem atinge bons resultados, e que o classificador multilíngue possui uma performance um pouco abaixo dos classificadores monolíngues. Como este trabalho utiliza *corpora* já anotados que não possuem *emoticons* ou outra característica do tipo, essa não é uma abordagem viável à nossa proposta.

Outra opção é a criação de dicionários de sentimentos a partir de algumas palavras sementes e da aplicação de algoritmos não-supervisionados (LIN et al., 2014; SOLAKIDIS; VAVLIAKIS; MITKAS, 2014). Lin et al. (2014) propõem um framework para análise de sentimentos multilíngues a partir de dados não rotulados, criando um léxico de polaridade a partir de algumas palavras sementes (advérbios de intensidade) que podem ser facilmente adaptadas para outros idiomas. Os idiomas considerados são inglês, francês e alemão. Algumas instâncias são rotuladas através de método não supervisionado, e as instâncias com maior confiança na classificação são selecionadas para treinar um classificador multilíngue. Como citado na Seção 2.4.1, o uso de dicionários de sentimentos não alcançou os resultados obtidos com aprendizado supervisionado, sendo descartado neste trabalho.

3.3 Criação de recursos para outros idiomas a partir de recursos já existentes

Dentre os recursos para AS disponíveis, como dicionários de sentimentos e *corpora* anotados, a maioria se encontra disponível apenas no idioma inglês, como o dicionário NRC, o *corpus* MPQA, entre outros recursos[1]. Uma das tarefas mais comuns de ASM é a criação de recursos para idiomas que possuem escassez de recursos para AS a partir dos recursos já existentes, geralmente voltados ao idioma inglês. Uma das formas mais comuns de fazer isso é através de tradução automática.

A tradução automática pode ser aplicada tanto sobre o texto sendo analisado quanto nos recursos de apoio à análise, tais como *corpora* anotados (BANEA et al., 2008) e dicionários especializados (MOLINA-GONZÁLEZ et al., 2013; SANTOS; BECKER; MOREIRA, 2014a). A vantagem desta abordagem é sua simplicidade, visto que existem diversos serviços de tradução automática de qualidade, muitos deles gratuitos e disponíveis na Web. A tradução automática já é considerada por muitos um método maduro para a criação de recursos para AS em outros idiomas (BALAHUR; TURCHI, 2014; SCHULZ; WOMSER-HACKER; MANDL, 2010).

[1] Disponíveis em http://mpqa.cs.pitt.edu/

Em (SANTOS; BECKER; MOREIRA, 2014a), foi utilizada a combinação de tradução automática com o uso de dicionário de sentimentos para classificação de emoções em diferentes idiomas. O *corpus* de teste se encontra no idioma português, enquanto o dicionário utilizado (NRC) contém termos em inglês. O objetivo era verificar qual tradução produzia os melhores resultados: se as instâncias para inglês ou se os termos do dicionário para português. Os resultados mostraram que a tradução das instâncias produziu melhores resultados, visto que a instância provê um contexto para a tradução, o que os termos do dicionário não provêm.

Molina-González et al. (2013) criam um dicionário de sentimentos para o idioma espanhol a partir de um dicionário existente em inglês. O dicionário Bing Liu English Lexicon (HU; LIU, 2004) é traduzido para o espanhol com o tradutor Reverso[2] e então revisado manualmente de forma a adicionar diferentes melhorias. A comparação do uso do dicionário melhorado com outros recursos para o espanhol já existentes mostra que os resultados obtidos com o novo dicionário superam os recursos existentes.

Entre os trabalhos com destaque para esta tarefa se encontram Banea et al. (2008) e Banea, Mihalcea e Wiebe (2010), nos quais é utilizada tradução automática em um *corpus* em inglês para gerar *corpora* de treinamento em espanhol e romeno, por exemplo. Ao analisar a combinação de idiomas em um mesmo classificador, para o uso de polaridade, Banea, Mihalcea e Wiebe (2010) mostram que os resultados de classificação melhoram quando se usam mais idiomas, visto que mais idiomas auxiliam na desambiguação de sentido de alguns termos. Nos experimentos apresentados nessa dissertação, procura-se verificar se esse resultado também é observado para o uso de emoção.

Balahur e Turchi (2014) propõem avaliar a combinação de tradução automática e aprendizado de máquina para ASM, junto com o impacto da tradução nos resultados, ou seja, se há correlação entre a qualidade da tradução e os resultados da classificação de sentimento medida em polaridade. Um dos principais objetivos do trabalho é avaliar o impacto que erros causados pela tradução automática podem causar na classificação de sentimentos em diferentes idiomas. Para isso, são comparadas duas formas de usar as traduções geradas: avaliando as traduções de cada tradutor separadamente e combinando as traduções de cada tradutor por idioma. Os idiomas considerados são inglês (da base de dados), e traduções para francês, alemão e espanhol. Foi definida a função de desempenho *sentiment classification performance scp*, com o propósito de maximizar o desempenho da classificação de sentimento, avaliado por Medida-F, escolhendo uma combinação ótima de parâmetros. A definição da função *scp* é $scp(fs, fr, l, ed)$, na qual: *feature set* (fs) indica o modo de representação dos atributos, cujas opções eram unigramas ou

[2] www.reverso.net

unigramas+bigramas; *feature representation* (fr) indica o tipo de peso (ou *weighting scheme*) a ser atribuído aos termos, se booleano ou TF-IDF; *learning algorithm* (l) indica o algoritmo usado para treinamento do classificador, se SVM SMO ou combinação de meta classificadores com SVM; *experiment design* (ed) indica o modo como os dados de teste e treinamento são divididos.

Dois grupos de experimentos for desenvolvidos por Balahur e Turchi (2014). No primeiro, usando os tipos de atributos citados anteriormente, um classificador é treinado a partir da tradução para cada idioma de cada um dos tradutores utilizados. É utilizado o algoritmo de classificação SVM, em isolado ou combinado com os meta classificadores Bagging e AdaBoost para atenuar o ruído gerado pela tradução. No segundo grupo, para cada idioma, são combinadas as traduções geradas por cada um dos sistemas de tradução automática para cada idioma. Para ambos os conjuntos, os resultados foram comparados com um *gold standard* para cada idioma. Os resultados mostraram que não há correlação entre a qualidade da tradução automática, medida pelo coeficiente BLEU, e os resultados da AS. O uso combinado de SVM com o meta classificador Bagging atingiu melhores resultados, considerando Medida-F ponderada, assim como o uso de unigramas com TF-IDF trouxe impacto positivo.

As traduções incorretas prejudicam a classificação do modelo SVM SMO, pois aumentam a quantidade de atributos e de esparsidade nos dados, dificultando a identificação de um hiperplano para separar as classes positiva e negativa. Logo, as traduções incorretas causaram uma queda na performance da classificação de sentimentos, visto que os atributos extraídos não são informativos o suficiente. O cenário do primeiro experimento, onde são usados os dados do mesmo tradutor, obteve melhores resultados, visto que os erros de tradução gerados se encontram tanto no conjunto de treinamento quanto no de testes, o que anula o efeito do erro no treinamento. A diferença de performance de classificação entre os dados no idioma original e suas traduções observada foi pequena, tornando tradução automática uma boa opção para geração de dados em outros idiomas.

Trabalhos como o de Balahur e Turchi (2014) mostram que a tradução automática é um recurso satisfatório para a criação de *corpora* em diferentes idiomas para ASM. A tradução automática é um dos recursos mais simples para criação de *corpora* em múltiplos idiomas, contando com diferentes opções de tradutores gratuitos na Web hoje e com suporte a diversos idiomas.

3.4 *Corpora* paralelos

Corpora paralelos são textos em um idioma acompanhados de sua tradução alinhada. Entre os exemplos mais usados de *corpus* paralelo se encontra o Europarl (KOEHN, 2005)[3].

Um exemplo do uso de *corpora* paralelos em conjunto com tradução automática em algum nível é o proposto por Steinberger et al. (2011), que visa a criação dicionários em outros idiomas com textos em inglês e espanhol para classificação de polaridade em nível de entidade. A partir de dicionários em inglês e espanhol, foram criados dicionários em outros idiomas usando tradução automática, e posteriormente melhorados manualmente. Os dicionários foram testados com um *gold standard* criado a partir do Europarl para os idiomas disponíveis (inglês, espanhol, francês, alemão, checo, italiano e húngaro). Os resultados foram promissores, chegando a uma concordância mínima de 66% para o idioma italiano e 74% para o idioma inglês.

Santos, Becker e Moreira (2014b) comparam duas abordagens para classificação de emoções em textos multilíngues usando um *corpus* paralelo composto pelos idiomas português e inglês. O uso da tradução automática é comparado em duas abordagens de AS: combinada com dicionário de sentimentos (NRC) e combinada com aprendizado supervisionado (uso do algoritmo Naïve Bayes). Compara-se também se um dos idiomas ou a combinação deles atinge resultados melhores. Os resultados mostraram que o uso de aprendizado de máquina obteve resultados consideravelmente superiores ao uso de dicionário de sentimento, visto que o dicionário é limitado pelos termos nele presentes. Quanto aos idiomas, a combinação dos idiomas obteve resultados minimamente melhores, porém a diferença não era significativa.

Considerando o esforço, geralmente manual, para obter um *corpus* paralelo alinhado de qualidade, além da dificuldade para adaptar para outros idiomas, e a ausência de opções voltadas à análise de emoções, essa é uma abordagem de difícil adaptação para a tarefa deste trabalho, não sendo explorada.

3.5 AS *Cross-lingual*

A classificação de sentimentos *cross-lingual* usa dados anotados em um idioma (geralmente inglês) para classificação de sentimentos em outro idioma. Wan (2011) mostrou uma abordagem de co-treinamento bilíngue para classificação de sentimentos na qual instâncias anotadas em inglês são traduzidas para o chinês e instâncias não-anotadas em chinês são traduzidas para o inglês, usando todos os atributos em inglês e chinês em um framework único. São com-

[3] Disponível em http://www.statmt.org/europarl/

paradas abordagens usando aprendizado supervisionado (com o modelo SVM) e usando um dicionário de sentimentos, considerando as diferentes direções de tradução (inglês→chinês e chinês→inglês). Os resultados mostraram que o modelo SVM em conjunto com a proposta de co-treinamento superaram os resultados obtidos com as abordagens léxicas.

A proposta de Hajmohammadi, Ibrahim e Selamat (2014) é criar um modelo para análise de sentimentos multilíngue usando dados rotulados de diversas fontes em diferentes idiomas, gerando múltiplas visões (em inglês: *views*) do conjunto de treinamento. Essas visões são usadas em um processo de aprendizado semi-supervisionado para incorporar dados não-rotulados no processo de treinamento. O tipo de sentimento usado é polaridade. As visões foram criadas através da tradução de dados rotulados a partir de dois idiomas fonte para um terceiro idioma alvo e da tradução de dados rotulados e não-rotulados do idioma alvo para ambos os idiomas fonte, gerando assim 3 visões contendo tanto dados rotulados quanto dados não-rotulados nos três idiomas. Foram considerados os idiomas inglês, francês, alemão, japonês e chinês. Os resultados mostraram que o modelo proposto obteve melhor performance que todos os *baselines* em relação à acurácia, obtendo ótimos resultados também para as medidas Precisão, Revocação e Medida-F. O uso dos dados de treinamento e teste, incluindo as traduções para os idiomas fonte e alvo, melhoram a classificação. Outro efeito do modelo proposto foi que diferentes exibições e diferentes idiomas fonte podem se complementar para conter os termos contidos nos dados de teste.

O presente trabalho apresenta um experimento onde textos em um idioma são utilizados para classificar emoções em textos de outro idioma, com o auxílio de tradução automática.

3.6 Considerações finais

Para uma melhor visualização dos trabalhos na área, a Tabela 3.1 apresenta os principais pontos dos trabalhos mais relevantes vistos neste capítulo. Diversos trabalhos combinam o uso de tradução automática com aprendizado supervisionado e obtêm ótimos resultados (BAUTIN; VIJAYARENU; SKIENA, 2008; BANEA et al., 2008; BANEA; MIHALCEA; WIEBE, 2010; BALAHUR; TURCHI, 2014; BALAHUR; PEREA-ORTEGA, 2015). Esta dissertação utiliza essa combinação para a classificação de emoções, o que não foi explorado nos trabalhos analisados. São comparados também a forma de usar as traduções automáticas geradas, como nos trabalhos de (BANEA; MIHALCEA; WIEBE, 2010) e de (NARR; HULFENHAUS; ALBAYRAK, 2012).

Considerando os resultados de Balahur e Turchi (2014), voltado à polaridade, este trabalho

Tabela 3.1: Trabalhos relacionados

Trabalho	**Técnica**	**Idiomas**	**Tipo de sentimento**	**Abordagem**	**Descrição**	**Textos utilizados**
Narr, Hulfenhaus e Albayrak (2012)	ASM sem o uso de recursos linguísticos específicos	Inglês, Francês, Alemão e Português	Polaridade	Aprendizado Supervisionado (Naïve Bayes)	Criar um *corpus* de treinamento rotulando *tweets* a partir de emoticons.	*Tweets*
Lin et al. (2014)	ASM sem o uso de recursos linguísticos específicos	Inglês, Francês e Alemão	Polaridade	*Self-Supervised Learning* : combinação de aprendizado supervisionado (Naïve Bayes) e não supervisionado	Framework para criação de dicionário de sentimento a partir de palavras semente sem uso de *corpora* anotados.	Comentários de diferentes produtos da Amazon
Banea et al. (2008) e Bane, Mihalcea e Wiebe (2010)	Criação de recursos para outros idiomas a partir de recursos já existentes	Inglês, Espanhol, Romeno (2008) mais Arábico, Francês e Alemão (2010)	Polaridade	Aprendizado Supervisionado (Naïve Bayes) e Tradução Automática.	Criação de *corpora* anotados em diferentes idiomas usando tradução automática.	MPQA
Molina-Gonzáles et al. (2013)	Criação de recursos para outros idiomas a partir de recursos já existentes	Inglês e Espanhol	Polaridade	Dicionário de sentimentos e Tradução Automática	Criação do dicionário de sentimentos eSOL a partir da tradução do Bing Liu English Lexicon para espanhol e posterior revisão manual.	Bing Liu English Lexicon
Balahur e Turchi (2014)	Criação de recursos para outros idiomas a partir de recursos já existentes	Inglês, Francês, Alemão e Espanhol	Polaridade	Aprendizado Supervisionado (SVM) e Tradução Automática	Comparação entre traduções automáticas de diferentes tradutores e a combinação deles para cada idioma; comparação de diferentes opções de pré-processamento e algoritmo.	MOAT
Steinberger et al. (2011)	*Corpu* s paralelo	Inglês, Espanhol, Francês, Alemão, Checo, Italiano e Húngaro	Polaridade	Dicionário de sentimentos e Tradução Automática em conjunto com um *corpus* paralelo	Tradução de listas de palavras nos idiomas inglês e espanhol para outros idiomas para criação de dicionários de sentimentos. Testes são feitos com um *gold standard.*	Europarl
Wan (2011)	AS *cross-lingual*	Inglês e Chinês	Polaridade	Dicionário de sentimentos e Aprendizado Supervisionado (SVM)	Co-treinamento bilíngue para classificação de sentimentos usando dados rotulados em inglês e dados não rotulados em chinês, comparadocom uso de dicionários de sentimentos	Comentários de produtos da Amazon
Hajmohammadi, Ibrahim e Selamat (2014)	AS *cross-lingual*	Inglês, Francês, Alemão, Japonês e Chinês	Polaridade	Aprendizado Supervisionado (SVM) e Tradução Automática	Aprendizado supervisionado usando dados de treinamento cross-lingual, criando diferentes exibições do conjunto de treinamento	Comentários de produtos da Amazon
Santos, Becker e Moreira (2014a)	Criação de recursos para outros idiomas a partir de recursos já existentes	Inglês e Português	Emoção	Dicionário de sentimentos (NRC) e Tradução Automática	Comparação da tradução das instâncias de teste ou do dicionário de sentimentos usado.	Comentários de ebooks da Amazon
Santos, Becker e Moreira (2014b)	*Corpu* s paralelo	Inglês e português	Emoção	Dicionário de sentimentos (NRC), Tradução Automática, *Corpus* Paralelo e Aprendizado Supervisionado (Naïve Bayes)	Comparação das combinações tradução automática + dicionário de sentimentos e tradução automática + aprendizado supervisionado; qual idioma (ou a combinação deles) produz melhor resultado	Estrofes de músicas

Fonte: a autora

usa tradução automática para gerar *corpora* em diferentes idiomas com a premissa de que as traduções geradas possuem boa qualidade, evitando o processo de extração e anotação de novos *corpora* em diferentes idiomas. Como diversos trabalhos mostram que a tradução automática é um método maduro para criação de recursos em outros idiomas, este trabalho assume que a qualidade da tradução gerada é aceitável, não fazendo parte do escopo a verificação da qualidade dela.

O trabalho de Balahur e Turchi (2014) apresenta uma função de desempenho para o classificador, que permite facilitar a comparação entre as diferentes opções testadas para algoritmos e pré-processamento dos textos para o classificador, além da estrutura do experimento. Essa função é adaptada para o presente trabalho, expandindo as opções de pré-processamento dos textos e algoritmos de classificação, visto serem poucas as opções testadas por Balahur e Turchi (2014).

O uso de meta algoritmos passou a ser considerado para a análise de emoções, visto a melhoria nos resultados obtidas para polaridade com o meta algoritmo Bagging por Balahur e Turchi (2014).

4 MINERAÇÃO DE EMOÇÕES EM TEXTOS MULTILÍNGUES

Este trabalho apresenta dois experimentos para classificar o sentimento de textos de diferentes idiomas em seis tipos de emoção, considerando a escassez de recursos para análise de emoção em diferentes idiomas. Para isso, é feito uso de tradução automática para gerar *corpora* anotados nos idiomas em que não há disponibilidade deste recurso em conjunto com aprendizado de máquina supervisionado para treinar classificadores. São usados dois *corpora* anotados nos idiomas inglês e português, que são traduzidos para o idioma do outro *corpus* e para os idiomas francês e espanhol. Uma das vantagens desta abordagem é que não depende de análises específicas dos idiomas, podendo ser adaptada para mais idiomas posteriormente.

As abordagens propostas na literatura usam polaridade como tipo de sentimento e classificam textos subjetivos nas classes positiva e negativa. A primeira hipótese testada é a de que tradução automática produz *corpora* satisfatórios para ASM utilizando emoção como tipo de sentimento.

Os experimentos têm foco em verificar a melhor forma de usar os *corpora* anotados gerados com tradução automática em conjunto com aprendizado de máquina para classificar emoções. Os experimentos buscam verificar se os resultados obtidos por Banea, Mihalcea e Wiebe (2010) para polaridade se confirmam também para emoção, isto é, se o uso combinado de idiomas melhora a classificação de sentimentos. A segunda hipótese testada pelos experimentos é a de que a combinação de textos de diferentes idiomas, mesmo que provenientes de tradução automática, produz resultados melhores para ASM do que o uso individual de cada idioma.

O Experimento 1 utiliza os dois *corpora* separadamente, avaliando a adaptação da abordagem supervisionada e de tradução automática; o Experimento 2 faz cruzamento dos *corpora*, considerando os melhores resultados obtidos no Experimento 1, usando textos originais e traduzidos para treinamento ou teste do classificador, avaliando a efetividade da tradução automática. O Experimento 2 busca confirmar a hipótese de que os *corpora* gerados a partir de tradução automática podem ser utilizados para ASM com outros *corpora* em idiomas originais sem grandes perdas.

Considerando as diversas opções disponíveis para pré-processamento de texto e de algoritmos para classificação, é definida a função de desempenho de classificação ecp (*emotion classification performance*), adaptada de Balahur e Turchi (2014). A função ecp se propõe a comparar as diferentes opções de pré-processamento, algoritmo e forma de usar os *corpora* traduzidos, com o objetivo de verificar se diferentes combinações produzem diferenças significativas nos resultados. Aqui se definem as formas de representação dos atributos, quais deles serão usados,

quais pesos cada atributo deve ter e qual algoritmo é o mais indicado para a tarefa proposta (a Seção 4.4 apresenta as opções utilizadas). Dessa forma são avaliadas todas as combinações de pré-processamento do texto e algoritmos, cobrindo todas as possibilidades conforme as opções analisadas. Outra hipótese a ser avaliada é a de que combinações de pré-processamento de texto e algoritmos de classificação que melhoram os resultados da ASM usando polaridade como tipo de sentimento também melhoram os resultados quando o tipo usado é emoção. A função *ecp* pode ser medida em termos do Precisão, Revocação e Medida-F. Exceto quando explicitamente indicado, assumiremos a Medida-F no restante deste trabalho.

A função *ecp* será utilizada como base nas comparações estabelecidas como objetivo deste trabalho:

1. Comparar as formas de usar os *corpora* gerados com tradução automática, entre treinar um classificador para cada idioma (classificadores monolíngues para os idiomas originais e traduzidos) ou combinando todos os idiomas para treinar um único classificador (classificador multilíngue).
2. Comparar os efeitos das diferentes opções de pré-processamento e algoritmos, verificando se há diferença significativa entre elas, considerando tanto os classificadores monolíngues quanto o multilíngue.
3. Comparar o desempenho entre duas direções de tradução: a tradução das instâncias de treinamento do classificador para o idioma das instâncias de teste contra a tradução das instâncias de teste para o idioma do classificador.

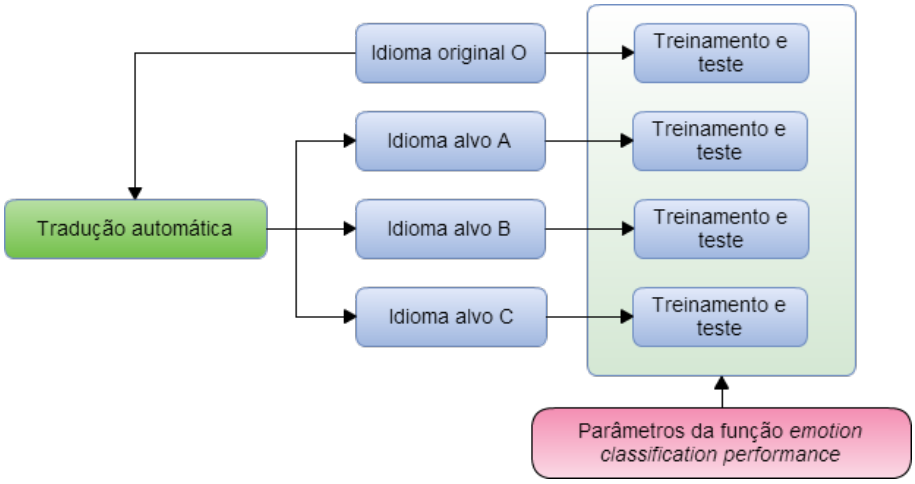
Considerando o processo de AS apresentado na Seção 2.1, este trabalho foca na etapa de classificação do sentimento, sem entrar nas etapas de identificação e de sumarização. O nível de análise utilizado é o nível de documento, sendo cada instância dos *corpora* um documento separado.

As contribuições deste trabalho são 1) mostrar que é possível produzir recursos para AS voltados a emoções em diferentes idiomas, aplicando as mesmas abordagens usadas para polaridade; 2) mostrar diferentes formas de usar *corpora* gerados por tradução automática e 3) como diferentes opções de pré-processamento do texto e algoritmos usados influenciam na classificação dos textos.

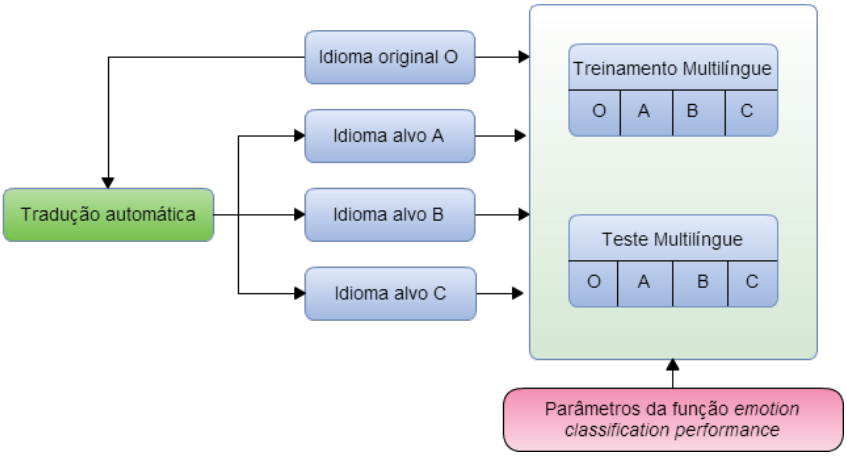
4.1 Experimento 1: análise individual dos *corpora*

Neste experimento, compara-se o uso de classificadores monolíngues, que consistem em treinar um classificador para cada idioma (original e traduções), com o uso de um único classificador multilíngue, treinado usando os textos do idioma original e das traduções geradas simultaneamente, cujas estruturas podem ser vistas na Figura 4.1. Procura-se verificar se o uso de tradução automática traz impacto significativo na qualidade da classificação, comparando os resultados obtidos com os classificadores treinados a partir do texto original e dos textos traduzidos, além de comparar com a combinação dos idiomas.

Figura 4.1: Classificadores monolíngues (a) e Classificador multilíngue (b)



(a) Classificadores monolíngues



(b) Classificador multilíngue

Fonte: a autora

Além de comparar a forma como a tradução é usada, são comparadas as opções de parâ-

metros para a função ecp, abordada na Seção 4.4, para saber como eles afetam os resultados de classificação. Os resultados reportados aqui referem-se ao uso individual de cada *corpus*, ou seja, não há influência dos atributos de um *corpus* no outro.

4.2 Experimento 2: cruzamento dos *corpora*

No Experimento 1, para os idiomas traduzidos, tanto as instâncias de teste quanto as de treinamento provêm de uma tradução, ou seja, possíveis erros introduzidos pela tradução no treinamento são repetidos no teste e se anulam. Como os dois *corpora* utilizados são anotados para as mesmas emoções e suas instâncias são do mesmo domínio (títulos de notícias), este experimento tem por objetivo comparar a tradução das instâncias de treinamento do classificador para o idioma das instâncias de teste com a tradução das instâncias de teste para o idioma do classificador.

Para os classificadores monolíngues são considerados apenas os idiomas originais dos *corpora* disponíveis, inglês e português; o classificador multilíngue contando com todos os quatro idiomas é testado, embora não seja possível analisar os resultados em cada idioma individualmente. A estrutura desse experimento é mostrada na Figura 4.2.

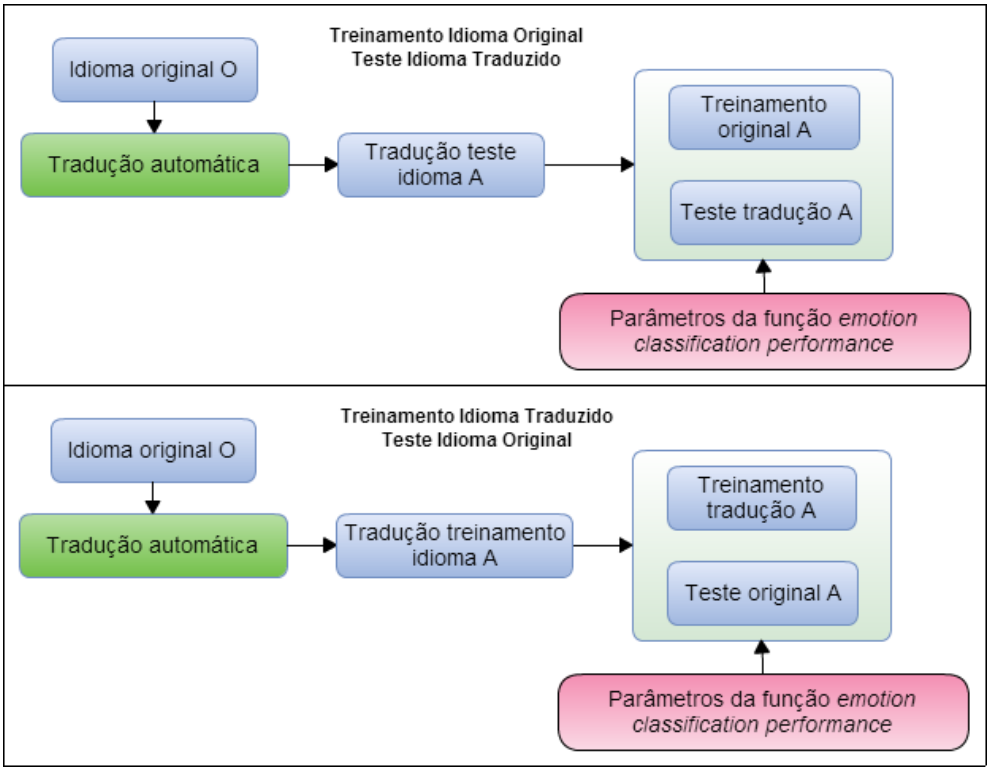
São avaliadas duas direções de uso dos idiomas originais e traduzidos, considerando a ordem treinamento $\rightarrow$ teste: original $\rightarrow$ tradução e tradução $\rightarrow$ original. Na direção original $\rightarrow$ tradução, o treinamento do classificador em inglês é feito com as instâncias no idioma original inglês do *corpus* SemEval e o teste é feito com a tradução para inglês das instâncias em português do Notícias BR; similarmente, o classificador em português treina com as instâncias no idioma original português do *corpus* Notícias BR enquanto o teste é feito com a tradução para português das instâncias em inglês do SemEval.

Já a direção tradução $\rightarrow$ original treina o classificador inglês com a tradução para inglês das instâncias do *corpus* Notícias BR e testa com as instâncias no idioma original inglês do SemEval; o classificador português treina com a tradução para português das instâncias do SemEval e testa com as instâncias originais no idioma português do *corpus* Notícias BR.

Com o classificador multilíngue não é possível analisar separadamente cada idioma, mas é possível avaliar o desempenho geral do experimento ao utilizar um conjunto de dados para teste e o outro para treinamento. Assim, similarmente à estrutura dos classificadores monolíngues, neste classificador há as traduções em espanhol e francês de ambos conjuntos de dados, sendo utilizadas para treinamento e teste.

Neste experimento, são usadas as opções de parâmetros e algoritmos que produziram os

Figura 4.2: Estrutura do experimento de cruzamento de *corpora*



Fonte: a autora

melhores resultados no Experimento 1.

4.3 *Corpora* utilizados

Foram utilizados dois *corpora* para os experimentos, descritos a seguir.

SemEval: adaptado de AffectiveTextSemeval.2007, que foi disponibilizado pelo SemEval 2007 para a tarefa de *Affective Text*[1] (STRAPPARAVA; MIHALCEA, 2007). Ele é composto de 1250 títulos de notícias em inglês extraídos de sites de notícias e/ou jornais de grande circulação (como Google News e CNN). A Figura 4.3 apresenta alguns exemplos de instâncias deste *corpus*.

Cada instância deste *corpus* é anotada para as emoções de Ekman por seis anotadores usando níveis de intensidade, em uma escala de 0 a 100, podendo expressar diferentes emoções simultaneamente. Os experimentos feitos neste trabalho não utilizam a intensidade presente na anotação, visto a dificuldade de definir limites para cada nível de intensidade (e quantidade de níveis,

[1] web.eecs.umich.edu/ mihalcea/affectivetext/

Figura 4.3: Exemplos de instâncias do *corpus* SemEval

```
1  Mortar assault leaves at least 18 dead (anger, disgust, fear, sadness)
2  Goal delight for Sheva (joy, surprise)
3  Nigeria hostage feared dead is freed (anger, fear, joy, sadness, surprise)
4  Bombers kill shoppers (anger, disgust, fear, sadness)
5  Vegetables, not fruit, slow brain decline (fear, joy, sadness, surprise)
```

Fonte: adaptada do *corpus* SemEval (STRAPPARAVA; MIHALCEA, 2007)

conforme a escala presente na anotação). É considerada apenas a presença ou ausência de cada emoção na instância, sendo que a intensidade anotada foi convertida para valores booleanos: acima de zero foi considerada presença e convertido para um, mantendo os valores zero para ausência.

Notícias BR: o Corpus_noticias[2] foi utilizado em (MARTINAZZO; DOSCIATTI; PARAISO, 2011). Este *corpus* é composto de 1002 títulos de notícias e sumários no idioma português, exemplificadas na Figura 4.4.

Figura 4.4: Exemplos de instâncias do *corpus* Notícias BR

```
1  Cientistas criam câmera mais rápida do mundo: Físicos querem aperfeiçoar técnica
   para ajudar a detectar tumores. (joy)
2  Governo do Irã proíbe homenagem a mortos em protestos: Dois ex-candidatos
   presidenciais queriam realizar evento na quinta-feira. (disgust)
3  Ativistas e procuradores temem que nova lei agrave conflitos na Amazônia:
   Regularização fundiária pode fragilizar pequenos agricultores sem acesso ao
   Estado. (fear)
4  Britânica que perdeu custódia de 13 filhos está grávida: Assistentes sociais da
   Grã-Bretanha dizem que Theresa Winters é negligente. (anger)
5  Energia do impacto sobre Felipe Massa foi 30% da letal, estima engenheiro: Perito
   em acidentes automotivos ressalva que adotou premissas. Detalhes que só a perícia
   vão revelar indicarão números exatos. (surprise)
6  Frade francês tem proteção policial 24 horas por dia no Pará: Religioso diz que
   governo brasileiro não tem 'infraestrutura' para impedir conflitos de terra.
   (sadness)
```

Fonte: adaptada do *corpus* Notícias BR (MARTINAZZO; DOSCIATTI; PARAISO, 2011)

Cada instância neste *corpus* foi anotada para apenas uma emoção por três anotadores. Essa anotação foi adaptada para considerar todas as emoções analisadas, resultando que a emoção anotada é considerada presente e as demais ausentes.

Comparando as instâncias dos dois *corpora*, nota-se que as instâncias do Notícias BR são mais longas e contêm mais termos que as instâncias do SemEval, o que pode gerar diferenças

[2] Disponibilizado em http://www.ppgia.pucpr.br/ paraiso/mineracaodeemocoes/recursos/

na classificação, visto que o Notícias BR provê mais atributos para cada instância.

Os *corpora* SemEval, em inglês, e Notícias BR, em português, foram traduzidos para os idiomas alvo, sendo suas anotações projetadas para as respectivas traduções. Para cada emoção, é treinado um classificador que identifica se a emoção está presente ou não na instância testada, caracterizando o problema de classificação como binário e tornando a classificação de uma emoção independente da classificação de outra.

A Tabela 4.1 contabiliza a quantidade de instâncias por emoção nos dois *corpora* usados. Pode-se notar que os *corpora* são opostos no sentido de representação das emoções: enquanto no SemEval a maioria das emoções está presente na maior parte das instâncias (com exceção de *disgust*), no Notícias BR a maior parte das instâncias não apresenta emoções. Outro detalhe a ser observado é o desbalanceamento das classes nos dois *corpora*, em especial a emoção *surprise*, que pode afetar de forma negativa os resultados ao classificar erroneamente instâncias de uma classe pouco representada.

Tabela 4.1: Distribuição das instâncias por emoções

		anger	**disgust**	**fear**	**joy**	**sadness**	**surprise**
SemEval	Presente	652	525	820	786	863	1102
	Ausente	598	725	430	464	387	148
Notícias BR	Presente	32	221	58	364	266	31
	Ausente	970	751	944	638	736	971

Fonte: a autora

As instâncias dos dois *corpora* passaram pelo mesmo pré-processamento básico. As instâncias foram traduzidas pelo Google Tradutor[3] para os idiomas espanhol, francês e português no *corpus* SemEval, e para inglês, espanhol e francês no *corpus* Notícias BR. Após a tradução, os termos das instâncias passaram pelo processo de conversão de letras para minúsculas e remoção de números, pontuações e outros caracteres especiais. Foi feita ainda a remoção de *stopwords*, utilizando uma lista de termos para cada idioma disponível na internet.

4.4 Parâmetros para função de classificação de emoções

Seguindo a proposta de Balahur e Turchi (2014), descrita na seção 3.3, foi adaptada a função *emotion classification performance* $ecp = fn(tk, ws, sa, alg, es)$, que permite a variação dos parâmetros, visando comparar seus efeitos na classificação. A maioria dos parâmetros foi

[3] https://translate.google.com

escolhida devido à melhoria que produzem em resultados de classificação de textos, com exceção do parâmetros es, que traduz uma das comparações feitas neste trabalho. Na função ecp, tk indica o tipo de tokenizador, ws representa o *weighting scheme* utilizado, sa indica se é feita a seleção de atributos, alg indica o algoritmo utilizado para o treinamento do classificador e es define a estrutura do experimento, ou seja, a forma como as traduções são utilizadas.

Entre as opções dos parâmetros estão:

- Tokenizadores: as opções testadas são unigramas, na qual cada token torna-se um atributo, e a combinação unigramas + bigramas (sequência de dois tokens).
- *Weighting scheme*: são testadas as opções booleano e TF-IDF.
- Seleção de atributos: este parâmetro indica se é feita a seleção de atributos do *corpus* através de InfoGain ou se são usados todos os atributos presentes no *corpus* - essa opção é identificada por null.
- Algoritmos: as opções de algoritmos testadas são Naïve Bayes, SVM, RBF e as combinações dos meta algoritmos AdaBoost e Bagging com RBF e SVM, obtendo AdaBoost(RBF), AdaBoost(SVM), Bagging(RBF) e Bagging(SVM).
- Estrutura do experimento: este parâmetro compara as formas de uso das traduções geradas, se separadas dos textos no idioma original, com classificadores separados para cada idioma, chamado **monolíngue**, e a combinação de todos os idiomas considerados em um único classificador, chamada **multilíngue**.

As opções dos parâmetros tk, ws, sa e es foram escolhidas devido a serem frequentemente usadas em tarefas de classificação de textos e pela disponibilidade na ferramenta usada para os testes. A função *ecp* extende a função *scp* ao considerar seleção de atributos como um dos parâmetros, além de considerar outras opções de algoritmos. Outra diferença fundamental entre as duas funções é entre os parâmetros es e ed: enquando na função *ecp* o parâmetro es avalia a combinação de idiomas em um classificador ou o uso individual deles, na função *scp* o parâmetro ed avalia a combinação de diferentes tradutores automáticos, porém, para um mesmo idioma.

4.5 Ambiente de testes

Os experimentos foram executados com o auxílio do pacote de software Weka (HALL et al., 2009). O Weka contém diversos algoritmos para classificação e diferentes opções de pré-processamento. Possui ainda a interface *Experimenter*, que permite a execução de experimentos

com diferentes configurações, através de arquivos arff, além de produzir diferentes métricas para análise.

Para cada emoção e idioma (e combinação de idiomas), foi gerado um conjunto de dados de acordo com a combinação dos parâmetros tk, sa, ws e es. Estes conjuntos de dados foram gerados automaticamente via código Java com o auxílio das funções do Weka (por exemplo, StringToWordVector, AttributeSelection, NumericToBinary, entre outras). Cada um dos conjuntos de dados, composto pela combinação das opções dos parâmetros da função ecp, é chamado nos experimentos de configuração.

Foram mantidos os parâmetros básicos do Weka, com exceção de: a) tokenizadores, nos quais foi removido o limite máximo de tokens a serem gerados, gerando todas as opções para unigramas e bigramas e b) *treshold* para o InfoGain, onde foi utilizado 0,001 para efetivo corte de atributos. A forma de validação do experimento é a *10-fold cross-validation*, visando aproveitar todas as instâncias ao máximo.

No Experimento 2, foi utilizada a função do Weka InputMapClassifier, que faz a ligação entre dados de treinamento e de teste incompatíveis através da construção de um mapeamento entre os dados de treinamento do classificador e a estrutura das instâncias de teste, de forma que possam ser usados na mesma execução.

Referente à quantidade de resultados obtidos, no Experimento 1, são considerados:

- 6 emoções: *anger*, *disgust*, *fear*, *joy*, *sadness* e *surprise*;
- 5 opções de idiomas: inglês, português, espanhol, francês e a combinação de todos eles (multilíngue);
- 2 opções de tokenizadores: unigramas e unigramas + bigramas;
- 2 opções para *weighting scheme*: booleano e TF-IDF;
- 2 opções para seleção de atributos: uso de InfoGain ou sem seleção de atributos (null);
- 7 algoritmos: AdaBoost(RBF), AdaBoost(SVM), Bagging(RBF), Bagging(SVM), Naïve Bayes, RBF e SVM SMO.

Todas essas opções geram um total de 1680 resultados para cada *corpus*. Considerando o grande volume de informações, para o Experimento 2, foram removidas algumas opções, como os idiomas espanhol e francês, uma das opções de seleção de atributos, e mantendo os três algoritmos que atingiram as melhores médias de Medida-F, conforme os resultados obtidos no Experimento 1, gerando um total de 216 resultados.

Os resultados são reportados usando médias ponderadas de Precisão, Revocação e Medida-F. A comparação entre os resultados é feita usando teste-t bi-caudal, com confiança de 0,95.

4.6 Resultados

A seguir são apresentados os resultados e análises dos experimentos descritos anteriormente. Os resultados são apresentados conforme os objetivos apresentados no início do capítulo. O Anexo apresenta as tabelas completas com os resultados que foram utilizados para criação dos gráficos.

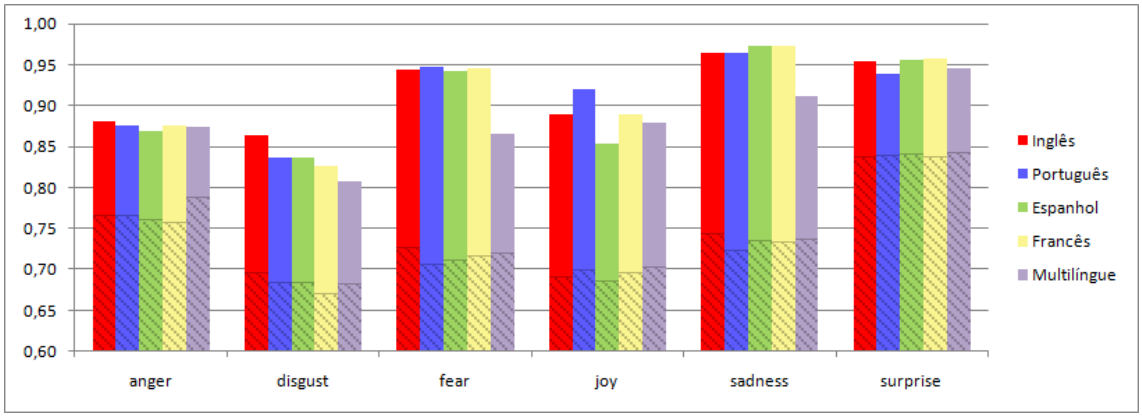
4.6.1 Experimento 1: análise por emoções

A Figura 4.5 mostra os resultados relativos à Medida-F para cada emoção para os *corpora* SemEval e Notícias BR, considerando a média de todas as configurações testadas. Para cada emoção, a base hachurada da barra indica a média de Medida-F para cada idioma referente a cada emoção, enquanto o topo da barra mostra o valor máximo atingido de Medida-F para cada idioma e emoção em uma dada configuração.

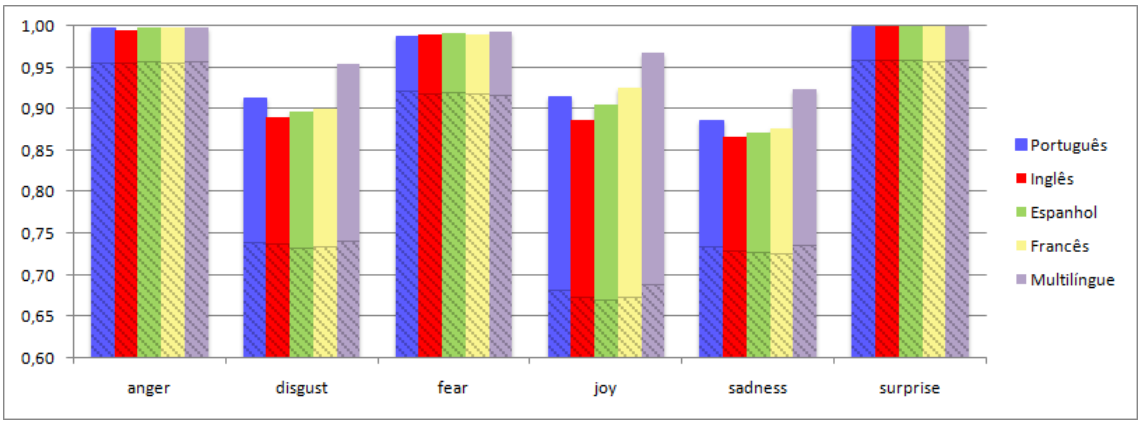
Pode-se notar que, independente do idioma, sendo original ou tradução, cada emoção apresenta comportamento similar dentro de cada *corpus*, obtendo médias bem próximas em cada idioma. Este comportamento foi observado não só para Medida-F, mas também para Precisão e Revocação. No SemEval as diferenças entre as emoções são menores (a média mais baixa foi obtida por *disgust*, com Medida-F em 0,67 para francês, e as mais alta por *surprise*, com Medida-F de 0,84 para espanhol). No Notícias BR as diferenças entre os classificadores das diferentes emoções são maiores: a média mais baixa de Medida-F foi atingida por *joy* para espanhol, com 0,67, e a mais alta foi atingida por *anger* e *surprise*, com 0,96 em todos idiomas. O *corpus* Notícias BR obtém médias mais altas de Medida-F do que o *corpus* SemEval, podendo ser justificada essa diferença ao analisar as instâncias dos dois *corpora*: as instâncias do SemEval apresentam o título da notícia apenas, sendo mais curtas que as instâncias do Notícias BR, que apresenta o título da notícia e o sumário, provendo mais informações para os classificadores.

As emoções melhores classificadas são *surprise* e *anger*, em ambos os *corpora*, enquanto *disgust* atinge os resultados mais baixos no SemEval e *joy* no Notícias BR. Ainda assim, nota-se que mesmo as emoções que atingiram as médias mais baixas atingiram um bom resultado (Medida-F acima de 0,80) em pelo menos uma configuração, como pode ser notado pelos topos das barras indicando os valores máximo. Logo, este experimento mostra que é possível obter bons resultados para classificação de emoção em diferentes idiomas usando aprendizado de máquina e tradução automática. Como os resultados para cada emoção diferem pouco entre

Figura 4.5: Resultados por emoções para os *corpora* SemEval (a) e Notícias BR (b) - Medida-F



(a) SemEval



(b) Notícias BR

Fonte: a autora

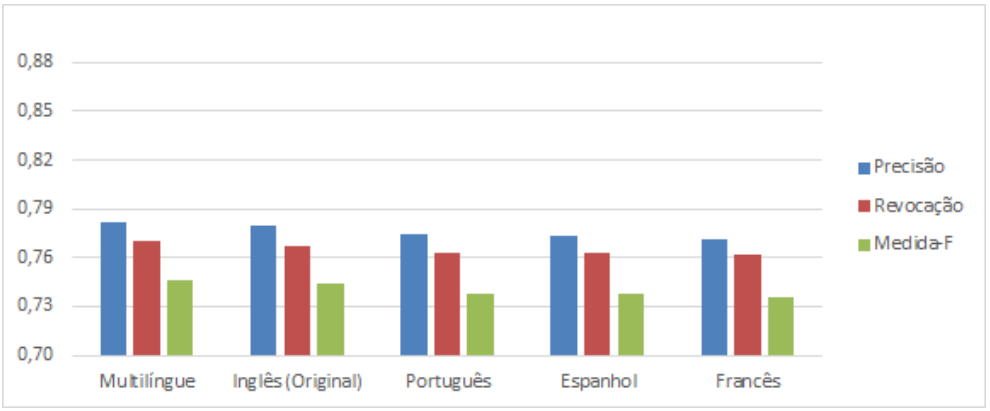
os idiomas, as próximas análises não irão ser discriminadas por emoção, usando a média entre elas.

4.6.2 Experimento 1: análise por estrutura do experimento

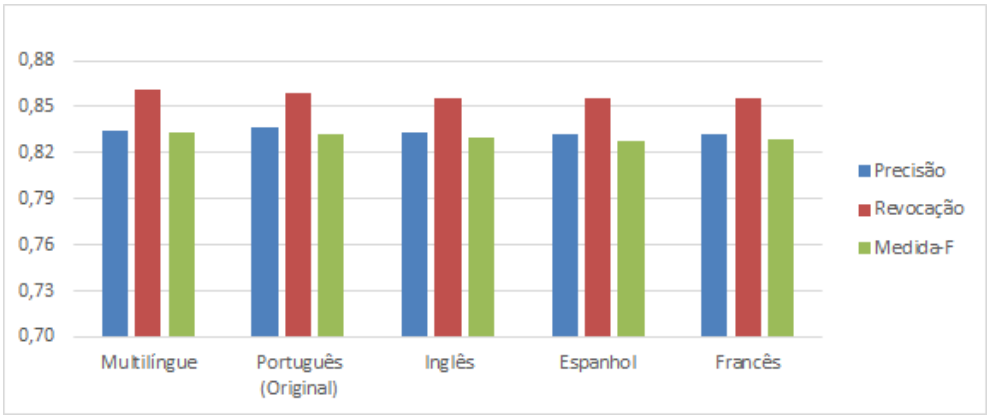
Nesse experimento, procurou-se verificar se o resultado obtido por Banea, Mihalcea e Wiebe (2010), referente ao uso de combinação de idiomas melhorar os resultados de classificação, se mantém para o uso de emoção. A Figura 4.6 apresenta as médias dos resultados de classificação considerando apenas os idiomas. Os gráficos mostram que os resultados para cada idioma são muito similares em cada *corpus*, o que nos leva a inferir que as características do *corpus* no idioma original são levadas também para as traduções.

Na comparação entre os idiomas, observa-se que, na média, as classificações baseadas em

Figura 4.6: Resultados das classificações por idiomas para os *corpora* SemEval (a) e Notícias BR (b)



(a) SemEval



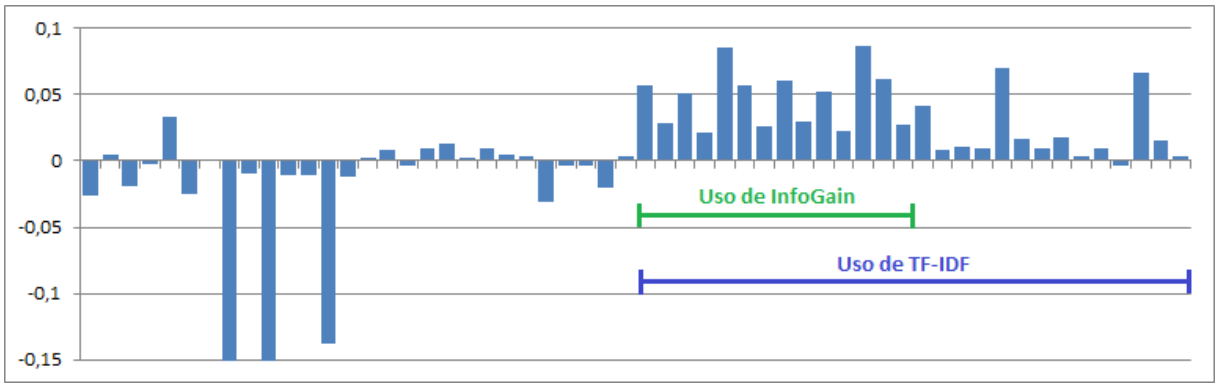
(b) Notícias BR

Fonte: a autora

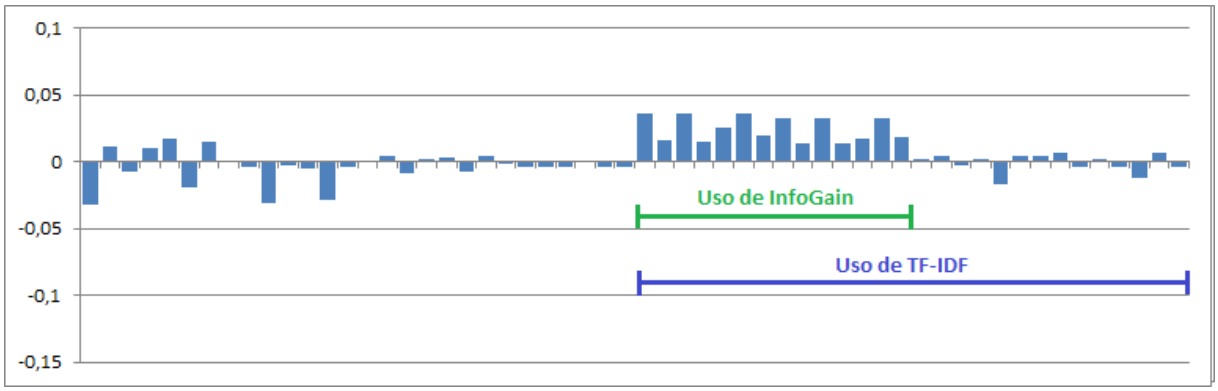
traduções chegam a resultados muito próximos dos idiomas originais: os resultados de Medida-F no *corpus* SemEval atingem média 0,746 para o classificador multilíngue, 0,744 para o inglês, 0,737 para o português, 0,737 para o espanhol e 0,736 para o francês; já para o *corpus* Notícias BR as médias de Medida-F atingem os valores 0,834 para o classificador multilíngue, 0,832 para o português, 0,830 para o inglês, 0,828 para o espanhol e 0,828 para o francês. Esses resultados mostram que há uma perda mínima, porém significativa estatisticamente, segundo o teste-t, nos resultados quando usada a tradução automática (em alguns casos a média mais alta é de um idioma traduzido, como português para *joy* no SemEval, que atinge Medida-F de 0,700, enquanto o idioma original inglês atinge 0,697 de Medida-F, sendo essas diferenças significativas), mas que confirmam a hipótese da viabilidade dos *corpora* provenientes de tradução automática.

Para uma melhor comparação entre as estruturas do experimento, ou seja, como são usados os dados obtidos através de tradução automática, pode-se observar os histogramas da Figura 4.7, que apresentam as diferenças de Medida-F entre o classificador multilíngue e os classificadores monolíngues. Para cada configuração (tk, ws, sa, alg) considerada, calculou-se a diferença entre o resultado de Medida-F obtido com a combinação das traduções com o idioma original (classificador multilíngue), e o respectivo resultado pela média obtida dos classificadores monolíngues (tanto o idioma original quanto as traduções).

Figura 4.7: Diferença entre classificador Multilíngue e classificadores Monolíngues para os *corpora* SemEval (a) e Notícias BR (b)



(a) SemEval



(b) Notícias BR

Fonte: a autora

As Tabelas A.13 e A.14 no Anexo apresentam as configurações de cada barra na ordem de apresentação das figuras. Pode-se observar que no *corpus* SemEval a maioria das barras é positiva (33 barras de um total de 56), indicando que os resultados multilíngues chegam a uma média mais alta, obtendo uma diferença positiva expressiva, embora três casos possuam uma diferença negativa expressiva nas médias de 0,15 e 0,13 em Medida-F, para as configu-

rações usando unigramas + bigramas, booleano, InfoGain e os algoritmos AdaBoost(RBF), Bagging(RBF) e RBF, indicando que os classificadores monolíngues chegaram em resultados bem melhores. Embora no *corpus* Notícias BR as diferenças não sejam tão expressivas (a diferença média é de 0,004 de Medida-F, enquanto a média de diferenças no SemEval é de 0,007), ainda assim nota-se comportamento similar: na maior parte das configurações o classificador multilíngue atinge melhores médias (39 barras), embora em alguns casos isolados os classificadores monolíngues chegam a melhores médias. Nos dois *corpora*, as maiores diferenças significativas positivas ocorrem quando é usado *weighting scheme* TF-IDF e InfoGain para seleção de atributos, indicando que nesses casos o classificador multilíngue atingiu resultados consideravelmente melhores que a média dos classificadores monolíngues.

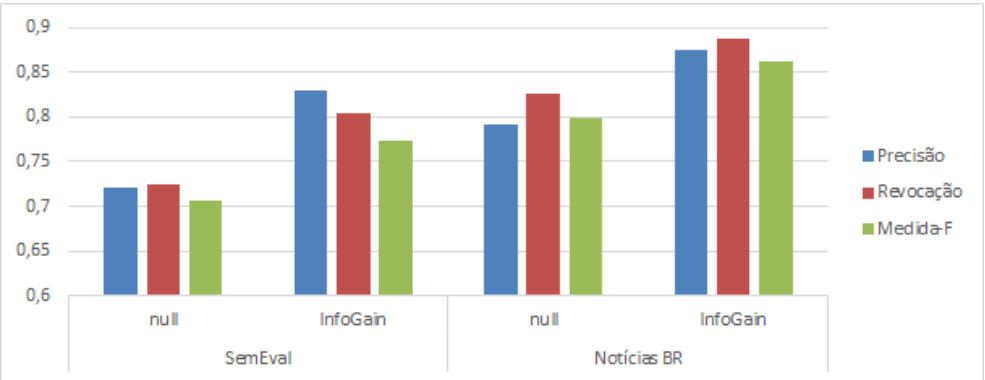
Em geral, os resultados mostram que o classificador multilíngue é superior aos classificadores monolíngues, apesar de a diferença não ser significativa estatisticamente, o que mostra que mais idiomas também melhora os resultados de classificação para emoção, confirmando a hipótese relacionada à estrutura do experimento.

4.6.3 Experimento 1: análise dos parâmetros da função *ecp*

Seleção de atributos

Considerando a função *ecp*, comparamos os resultados obtidos com os diferentes parâmetros testados, considerando as médias de todas emoções e idiomas. Dentre as opções testadas, a seleção de atributos foi a que trouxe a maior diferença, comparando o uso de InfoGain para seleção de atributos e o uso de todos os atributos (identificado como null), como pode ser visto na Figura 4.8.

Figura 4.8: Resultados por seleção de atributos



Fonte: a autora

Como pode ser observado na Figura 4.8, o uso de InfoGain obtém resultados melhores para Precisão, Revocação e Medida-F nos dois *corpora*, sendo essa melhora estatisticamente significativa (a diferença do uso de InfoGain e o uso de todos os atributos nos dois *corpora* foi de 0,07 na Medida-F). Isto indica que foram selecionados os atributos mais significativos para cada classe. O uso de InfoGain reduziu consideravelmente a quantidade de atributos (o Anexo apresenta uma tabela com uma média da quantidade de atributos selecionados para cada *corpus*). Em ambos os *corpora*, o *weighting scheme* TF-IDF provê uma redução mais substancial de atributos do que quando o peso usado é o booleano, com maior efetividade para o *corpus* Notícias BR. Quando é utilizado unigramas como tokenizador, a quantidade de atributos selecionados também é menor do que quando são usados unigramas + bigramas. Tanto para a estrutura monolíngues quanto para a multilíngue, a percentagem de atributos selecionados é similar, variando em torno de 2% no máximo por emoção. Como o *corpus* SemEval possui menos atributos do que o Notícias BR, ele manteve uma percentagem maior de atributos em relação à quantidade original, em especial quando a opção de tokenizador utilizada é a combinação unigramas + bigramas: apenas para essa opção, a opção de *weighting scheme* booleano obteve uma redução de atributos maior.

Weighting Scheme

No parâmetro *weighting scheme*, podemos ver na Figura 4.9 que as diferenças entre as opções booleano e TF-IDF são mínimas, mas ainda assim estatisticamente significativas: para o SemEval, a opção booleano obteve resultado melhores, enquanto no Notícias BR a melhor opção foi TF-IDF. A diferença está nas instâncias dos *corpora*: o Notícias BR possui instâncias mais longas do que o SemEval (quase o dobro do tamanho). Logo, o TF-IDF se torna uma opção mais adequada para o *corpus* Notícias BR, visto que o peso TF-IDF é mais indicado para textos mais longos, obtendo vantagem sobre o peso booleano nesse corpus. No SemEval, como as instâncias são mais curtas, o peso booleano não sofre a desvantagem do outro *corpus*, e as duas opções de *weighting scheme* atingem quase os mesmos resultados, embora o peso booleano seja um pouco melhor: a média de Medida-F do booleano é 0,744 e a do TF-IDF é 0,735.

Tokenizadores

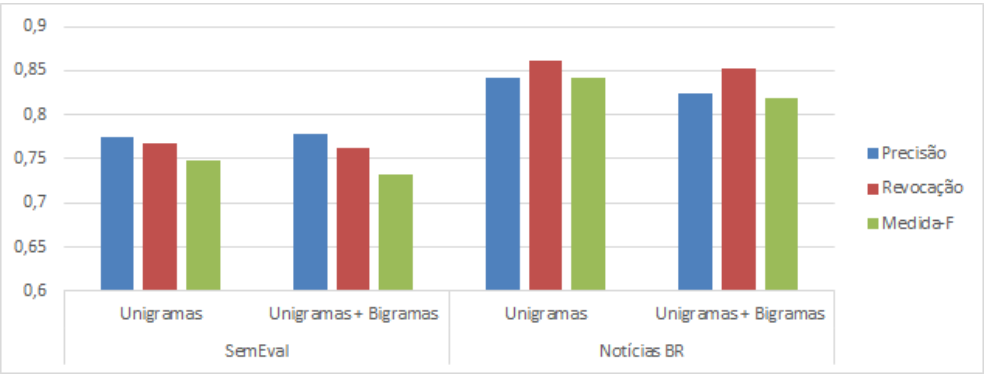
As opções testadas para tokenizadores são unigramas e unigramas + bigramas, como mostrado na Figura 4.10. A figura mostra que as diferenças entre as opções de tokenizadores são pequenas, porém estatisticamente significativas: no SemEval, o uso de unigramas atingiu Medida-F de 0,75 e unigramas + bigramas atingiu 0,73, enquanto no Notícias BR unigramas

Figura 4.9: Resultados por *weighting schemes*

Fonte: a autora

atingiu Medida-F de 0,84 e a combinação unigramas + bigramas atingiu 0,82. Assim, o uso de apenas unigramas se mostrou mais efetivo que o uso de combinado de unigramas e bigramas, o que indica que a maior parte dos bigramas não deve ser discriminativa nos *corpora*.

Figura 4.10: Resultados por tokenizadores



Fonte: a autora

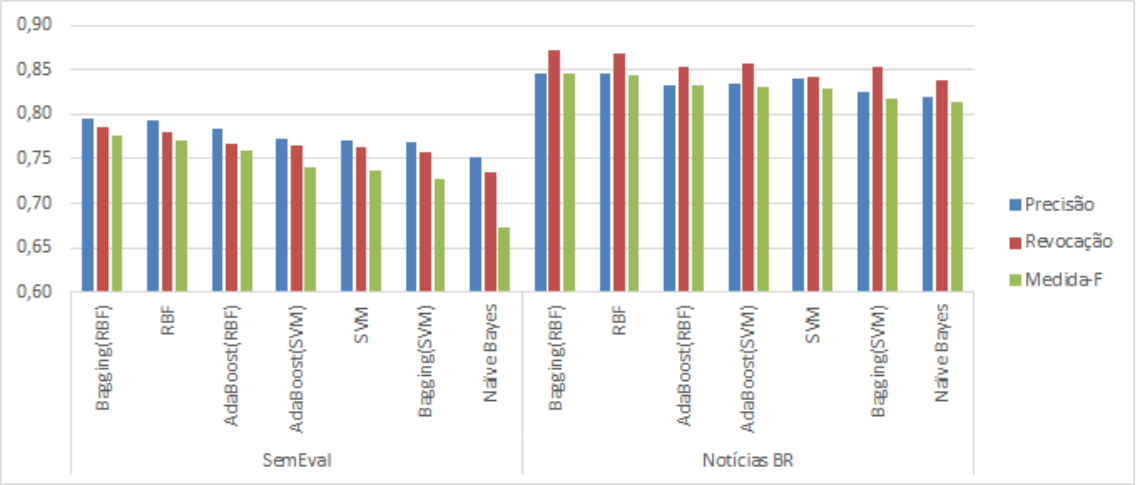
Ao analisar as opções de tokenizador junto ao efeito de *weighting scheme*, o uso de unigramas + bigramas obteve resultados superiores ao uso de apenas unigramas quando o *weighting scheme* utilizado era TF-IDF, indicando que alguns bigramas devem ser mais discriminativos e foram favorecidos pelo peso TF-IDF.

Algoritmos

Quanto ao algoritmo utilizado, conforme a Figura 4.11, os melhores resultados foram atingidos quando foi utilizado o algoritmo RBF, sendo que o uso em conjunto com Bagging trouxe

um ganho de 0,002 na média de Medida-F, sendo significativo segundo teste-t, enquanto o uso de AdaBoost em conjunto com RBF trouxe uma perda de 0,01 na média de Medida-F em comparação com RBF. O algoritmo SVM SMO e suas combinações com meta algoritmos obteve resultados apenas um pouco abaixo dos resultados de RBF, enquanto o Naïve Bayes obteve os resultados mais baixos, atingindo Medida-F média de 0,67.

Figura 4.11: Resultados por algoritmos



Fonte: a autora

Considerando outros parâmetros da função *ecp*, quando o algoritmo usado era o SVM SMO, incluindo o uso dos meta algoritmos com SVM SMO como base, a opção que obteve resultados superiores foi TF-IDF, enquanto que no uso de Naïve Bayes e RBF, incluindo os meta algoritmos com RBF como base, a melhor opção de *weighting scheme* foi booleano. Foi observado que o meta algoritmo AdaBoost melhorou os resultados do algoritmo SVM SMO, enquanto piorou os resultados do RBF; em contrapartida, o meta algoritmo Bagging melhorou os resultados do RBF, mas piorou os resultados do SVM SMO.

Considerando os parâmetros avaliados nesse experimento, podemos confirmar a hipótese de que as opções de pré-processamento e algoritmo que geralmente trazem ganho nos resultados de classificação de polaridade também trazem ganhos para classificação de emoção.

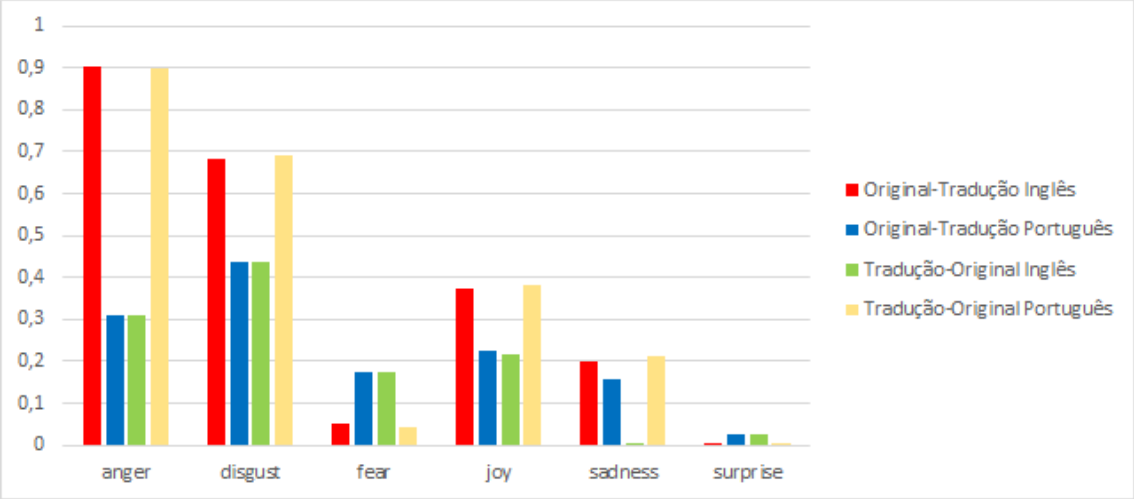
4.6.4 Experimento 2: tradução das instâncias de treinamento do classificador *vs.* tradução das instâncias de teste

O Experimento 2 faz uso de cruzamento de *corpora*, usando um para treinamento e o outro para teste, buscando avaliar se uma das traduções, entre treinamento e teste, produz resultados

melhores, através do uso de textos no idioma original. Este experimento utiliza RBF e suas combinações com meta algoritmos como opções de algoritmo e usa InfoGain para seleção de atributos, visto que essas opções trouxeram melhorias mais significativas no experimento anterior. Os parâmetros tk e ws, que obtiveram as menores diferenças entre as opções testadas, permanecem com as mesma opções do Experimento 1.

Emoção e Idioma

Figura 4.12: Resultados por emoção - Cruzamento de *corpora* - Inglês e Português - Medida-F



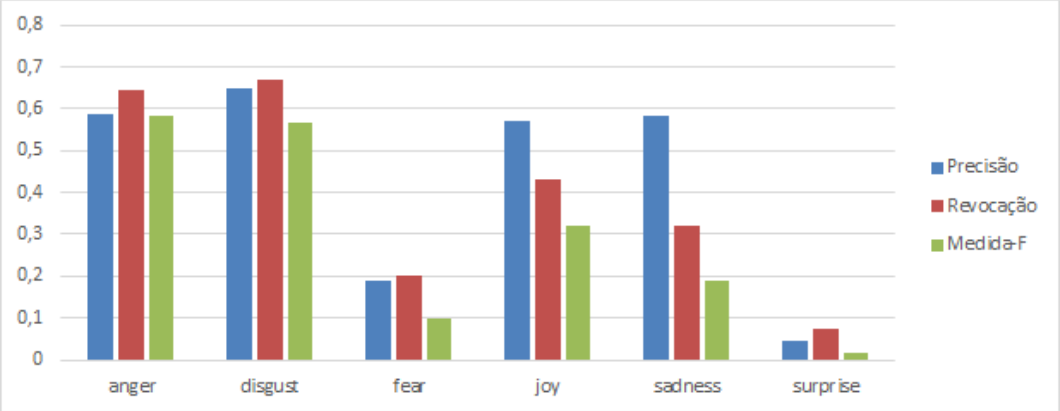
Fonte: a autora

A Figura 4.12 mostra os resultados da média de Medida-F para os idiomas português e inglês, identificando quando o idioma é original ou tradução no sentido Treinamento→Teste. Como pode-se ver, a emoção melhor classificada é *anger*, seguida de *disgust* e *joy*. A emoção que teve os piores resultados foi *surprise*, que não atinge 0,03 de Medida-F. Pode ser observado que os resultados dependem mais do *corpus* utilizado do que do idioma ou da direção testada: quando o treinamento é feito com o SemEval são obtidos os melhores resultados para *anger*, *disgust*, *joy* e *sadness*; quando o treinamento é feito com o Notícias BR, *fear* e *surprise* atingem resultados um pouco melhores. Nota-se que *sadness* é a única emoção que não segue completamente o comportamento das outras relativo ao conjunto de treinamento: quando o treinamento é feito com o Notícias BR, os resultados obtidos possuem uma diferença de 0,15 de Medida-F, enquanto para as outras emoções a diferença entre os resultados dos casos onde é usado o mesmo conjunto de treinamento é de no máximo 0,01 de Medida-F.

A Figura 4.13 mostra os resultados por emoção para a estrutura multilíngue do experimento.

Apesar de não ser possível identificar os idiomas nessa estrutura, ele atinge resultados muito similares à estrutura monolíngue, mantendo o comportamento em relação às emoções: *surprise* ainda é a pior emoção classificada, enquanto *anger*, apesar de uma piora em relação aos melhores resultados na estrutura monolíngue, ainda é a emoção com melhor Medida-F, seguida de perto por *disgust*.

Figura 4.13: Resultados por emoção - Cruzamento de *corpora* - Multilíngue



Fonte: a autora

Assim, considerando a estrutura do experimento, o classificador monolíngue atinge melhores resultados que os classificadores multilíngues em pelo menos uma das duas direções testadas (original → tradução para *anger*, *disgust*, *joy* e *sadness* e tradução → original para *fear* e *surprise*), embora essa melhora não seja estatisticamente significativa.

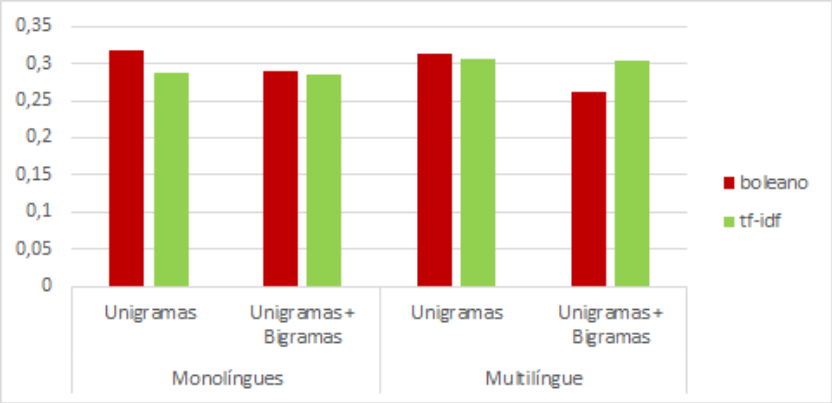
O desbalanceamento oposto das classes nos dois *corpora* pode ter afetado significativamente este experimento, em especial a emoção *surprise*, que é majoritariamente presente em um *corpus* e ausente no outro.

Parâmetros da função *ecp*

Analisando os parâmetros da função *ecp* que foram variados nesse experimento, pode-se observar na Figura 4.14 que o uso de unigramas supera significativamente o uso em conjunto com bigramas para a maioria das configurações, com exceção da estrutura multilíngue, onde o uso de unigramas e bigramas em conjunto com peso TF-IDF atinge melhores resultados.

A Figura 4.15 apresenta os resultados para os classificadores monolíngues e o classificador multilíngue. Tanto no cenário monolíngue quanto no cenário multilíngue o algoritmo que obteve melhores resultados foi o AdaBoost com RBF como algoritmo base, sendo superior significativamente às outras duas opções, chegando a melhores resultados nos classificadores

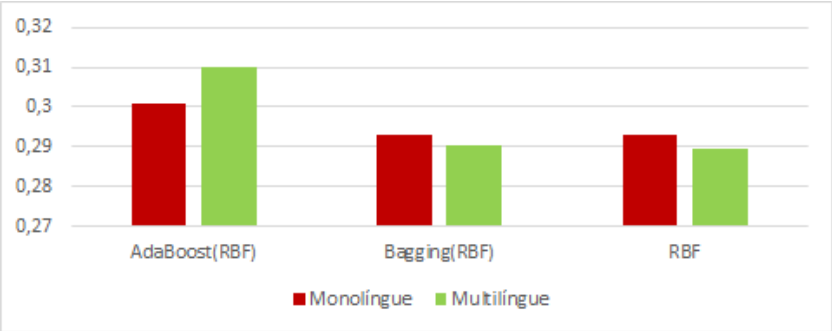
Figura 4.14: Resultados por parâmetros - Cruzamento de *corpora* - Medida-F



Fonte: a autora

multilíngues. Nos outros algoritmos, os classificadores monolíngues obtiveram melhores resultados, embora menos expressivos.

Figura 4.15: Resultados por algoritmos - Cruzamento de *corpora* - Medida-F



Fonte: a autora

Em geral, o uso de unigramas permanece sendo a melhor escolha para tokenizador, porém, o melhor algoritmo difere do resultado obtido no Experimento 1.

Os resultados obtidos no Experimento 2 são inferiores aos obtidos no Experimento 1. Dentre os fatores que podem justificar esses resultados se encontram:

- Domínio: mesmo que ambos os *corpora* estejam no mesmo domínio global, que é títulos de notícias, diferenças de regiões podem afetar o conteúdo de tal forma que os atributos de cada classe sejam distintos entre os *corpora*, como por exemplo notícias regionais; os assuntos das notícias também podem influenciar essa falta de alinhamento entre os *corpora*.

- Desbalanceamento das classes: dentro de cada *corpus* as classes são desbalanceadas, fazendo com que algumas instâncias sejam classificadas na classe errada. Como os *corpora* SemEval e Notícias BR são inversamente desbalanceados, quando foram usados juntos no mesmo experimento o efeito foi potencializado, como pode-se notar especialmente com a emoção *surprise*, e com as emoções *sadness* e *fear*.
- Anotação: a atribuição de rótulos originalmente foi feita de forma diferente nos *corpora* - o SemEval rotula cada instância de acordo com uma intensidade para cada emoção representada por um número inteiro, o que faz com que qualquer valor maior que 0 indique que o termo representa a emoção; já o Notícias BR apresenta apenas uma emoção por instância, que foi escolhida por ter a melhor representatividade, o que faz com que emoções não tão representativas, mesmo que presentes em algum grau pelo processo original de anotação, fossem descartadas, ou seja, os termos da instância não representam essas emoções e o peso delas passa a ser 0; ainda, cada *corpora* foi anotado por um grupo de pessoas diferentes, e o conceito de subjetividade varia para cada pessoa.
- Tradução: embora tradução automática já seja reconhecida como um método maduro para geração de recursos em outros idioma, ela ainda não é um método perfeito, podendo gerar erros que prejudicam a análise de sentimento. No Experimento 1 os erros de tradução se anulam, por se tratar do mesmo *corpus*: os erros presentes no treinamento se encontram também no teste.

Assim, pode-se afirmar que as instâncias de teste precisam estar fortemente relacionadas às de treinamento, de forma que as classes de emoções estejam bem representadas em ambos conjuntos pelos mesmos atributos para que seja possível classificar emoções em um classificador traduzido ou traduzir as instâncias de teste. Logo, não é possível confirmar ou rejeitar a hipótese de que os *corpora* gerados por tradução automática podem ser usados em conjunto com *corpora* em idiomas originais sem perda significativa.

4.7 Considerações finais

Os experimentos realizados mostram que uma abordagem envolvendo aprendizado supervisionado e tradução automática para a criação de *corpora* em diferentes idiomas obtém ótimos resultados para classificar emoção em textos de diferentes idiomas. Os resultados obtidos com as traduções ficam muito próximos dos resultados nos idiomas originais e mesmo que a perda nos resultados das traduções seja significativa, considerando a escassez de recursos nos idiomas traduzidos, ela é aceitável. Esses resultados abrem caminho para que sejam feitos testes com

mais idiomas, para verificar se/quando a inclusão de idiomas afeta negativamente os resultados de classificação. O classificador que reúne todos os idiomas obteve resultados na média superiores aos idiomas originais e traduzidos, indicando que mais idiomas auxiliam a tarefa de classificação de emoção.

A variedade de parâmetros utilizada nos experimentos produziu bons resultados, como combinações específicas de algoritmos e *weighting schemes* (RBF e Naïve Bayes com booleano, SVM SMO com TF-IDF) que não seriam possíveis de identificar caso não houvesse tantas opções de variações.

A escolha de parâmetros que trouxe os melhores resultados em geral, isto é, maior Medida-F considerando as médias de todos os idiomas e emoções, foi unigramas como tokenizador, uso de pesos booleanos, seleção de atributos com InfoGain e uso do meta algoritmo Bagging com RBF como algoritmo base.

Quanto ao Experimento 2, ele mostra que os *corpora* não são muito parecidos entre si, obtendo resultados inconclusivos. Não é possível descartar o uso de classificadores treinados a partir de textos provenientes de tradução automática ainda, visto que para alguns casos estes obtiveram melhores resultados do que classificadores treinados com textos no idioma original.

Considerando os dois experimentos, a emoção melhor classificada é *anger*, indicando que os atributos que identificam esta emoção são os menos ambíguos, independente do idioma.

5 CONCLUSÃO

Este trabalho teve como objetivo comparar diferentes formas de usar *corpora* anotados obtidos de tradução automática para mineração de emoções em textos multilíngues usando aprendizado supervisionado. Foram comparados classificadores monolíngues para cada idioma e um classificador multilíngue combinando todos os idiomas, para verificar se mais idiomas melhoravam a classificação de emoções e se os idiomas traduzidos apresentavam bons resultados. A função ecp foi definida para facilitar a comparação entre as diversas configurações testadas, com o objetivo de verificar os efeitos de diferentes opções de pré-processamento, algoritmos e estrutura dos experimentos. Por último, foi feito um experimento usando os dois *corpora* juntos, utilizando um para teste e o outro para treinamento, utilizando um no idioma original e traduzindo o outro para o idioma em questão.

Os resultados mostram que o na média o classificador multilíngue obteve resultados melhores nos dois *corpora*, indicando que mais idiomas melhoram a classificação de emoções. Os parâmetros da função ecp que mais contribuem positivamente para os resultados são o uso de unigramas como tokenizador, pesos booleanos para o *weighting scheme*, seleção de atributos com InfoGain e uso do meta algoritmo Bagging em conjunto com o algoritmo RBF. O experimento cruzando os *corpora* não obteve resultados conclusivos na maioria das análises feitas, possivelmente devido às diferenças nos *corpora*. Assim, foram confirmadas três das quatro hipóteses trabalhadas nesta dissertação.

As análises presentes neste trabalho estão limitadas à disponibilidade de *corpora* anotados em diferentes idiomas para as emoções de Ekman. Outras opções de pré-processamento e mais opções de algoritmos poderiam ser utilizadas, mas o tempo de processamento se tornaria extremamente elevado, não sendo hábil para este trabalho.

Considerando os resultados obtidos por Balahur e Turchi (2014), indicando que não há correlação entre qualidade da tradução automática e os resultados de classificação de sentimentos, não foi avaliada a qualidade da tradução automática gerada pelo Google Tradutor nesse experimento. Uma análise posterior utilizando tradução manual pode ser feita para avaliar se houve impacto na classificação das emoções por causa de erros de tradução.

Durante o curso de mestrado, foram estudadas diferentes abordagens para classificar emoções em textos multilíngues, resultando em um artigo para o *III Brazilian Workshop on Social Network Analysis and Mining* (BraSNAM 2014) (SANTOS; BECKER; MOREIRA, 2014a) e em um artigo para o Simpósio Brasileiro de Banco de Dados (SBBD 2014) (SANTOS; BECKER; MOREIRA, 2014b). Os resultados obtidos nesses artigos abriram caminho para a pro-

posta abordada nesta dissertação.

Dentre as possíveis expansões para esse trabalho se encontram o estudo da viabilidade da abordagem para idiomas de diferentes etimologias, como alemão e chinês. Também pode ser estudada a possibilidade de criar um classificador sem a necessidade de identificação do idioma a ser testado, ou seja, um classificador multilíngue treinado com diversos idiomas e que aceite qualquer um deles para teste, sendo uma evolução da estrutura multilíngue do experimento apresentado. Ainda, seria interessante a criação de *corpora* anotados que sigam o mesmo processo de criação e anotação do SemEval e do Notícias BR, para refazer o Experimento 2 e obter resultados mais conclusivos.

REFERÊNCIAS

ALM, C. O.; ROTH, D.; SPROAT, R. Emotions from text: machine learning for text-based emotion prediction. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. [S.l.], 2005. p. 579–586.

ALMEIDA, M. S. et al. Aligning opinions: Cross-lingual opinion mining with dependencies. In: *Proc. of the Annual Meeting of the Association for Computational Linguistics*. [S.l.: s.n.], 2015.

ALVES, A. L. et al. Temporal analysis of sentiment in tweets: A case study with fifa confederations cup in brazil. In: DECKER, H. et al. (Ed.). *Database and Expert Systems Applications*. [S.l.]: Springer International Publishing, 2014, (Lecture Notes in Computer Science, v. 8644). p. 81–88.

ASUR, S.; HUBERMAN, B. A. Predicting the future with social media. In: *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*. [S.l.: s.n.], 2010. v. 1, p. 492–499.

BADER, B. W. et al. Multilingual sentiment analysis using latent semantic indexing and machine learning. In: IEEE. *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*. [S.l.], 2011. p. 45–52.

BALAHUR, A. et al. Detecting implicit expressions of affect in text using emotinet and its extensions. *Data & Knowledge Engineering*, v. 88, p. 113–125, 2013.

BALAHUR, A.; PEREA-ORTEGA, J. M. Sentiment analysis system adaptation for multilingual processing: The case of tweets. *Information Processing & Management*, Elsevier, v. 51, n. 4, p. 547–556, 2015.

BALAHUR, A.; TURCHI, M. Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. *Computer Speech & Language*, v. 28, n. 1, p. 56–75, 2014.

BALAHUR, A. et al. Resource creation and evaluation for multilingual sentiment analysis in social media texts. In: *Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC)*. [S.l.: s.n.], 2014.

BANEA, C.; MIHALCEA, R.; WIEBE, J. Multilingual subjectivity: are more languages better? In: *Proceedings of the 23rd international conference on computational linguistics*. [S.l.: s.n.], 2010. p. 28–36.

BANEA, C. et al. Multilingual subjectivity analysis using machine translation. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. [S.l.: s.n.], 2008. p. 127–135.

BAUTIN, M.; VIJAYARENU, L.; SKIENA, S. International sentiment analysis for news and blogs. In: *ICWSM*. [S.l.: s.n.], 2008.

BECKER, K.; TUMITAN, D. Introdução à mineração de opiniões: Conceitos, aplicações e desafios. *Simpósio Brasileiro de Banco de Dados*, 2013.

BOLLEN, J.; MAO, H.; ZENG, X. Twitter mood predicts the stock market. *Journal of Computational Science*, v. 2, n. 1, p. 1–8, 2011.

BRAVO-MARQUEZ, F.; MENDOZA, M.; POBLETE, B. Combining strengths, emotions and polarities for boosting twitter sentiment analysis. In: ACM. *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*. [S.l.], 2013. p. 2.

BROOKS, M. et al. Statistical affect detection in collaborative chat. In: ACM. *Proceedings of the 2013 conference on Computer supported cooperative work*. [S.l.], 2013. p. 317–328.

CHEN, L.; WANG, F. Sentiment-enhanced explanation of product recommendations. In: *Proceedings of the 23rd International World Wide Web Conference(WWW) - Companion Volume*. [S.l.: s.n.], 2014. p. 239–240.

CHERRY, C.; MOHAMMAD, S. M.; BRUIJN, B. D. Binary classifiers and latent sequence models for emotion detection in suicide notes. *Biomedical informatics insights*, Libertas Academica, v. 5, n. Suppl 1, p. 147, 2012.

DEMIRTAS, E. Cross-lingual sentiment analysis with machine translation. 2013.

EKMAN, P. Are there basic emotions? American Psychological Association, 1992.

EKMAN, P. An argument for basic emotions. *Cognition & emotion*, v. 6, n. 3-4, p. 169–200, 1992.

ESULI, A.; SEBASTIANI, F. Sentiwordnet: A publicly available lexical resource for opinion mining. In: *Proceedings of LREC*. [S.l.: s.n.], 2006. v. 6, p. 417–422.

GHAZI, D.; INKPEN, D.; SZPAKOWICZ, S. Prior and contextual emotion of words in sentential context. *Computer Speech & Language*, Elsevier, v. 28, n. 1, p. 76–92, 2014.

GO, A.; BHAYANI, R.; HUANG, L. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, p. 1–12, 2009.

HAJMOHAMMADI, M. S.; IBRAHIM, R.; SELAMAT, A. Cross-lingual sentiment classification using multiple source languages in multi-view semi-supervised learning. *Engineering Applications of Artificial Intelligence*, v. 36, p. 195–203, 2014.

HALL, M. et al. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, ACM, v. 11, n. 1, p. 10–18, 2009.

HU, M.; LIU, B. Mining and summarizing customer reviews. In: ACM. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.], 2004. p. 168–177.

KOEHN, P. Europarl: A parallel corpus for statistical machine translation. In: CITESEER. *MT summit*. [S.l.], 2005. v. 5, p. 79–86.

LIN, Z. et al. Make it possible: Multilingual sentiment analysis without much prior knowledge. In: *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Joint Conferences on*. [S.l.: s.n.], 2014. v. 2, p. 79–86.

LIN, Z.; TAN, S.; CHENG, X. Language-independent sentiment classification using three common words. In: ACM. *Proceedings of the 20th ACM international conference on Information and knowledge management*. [S.l.], 2011. p. 1041–1046.

LIU, B. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, v. 5, n. 1, p. 1–167, 2012.

LIU, B. *Opinions, Sentiment, and Emotion in Text*. [S.l.]: Cambridge University Press, 2015.

LIU, H.; LIEBERMAN, H.; SELKER, T. A model of textual affect sensing using real-world knowledge. In: ACM. *Proceedings of the 8th international conference on Intelligent user interfaces*. [S.l.], 2003. p. 125–132.

MARTINAZZO, B.; DOSCIATTI, M. M.; PARAISO, E. C. Identifying emotions in short texts for brazilian portuguese. 2011.

MARTINAZZO, B.; PARAISO, E. C. Identificação de emoções em notícias curtas. In: *CLEI-Conferência Latino-Americana de Informática*. [S.l.: s.n.], 2010. v. 1, p. 1–10.

MOHAMMAD, S. M. From once upon a time to happily ever after: Tracking emotions in mail and books. *Decision Support Systems*, v. 53, n. 4, p. 730–741, 2012.

MOHAMMAD, S. M. Imagisaurus: An interactive visualizer of valence and emotion in the roget's thesaurus. In: *6TH WORKSHOP ON COMPUTATIONAL APPROACHES TO SUBJECTIVITY, SENTIMENT AND SOCIAL MEDIA ANALYSIS WASSA 2015*. [S.l.: s.n.], 2015. p. 85.

MOHAMMAD, S. M. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. *Emotion Measurement*, 2015.

MOHAMMAD, S. M.; TURNEY, P. D. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In: *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*. [S.l.: s.n.], 2010. (CAAGET '10), p. 26–34.

MOLINA-GONZÁLEZ, M. D. et al. Semantic orientation for polarity classification in spanish reviews. *Expert Systems with Applications*, v. 40, n. 18, p. 7250–7257, 2013.

MOSTAFA, M. M. More than words: Social networks' text mining for consumer brand sentiments. *Expert Systems with Applications*, v. 40, n. 10, p. 4241 – 4251, 2013. ISSN 0957-4174.

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

NEVIAROUSKAYA, A.; PRENDINGER, H.; ISHIZUKA, M. Analysis of affect expressed through the evolving language of online communication. In: ACM. *Proceedings of the 12th international conference on Intelligent user interfaces*. [S.l.], 2007. p. 278–281.

NJOLSTAD, P. et al. Evaluating feature sets and classifiers for sentiment analysis of financial news. In: *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Joint Conferences on*. [S.l.: s.n.], 2014. v. 2, p. 71–78.

ORTONY, A.; TURNER, T. J. What's basic about basic emotions? *Psychological review*, v. 97, n. 3, p. 315, 1990.

PANG, B.; LEE, L. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, Now Publishers Inc., v. 2, n. 1-2, p. 1–135, 2008.

PANG, B.; LEE, L.; VAITHYANATHAN, S. Thumbs up?: sentiment classification using machine learning techniques. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. [S.l.], 2002. p. 79–86.

SANTOS, A. G. L. dos; BECKER, K.; MOREIRA, V. Um estudo de caso de mineração de emoções em textos multilíngues. In: *Proceedings of the Brazilian Workshop on Social Network Analysis and Mining (BraSNAM'13)*. [S.l.: s.n.], 2014.

SANTOS, A. L. dos; BECKER, K.; MOREIRA, V. Multilingual emotion mining using a parallel corpus. In: *Proceedings of 29th Brazilian Symposium on Databases (SBBD)*. [S.l.: s.n.], 2014. p. 127–136. In Portuguese.

SCHULZ, J. M.; WOMSER-HACKER, C.; MANDL, T. Multilingual corpus development for opinion mining. In: *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*. [s.n.], 2010. Available from Internet: <http://www.lrec-conf.org/proceedings/lrec2010/summaries/689.html>.

SILVA, M. J.; CARVALHO, P.; SARMENTO, L. Building a sentiment lexicon for social judgement mining. In: *Computational Processing of the Portuguese Language*. [S.l.]: Springer, 2012. p. 218–228.

SOLAKIDIS, G.; VAVLIAKIS, K.; MITKAS, P. Multilingual sentiment analysis using emoticons and keywords. In: *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Joint Conferences on*. [S.l.: s.n.], 2014. v. 2, p. 102–109.

STEINBERGER, J. et al. Creating sentiment dictionaries via triangulation. *Decision Support Systems*, v. 53, n. 4, p. 689 – 694, 2012.

STEINBERGER, J. et al. Multilingual entity-centered sentiment analysis evaluated by parallel corpora. In: CITESEER. *RANLP*. [S.l.], 2011. p. 770–775.

STRAPPARAVA, C.; MIHALCEA, R. Semeval-2007 task 14: Affective text. In: *Proceedings of the 4th International Workshop on Semantic Evaluations*. [S.l.: s.n.], 2007. p. 70–74.

STRAPPARAVA, C.; VALITUTTI, A. et al. Wordnet affect: an affective extension of wordnet. In: *LREC*. [S.l.: s.n.], 2004. v. 4, p. 1083–1086.

TAUSCZIK, Y. R.; PENNEBAKER, J. W. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, v. 29, n. 1, p. 24–54, 2010.

TSAKALIDIS, A. et al. Predicting elections for multiple countries using twitter and polls. *Intelligent Systems, IEEE*, v. 30, n. 2, p. 10–17, Mar 2015.

TSYTSARAU, M.; PALPANAS, T. Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery*, v. 24, n. 3, p. 478–514, 2012.

TUMASJAN, A. et al. Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, v. 10, p. 178–185, 2010.

TUMITAN, D.; BECKER, K. Sentiment-based features for predicting election polls: A case study on the brazilian scenario. In: *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Joint Conferences on*. [S.l.: s.n.], 2014. v. 2, p. 126–133.

VILARES, D.; ALONSO, M. A.; GÓMEZ-RODRIGUEZ, C. Sentiment analysis on monolingual, multilingual and code-switching twitter corpora. In: *6TH WORKSHOP ON COMPUTATIONAL APPROACHES TO SUBJECTIVITY, SENTIMENT AND SOCIAL MEDIA ANALYSIS WASSA 2015*. [S.l.: s.n.], 2015. p. 2.

VO, B.; COLLIER, N. Twitter emotion analysis in earthquake situations. *International Journal of Computational Linguistics and Applications*, v. 4, n. 1, p. 159–173, 2013.

WAN, X. Bilingual co-training for sentiment classification of chinese product reviews. *Computational Linguistics*, v. 37, n. 3, p. 587–616, 2011.

WARRINER, A. B.; KUPERMAN, V.; BRYSBAERT, M. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, Springer, v. 45, n. 4, p. 1191–1207, 2013.

A APÊNDICE

Este apêndice apresenta as tabelas usadas para as análises dos resultados apresentados nesta dissertação. Se encontram aqui as tabelas de contagem de atributos para cada *corpus* e algumas das tabelas dos resultados gerados nos experimentos 1 e 2. Os valores consideram as médias das execuções. No Experimento 1 foram 1680 execuções para cada *corpus*, enquanto no Experimento 2 foram 216 execuções, sendo inviável apresentar todos os resultados aqui.

A.1 Tabelas de resultados dos Experimentos

Tabela A.1: Resultados por emoções para o *corpus* SemEval - Medida-F

Emoção	Inglês	Max IN	Português	Max PT	Espanhol	Max ES	Francês	Max FR	Multilíngue	Max MULTI
anger	0,77	0,88	0,77	0,88	0,76	0,87	0,76	0,88	0,79	0,87
disgust	0,70	0,86	0,69	0,84	0,68	0,84	0,67	0,83	0,68	0,81
fear	0,73	0,94	0,71	0,95	0,71	0,94	0,72	0,95	0,72	0,87
joy	0,69	0,89	0,70	0,92	0,69	0,85	0,70	0,89	0,70	0,88
sadness	0,75	0,97	0,72	0,97	0,74	0,97	0,73	0,97	0,74	0,91
surprise	0,84	0,95	0,84	0,94	0,84	0,96	0,84	0,96	0,84	0,95

Fonte: a autora

Tabela A.2: Resultados por emoções para o *corpus* Notícias BR - Medida-F

Emoção	Português	Max PT	Inglês	Max IN	Espanhol	Max ES	Francês	Max FR	Multilíngue	Max MULTI
anger	0,77	0,88	0,77	0,88	0,76	0,87	0,76	0,88	0,79	0,87
disgust	0,70	0,86	0,69	0,84	0,68	0,84	0,67	0,83	0,68	0,81
fear	0,73	0,94	0,71	0,95	0,71	0,94	0,72	0,95	0,72	0,87
joy	0,69	0,89	0,70	0,92	0,69	0,85	0,70	0,89	0,70	0,88
sadness	0,75	0,97	0,72	0,97	0,74	0,97	0,73	0,97	0,74	0,91
surprise	0,96	1,00	0,96	1,00	0,96	1,00	0,96	1,00	0,96	1,00

Fonte: a autora

Tabela A.3: Resultados das classificações por idiomas para o *corpus* SemEval

	Precisão	**Revocação**	**Medida-F**
Multilíngue	0,78223	0,77076	0,74625
Inglês (Original)	0,77919	0,76760	0,74429
Português	0,77401	0,76253	0,73728
Espanhol	0,77291	0,76281	0,73731
Francês	0,77176	0,76154	0,73602

Fonte: a autora

Tabela A.4: Resultados das classificações por idiomas para o *corpus* Notícias BR

	Precisão	**Revocação**	**Medida-F**
Multilíngue	0,83482	0,86093	0,83376
Português (Original)	0,83611	0,85894	0,83248
Inglês	0,83346	0,85525	0,82961
Espanhol	0,83185	0,85519	0,82817
Francês	0,83179	0,85549	0,82841

Fonte: a autora

Tabela A.5: Resultados por seleção de atributos

		Precisão	**Revocação**	**Medida-F**
SemEval	null	0,72155	0,72516	0,70726
	InfoGain	0,83049	0,80494	0,77320
Notícias BR	null	0,79175	0,82632	0,79812
	InfoGain	0,87546	0,88800	0,86286

Fonte: a autora

Tabela A.6: Resultados por *weighting schemes*

		Precisão	**Revocação**	**Medida-F**
SemEval	booleano	0,78055	0,77146	0,74471
	TF-IDF	0,77149	0,75864	0,73575
Notícias BR	booleano	0,82321	0,85327	0,81693
	TF-IDF	0,84400	0,86105	0,84404

Fonte: a autora

Tabela A.7: Resultados por tokenizadores

		Precisão	**Revocação**	**Medida-F**
SemEval	Unigramas	0,77452	0,76718	0,74759
	Unigramas + Bigramas	0,77751	0,76292	0,73287
Notícias BR	Unigramas	0,84242	0,86134	0,84202
	Unigramas + Bigramas	0,82479	0,85298	0,81895

Fonte: a autora

Tabela A.8: Resultados por algoritmos

		Precisão	**Revocação**	**Medida-F**
SemEval	AdaBoost(RBF)	0,78325	0,76690	0,75935
	AdaBoost(SVM)	0,77223	0,76583	0,73968
	Bagging(RBF)	0,79458	0,78643	0,77537
	Bagging(SVM)	0,76816	0,75764	0,72655
	Naïve Bayes	0,75160	0,73427	0,67314
	RBF	0,79234	0,78088	0,77022
	SVM	0,76997	0,76342	0,73730
Notícias BR	AdaBoost(RBF)	0,83317	0,85244	0,83330
	AdaBoost(SVM)	0,83401	0,85786	0,83069
	Bagging(RBF)	0,84615	0,87254	0,84487
	Bagging(SVM)	0,82510	0,85362	0,81785
	Naïve Bayes	0,81851	0,83906	0,81435
	RBF	0,84538	0,86800	0,84354
	SVM	0,83944	0,84270	0,82880

Fonte: a autora

Tabela A.9: Resultados por emoção - Cruzamento de *corpora* - Inglês e Português - Medida-F

	Original-Tradução		**Tradução-Original**	
	Inglês	**Português**	**Inglês**	**Português**
anger	0,90499	0,30909	0,30856	0,89989
disgust	0,68396	0,43782	0,43508	0,69080
fear	0,05233	0,17610	0,17610	0,04250
joy	0,37158	0,22354	0,21703	0,38174
sadness	0,19980	0,15680	0,00228	0,21091
surprise	0,00228	0,02499	0,02493	0,00183

Fonte: a autora

Tabela A.10: Resultados por emoção - Cruzamento de *corpora* - Multilíngue

Emoção	**Precisão**	**Revocação**	**Medida-F**
anger	0,58630	0,64576	0,58512
disgust	0,64853	0,66899	0,56721
fear	0,18752	0,20128	0,09790
joy	0,56955	0,43088	0,32232
sadness	0,58321	0,31841	0,18955
surprise	0,04381	0,07357	0,01685

Fonte: a autora

Tabela A.11: Resultados por parâmetros - Cruzamento de *corpora* - Medida-F

		booleano	**TF-IDF**
Monolíngues	Unigramas	0,31854	0,28813
	Unigramas + Bigramas	0,29044	0,28531
Multilíngue	Unigramas	0,31413	0,30570
	Unigramas + Bigramas	0,26227	0,30387

Fonte: a autora

Tabela A.12: Resultados por algoritmos - Cruzamento de *corpora* - Medida-F

	AdaBoost(RBF)	**Bagging(RBF)**	**RBF**
Monolíngue	0,30075	0,29290	0,29317
Multilíngue	0,30984	0,29028	0,28936

Fonte: a autora

Tabela A.13: Diferença entre classificador Multilíngue e classificadores Monolíngues para o *corpus* SemEval

Tokenizador	Weighting Scheme	Seleção de Atributos	Algoritmo	Multilíngue	Monolíngues	Diferença
Unigramas	booleano	InfoGain	AdaBoost(RBF)	0,84411	0,87631	-0,03219
Unigramas	booleano	InfoGain	AdaBoost(SVM)	0,86515	0,85363	0,01152
Unigramas	booleano	InfoGain	Bagging(RBF)	0,92433	0,93188	-0,00755
Unigramas	booleano	InfoGain	Bagging(SVM)	0,84270	0,83225	0,01045
Unigramas	booleano	InfoGain	Naive Bayes	0,91083	0,89299	0,01784
Unigramas	booleano	InfoGain	RBF	0,90717	0,92643	-0,01926
Unigramas	booleano	InfoGain	SVM	0,86515	0,85030	0,01485
Unigramas + Bigramas	booleano	InfoGain	AdaBoost(RBF)	0,78655	0,78639	0,00015
Unigramas + Bigramas	booleano	InfoGain	AdaBoost(SVM)	0,77733	0,78124	-0,00391
Unigramas + Bigramas	booleano	InfoGain	Bagging(RBF)	0,80487	0,83617	-0,03130
Unigramas + Bigramas	booleano	InfoGain	Bagging(SVM)	0,76816	0,77075	-0,00260
Unigramas + Bigramas	booleano	InfoGain	Naive Bayes	0,80016	0,80502	-0,00487
Unigramas + Bigramas	booleano	InfoGain	RBF	0,80612	0,83481	-0,02868
Unigramas + Bigramas	booleano	InfoGain	SVM	0,77733	0,78124	-0,00391
Unigramas	booleano	null	AdaBoost(RBF)	0,79648	0,79574	0,00074
Unigramas	booleano	null	AdaBoost(SVM)	0,81527	0,81126	0,00400
Unigramas	booleano	null	Bagging(RBF)	0,79706	0,80514	-0,00808
Unigramas	booleano	null	Bagging(SVM)	0,81200	0,81029	0,00172
Unigramas	booleano	null	Naive Bayes	0,79013	0,78645	0,00368
Unigramas	booleano	null	RBF	0,79755	0,80481	-0,00726
Unigramas	booleano	null	SVM	0,81527	0,81126	0,00400
Unigramas + Bigramas	booleano	null	AdaBoost(RBF)	0,78609	0,78720	-0,00110
Unigramas + Bigramas	booleano	null	AdaBoost(SVM)	0,80731	0,81170	-0,00440
Unigramas + Bigramas	booleano	null	Bagging(RBF)	0,76355	0,76757	-0,00402
Unigramas + Bigramas	booleano	null	Bagging(SVM)	0,79629	0,79991	-0,00362
Unigramas + Bigramas	booleano	null	Naive Bayes	0,76399	0,76392	0,00008
Unigramas + Bigramas	booleano	null	RBF	0,76399	0,76816	-0,00416
Unigramas + Bigramas	booleano	null	SVM	0,80731	0,81170	-0,00440
Unigramas	TF-IDF	InfoGain	AdaBoost(RBF)	0,94387	0,90780	0,03607
Unigramas	TF-IDF	InfoGain	AdaBoost(SVM)	0,89214	0,87604	0,01610
Unigramas	TF-IDF	InfoGain	Bagging(RBF)	0,94081	0,90405	0,03676
Unigramas	TF-IDF	InfoGain	Bagging(SVM)	0,86945	0,85445	0,01500
Unigramas	TF-IDF	InfoGain	Naive Bayes	0,85288	0,82705	0,02583
Unigramas	TF-IDF	InfoGain	RBF	0,94387	0,90783	0,03604
Unigramas	TF-IDF	InfoGain	SVM	0,88820	0,86846	0,01975
Unigramas + Bigramas	TF-IDF	InfoGain	AdaBoost(RBF)	0,95258	0,92015	0,03243
Unigramas + Bigramas	TF-IDF	InfoGain	AdaBoost(SVM)	0,89531	0,88127	0,01404
Unigramas + Bigramas	TF-IDF	InfoGain	Bagging(RBF)	0,94895	0,91576	0,03320
Unigramas + Bigramas	TF-IDF	InfoGain	Bagging(SVM)	0,87224	0,85840	0,01384
Unigramas + Bigramas	TF-IDF	InfoGain	Naive Bayes	0,84804	0,83068	0,01736
Unigramas + Bigramas	TF-IDF	InfoGain	RBF	0,95258	0,92022	0,03235
Unigramas + Bigramas	TF-IDF	InfoGain	SVM	0,89363	0,87472	0,01891
Unigramas	TF-IDF	null	AdaBoost(RBF)	0,79291	0,79038	0,00254
Unigramas	TF-IDF	null	AdaBoost(SVM)	0,81527	0,81126	0,00400
Unigramas	TF-IDF	null	Bagging(RBF)	0,79296	0,79560	-0,00264
Unigramas	TF-IDF	null	Bagging(SVM)	0,81200	0,81029	0,00172
Unigramas	TF-IDF	null	Naive Bayes	0,77918	0,80007	-0,02089
Unigramas	TF-IDF	null	RBF	0,79291	0,78870	0,00421
Unigramas	TF-IDF	null	SVM	0,81527	0,81126	0,00400
Unigramas + Bigramas	TF-IDF	null	AdaBoost(RBF)	0,80003	0,79339	0,00664
Unigramas + Bigramas	TF-IDF	null	AdaBoost(SVM)	0,80731	0,81170	-0,00440
Unigramas + Bigramas	TF-IDF	null	Bagging(RBF)	0,80143	0,79903	0,00239
Unigramas + Bigramas	TF-IDF	null	Bagging(SVM)	0,79629	0,79991	-0,00362
Unigramas + Bigramas	TF-IDF	null	Naive Bayes	0,79111	0,80320	-0,01209
Unigramas + Bigramas	TF-IDF	null	RBF	0,80003	0,79339	0,00664
Unigramas + Bigramas	TF-IDF	null	SVM	0,80731	0,81170	-0,00440
			Média	**0,83376**	**0,82967**	**0,00410**

Fonte: a autora

Tabela A.14: Diferença entre classificador Multilíngue e classificadores Monolíngues para o *corpus* Notícias BR

Tokenizador	Weighting Scheme	Seleção de Atributos	Algoritmo	Multilíngue	Monolíngues	Diferença
Unigramas	booleano	InfoGain	AdaBoost(RBF)	0,82330	0,84954	-0,02623
Unigramas	booleano	InfoGain	AdaBoost(SVM)	0,75866	0,75443	0,00423
Unigramas	booleano	InfoGain	Bagging(RBF)	0,86042	0,87926	-0,01885
Unigramas	booleano	InfoGain	Bagging(SVM)	0,73303	0,73594	-0,00291
Unigramas	booleano	InfoGain	Naive Bayes	0,79292	0,75968	0,03324
Unigramas	booleano	InfoGain	RBF	0,85428	0,87974	-0,02546
Unigramas	booleano	InfoGain	SVM	0,75603	0,75496	0,00107
Unigramas + Bigramas	booleano	InfoGain	AdaBoost(RBF)	0,69364	0,84455	-0,15091
Unigramas + Bigramas	booleano	InfoGain	AdaBoost(SVM)	0,68495	0,69423	-0,00928
Unigramas + Bigramas	booleano	InfoGain	Bagging(RBF)	0,74992	0,90146	-0,15154
Unigramas + Bigramas	booleano	InfoGain	Bagging(SVM)	0,65871	0,66981	-0,01111
Unigramas + Bigramas	booleano	InfoGain	Naive Bayes	0,66965	0,68083	-0,01118
Unigramas + Bigramas	booleano	InfoGain	RBF	0,75522	0,89294	-0,13772
Unigramas + Bigramas	booleano	InfoGain	SVM	0,68209	0,69342	-0,01134
Unigramas	booleano	null	AdaBoost(RBF)	0,71585	0,71314	0,00271
Unigramas	booleano	null	AdaBoost(SVM)	0,73461	0,72601	0,00860
Unigramas	booleano	null	Bagging(RBF)	0,72115	0,72464	-0,00349
Unigramas	booleano	null	Bagging(SVM)	0,73264	0,72318	0,00946
Unigramas	booleano	null	Naive Bayes	0,70627	0,69269	0,01357
Unigramas	booleano	null	RBF	0,72510	0,72326	0,00185
Unigramas	booleano	null	SVM	0,73313	0,72391	0,00922
Unigramas + Bigramas	booleano	null	AdaBoost(RBF)	0,70857	0,70377	0,00480
Unigramas + Bigramas	booleano	null	AdaBoost(SVM)	0,72872	0,72570	0,00302
Unigramas + Bigramas	booleano	null	Bagging(RBF)	0,67489	0,70617	-0,03128
Unigramas + Bigramas	booleano	null	Bagging(SVM)	0,71171	0,71586	-0,00416
Unigramas + Bigramas	booleano	null	Naive Bayes	0,64920	0,65226	-0,00306
Unigramas + Bigramas	booleano	null	RBF	0,69026	0,71001	-0,01976
Unigramas + Bigramas	booleano	null	SVM	0,72805	0,72510	0,00295
Unigramas	TF-IDF	InfoGain	AdaBoost(RBF)	0,84970	0,79312	0,05658
Unigramas	TF-IDF	InfoGain	AdaBoost(SVM)	0,80476	0,77596	0,02880
Unigramas	TF-IDF	InfoGain	Bagging(RBF)	0,84971	0,79810	0,05161
Unigramas	TF-IDF	InfoGain	Bagging(SVM)	0,78209	0,76077	0,02132
Unigramas	TF-IDF	InfoGain	Naive Bayes	0,70690	0,62120	0,08569
Unigramas	TF-IDF	InfoGain	RBF	0,85083	0,79330	0,05753
Unigramas	TF-IDF	InfoGain	SVM	0,79627	0,77003	0,02624
Unigramas + Bigramas	TF-IDF	InfoGain	AdaBoost(RBF)	0,85314	0,79264	0,06050
Unigramas + Bigramas	TF-IDF	InfoGain	AdaBoost(SVM)	0,80354	0,77407	0,02947
Unigramas + Bigramas	TF-IDF	InfoGain	Bagging(RBF)	0,85151	0,79910	0,05241
Unigramas + Bigramas	TF-IDF	InfoGain	Bagging(SVM)	0,78195	0,75976	0,02219
Unigramas + Bigramas	TF-IDF	InfoGain	Naive Bayes	0,71654	0,62983	0,08671
Unigramas + Bigramas	TF-IDF	InfoGain	RBF	0,85416	0,79252	0,06164
Unigramas + Bigramas	TF-IDF	InfoGain	SVM	0,79581	0,76849	0,02732
Unigramas	TF-IDF	null	AdaBoost(RBF)	0,71959	0,67798	0,04161
Unigramas	TF-IDF	null	AdaBoost(SVM)	0,73461	0,72601	0,00860
Unigramas	TF-IDF	null	Bagging(RBF)	0,71582	0,70570	0,01012
Unigramas	TF-IDF	null	Bagging(SVM)	0,73264	0,72318	0,00946
Unigramas	TF-IDF	null	Naive Bayes	0,70602	0,63613	0,06989
Unigramas	TF-IDF	null	RBF	0,70354	0,68676	0,01678
Unigramas	TF-IDF	null	SVM	0,73313	0,72391	0,00922
Unigramas + Bigramas	TF-IDF	null	AdaBoost(RBF)	0,71696	0,69862	0,01835
Unigramas + Bigramas	TF-IDF	null	AdaBoost(SVM)	0,72872	0,72570	0,00302
Unigramas + Bigramas	TF-IDF	null	Bagging(RBF)	0,71468	0,70471	0,00997
Unigramas + Bigramas	TF-IDF	null	Bagging(SVM)	0,71171	0,71586	-0,00416
Unigramas + Bigramas	TF-IDF	null	Naive Bayes	0,71049	0,64426	0,06623
Unigramas + Bigramas	TF-IDF	null	RBF	0,70437	0,68929	0,01508
Unigramas + Bigramas	TF-IDF	null	SVM	0,72805	0,72510	0,00295
			Média	**0,74625**	**0,73872**	**0,00753**

Fonte: a autora

A.2 Tabelas gerais

As Tabelas A.15 e A.16 apresentam as médias de Medida-F para as configurações apresentadas, separadas por idioma.

Tabela A.15: Resultados gerais - Pré-processamento vs. Classificadores - Medida-F - SemEval

				Bagging(RBF)	AdaBoost(RBF)	AdaBoost(SVM)	Bagging(SVM)	Naïve Bayes	RBF	SVM
Multilíngue	Unigramas	boolean	InfoGain	0,86042	0,82330	0,75866	0,73303	0,79292	0,85428	0,75603
			null	0,72115	0,71585	0,73461	0,73264	0,70627	0,72510	0,73313
		TF-IDF	InfoGain	0,84971	0,84970	0,80476	0,78209	0,70690	0,85083	0,79627
			null	0,71582	0,71959	0,73461	0,73264	0,70602	0,70354	0,73313
	Unigramas + Bigramas	boolean	InfoGain	0,74992	0,69364	0,68495	0,65871	0,66965	0,75522	0,68209
			null	0,67489	0,70857	0,72872	0,71171	0,64920	0,69026	0,72805
		TF-IDF	InfoGain	0,85151	0,85314	0,80354	0,78195	0,71654	0,85416	0,79581
			null	0,71468	0,71696	0,72872	0,71171	0,71049	0,70437	0,72805
Inglês	Unigramas	boolean	InfoGain	0,87804	0,86315	0,76620	0,74562	0,77238	0,87918	0,76487
			null	0,73266	0,71717	0,72988	0,72791	0,70109	0,73213	0,72662
		TF-IDF	InfoGain	0,80063	0,79572	0,77982	0,76399	0,62086	0,79550	0,77372
			null	0,71253	0,67304	0,72988	0,72791	0,64081	0,69270	0,72662
	Unigramas + Bigramas	boolean	InfoGain	0,90953	0,87391	0,69944	0,67692	0,68480	0,90674	0,69971
			null	0,71875	0,70540	0,72953	0,72486	0,65597	0,72103	0,72875
		TF-IDF	InfoGain	0,80248	0,79366	0,77728	0,76163	0,62796	0,79329	0,77240
			null	0,71512	0,70851	0,72953	0,72486	0,64609	0,69252	0,72875
Português	Unigramas	boolean	InfoGain	0,87834	0,83969	0,74451	0,72920	0,75226	0,88199	0,74641
			null	0,72014	0,71416	0,72479	0,71786	0,68621	0,72092	0,72265
		TF-IDF	InfoGain	0,79396	0,79046	0,77358	0,76006	0,62063	0,79117	0,76882
			null	0,70686	0,68539	0,72479	0,71786	0,63756	0,68069	0,72265
	Unigramas + Bigramas	boolean	InfoGain	0,90884	0,86116	0,69051	0,66899	0,68146	0,90846	0,69104
			null	0,70237	0,69425	0,72262	0,71312	0,65455	0,70982	0,72255
		TF-IDF	InfoGain	0,79632	0,79202	0,77174	0,75986	0,63090	0,79287	0,76717
			null	0,69862	0,70157	0,72262	0,71312	0,64607	0,68913	0,72255
Espanhol	Unigramas	boolean	InfoGain	0,88059	0,84938	0,75572	0,73574	0,75493	0,88092	0,75575
			null	0,72398	0,71115	0,72606	0,72315	0,69320	0,72085	0,72628
		TF-IDF	InfoGain	0,79931	0,79596	0,77295	0,75910	0,61197	0,79603	0,76674
			null	0,70754	0,68541	0,72606	0,72315	0,63068	0,69815	0,72628
	Unigramas + Bigramas	boolean	InfoGain	0,89230	0,81452	0,69404	0,66628	0,67695	0,87479	0,69135
			null	0,70398	0,70704	0,72783	0,71340	0,65010	0,70693	0,72685
		TF-IDF	InfoGain	0,80024	0,79373	0,77247	0,75903	0,62118	0,79282	0,76605
			null	0,70608	0,69459	0,72783	0,71340	0,63950	0,69217	0,72685
Francês	Unigramas	boolean	InfoGain	0,88009	0,84593	0,75130	0,73318	0,75917	0,87686	0,75280
			null	0,72178	0,71007	0,72333	0,72380	0,69027	0,71911	0,72009
		TF-IDF	InfoGain	0,79849	0,79035	0,77749	0,75994	0,63135	0,79051	0,77083
			null	0,69588	0,66809	0,72333	0,72380	0,63547	0,67548	0,72009
	Unigramas + Bigramas	boolean	InfoGain	0,89518	0,82861	0,69292	0,66707	0,68009	0,88178	0,69159
			null	0,69960	0,70840	0,72282	0,71207	0,64843	0,70227	0,72225
		TF-IDF	InfoGain	0,79733	0,79116	0,77480	0,75853	0,63927	0,79109	0,76832
			null	0,69900	0,68978	0,72282	0,71207	0,64539	0,68332	0,72225

Fonte: a autora

Tabela A.16: Resultados gerais - Pré-processamento vs. Classificadores - Medida-F - Notícias BR

				Bagging(RBF)	AdaBoost(RBF)	AdaBoost(SVM)	Bagging(SVM)	Naïve Bayes	RBF	SVM
Multilíngue	Unigramas	boolean	InfoGain	0,92433	0,84411	0,86515	0,84270	0,91083	0,90717	0,86515
			null	0,79706	0,79648	0,81527	0,81200	0,79013	0,79755	0,81527
		TF-IDF	InfoGain	0,94081	0,94387	0,89214	0,86945	0,85288	0,94387	0,88820
			null	0,79296	0,79291	0,81527	0,81200	0,77918	0,79291	0,81527
	Unigramas + Bigramas	boolean	InfoGain	0,80487	0,78655	0,77733	0,76816	0,80016	0,80612	0,77733
			null	0,76355	0,78609	0,80731	0,79629	0,76399	0,76399	0,80731
		TF-IDF	InfoGain	0,94895	0,95258	0,89531	0,87224	0,84804	0,95258	0,89363
			null	0,80143	0,80003	0,80731	0,79629	0,79111	0,80003	0,80731
Inglês	Unigramas	boolean	InfoGain	0,93936	0,87075	0,85273	0,83091	0,89185	0,92982	0,85108
			null	0,80865	0,80102	0,82203	0,81364	0,78648	0,81129	0,82203
		TF-IDF	InfoGain	0,90790	0,91285	0,87949	0,85822	0,83141	0,91233	0,87275
			null	0,79677	0,79326	0,82203	0,81364	0,80538	0,79155	0,82203
	Unigramas + Bigramas	boolean	InfoGain	0,83805	0,78551	0,78310	0,77060	0,80479	0,84248	0,78310
			null	0,76691	0,78689	0,81408	0,80125	0,76371	0,76798	0,81408
		TF-IDF	InfoGain	0,91865	0,92372	0,88463	0,85877	0,83238	0,92370	0,87602
			null	0,80098	0,79444	0,81408	0,80125	0,80791	0,79444	0,81408
Português	Unigramas	boolean	InfoGain	0,92705	0,88448	0,85450	0,83552	0,89797	0,92444	0,84905
			null	0,80941	0,79443	0,80709	0,81075	0,79035	0,80413	0,80709
		TF-IDF	InfoGain	0,89769	0,90140	0,87313	0,84958	0,82924	0,90140	0,86204
			null	0,79620	0,78961	0,80709	0,81075	0,80408	0,78695	0,80709
	Unigramas + Bigramas	boolean	InfoGain	0,84091	0,78101	0,78327	0,77244	0,80691	0,83840	0,78327
			null	0,76960	0,79026	0,81499	0,80432	0,76369	0,77018	0,81499
		TF-IDF	InfoGain	0,91003	0,91526	0,87794	0,85503	0,83360	0,91526	0,87060
			null	0,79904	0,79598	0,81499	0,80432	0,80857	0,79598	0,81499
Espanhol	Unigramas	boolean	InfoGain	0,93028	0,86702	0,85721	0,83230	0,88760	0,92580	0,85310
			null	0,80082	0,79693	0,81041	0,81061	0,78298	0,80019	0,81041
		TF-IDF	InfoGain	0,90565	0,90780	0,87511	0,85546	0,82264	0,90812	0,86912
			null	0,79316	0,78871	0,81041	0,81061	0,79941	0,78818	0,81041
	Unigramas + Bigramas	boolean	InfoGain	0,83306	0,79060	0,77854	0,76959	0,80439	0,83075	0,77854
			null	0,76689	0,78556	0,80960	0,79896	0,76434	0,76688	0,80960
		TF-IDF	InfoGain	0,91641	0,91986	0,87763	0,85866	0,82855	0,92001	0,87300
			null	0,79284	0,78657	0,80960	0,79896	0,80136	0,78657	0,80960
Francês	Unigramas	boolean	InfoGain	0,93083	0,88297	0,85010	0,83026	0,89452	0,92567	0,84795
			null	0,80169	0,79058	0,80552	0,80615	0,78599	0,80362	0,80552
		TF-IDF	InfoGain	0,90497	0,90915	0,87643	0,85455	0,82490	0,90947	0,86993
			null	0,79626	0,78993	0,80552	0,80615	0,79141	0,78812	0,80552
	Unigramas + Bigramas	boolean	InfoGain	0,83267	0,78845	0,78008	0,77038	0,80400	0,82760	0,78008
			null	0,76690	0,78607	0,80815	0,79510	0,76392	0,76759	0,80815
		TF-IDF	InfoGain	0,91793	0,92177	0,88488	0,86113	0,82819	0,92192	0,87928
			null	0,80327	0,79658	0,80815	0,79510	0,79497	0,79658	0,80815

Fonte: a autora

A.3 Tabelas de contagem de atributos

As Tabelas A.17 e A.18 apresentam as contagens dos atributos para as estruturas monolíngue (médias da quantidade de atributos dos quatro idiomas) e da estrutura multilíngue para os *corpora* SemEval e Notícias BR.

Tabela A.17: Contagem de atributos - SemEval

Tokenizador	**Idioma**	**Atributos Originais**	**Emoção e Estrutura**	**Booleano**		**TF-IDF**	
				Média de atributos	% média de atributos	Média de atributos	% média de atributos
Unigramas	Monolíngues	3294,25	anger-Monolíngues	653,50	19,87	235,75	7,17
	Multilíngue	10042	anger-Multilíngue	2044	20,35	756	7,53
			fear-Monolíngues	1218,75	37,01	204,75	6,23
			fear-Multilíngue	3518	35,03	608	6,05
			sadness-Monolíngues	1117,75	33,96	164,25	5,00
			sadness-Multilíngue	3292	32,78	485	4,83
			joy-Monolíngues	1338,75	40,63	170,75	5,19
			joy-Multilíngue	4130	41,13	550	5,48
			surprise-Monolíngues	440,25	13,37	277,75	8,43
			surprise-Multilíngue	1400	13,94	863	8,59
			disgust-Monolíngues	1487,00	45,14	120,50	3,67
			disgust-Multilíngue	4626	46,07	384	3,82
Unigramas + Bigramas	Monolíngues	7749,75	anger-Monolíngues	861,75	88,87	270,75	96,51
	Multilíngue	30160	anger-Multilíngue	2899	90,39	902	97,01
			fear-Monolíngues	2624,00	66,13	270,75	96,50
			fear-Multilíngue	9827	67,42	850	97,18
			sadness-Monolíngues	2410,00	68,89	232,50	97,00
			sadness-Multilíngue	9114	69,78	753	97,50
			joy-Monolíngues	3102,00	59,97	234,75	96,97
			joy-Multilíngue	12024	60,13	820	97,28
			surprise-Monolíngues	978,00	87,38	778,50	89,96
			surprise-Multilíngue	3845	87,25	3153	89,55
			disgust-Monolíngues	3540,25	54,32	136,75	98,23
			disgust-Multilíngue	13831	54,14	451	98,50

Fonte: a autora

Tabela A.18: Contagem de atributos - Notícias BR

Tokenizador	**Idioma**	**Atributos Originais**	**Emoção e Estrutura**	**Booleano**		**TF-IDF**	
				Média de atributos	% média de atributos	Média de atributos	% média de atributos
Unigramas	Monolíngues	5988,25	anger-Monolíngues	367,00	6,14	147,00	2,45
	Multilíngue	18980	anger-Multilíngue	1221	6,43	457	2,41
			fear-Monolíngues	578,50	9,67	220,50	3,67
			fear-Multilíngue	1896	9,99	706	3,72
			sadness-Monolíngues	1623,50	27,11	159,75	2,67
			sadness-Multilíngue	5113	26,94	520	2,74
			joy-Monolíngues	2487,00	41,52	316,00	5,29
			joy-Multilíngue	7890	41,57	1039	5,47
			surprise-Monolíngues	393,75	6,59	165,25	2,76
			surprise-Multilíngue	1264	6,66	506	2,67
			disgust-Monolíngues	1479,00	24,72	149,50	2,50
			disgust-Multilíngue	4705	24,79	446	2,35
Unigramas + Bigramas	Monolíngues	20180,50	anger-Monolíngues	832,00	4,13	585,00	2,90
	Multilíngue	76107	anger-Multilíngue	3084	4,05	2216	2,91
			fear-Monolíngues	1429,50	7,08	1007,00	4,99
			fear-Multilíngue	5307	6,97	3886,00	5,11
			sadness-Monolíngues	5296,25	26,25	235,00	1,17
			sadness-Multilíngue	19938	26,20	805,00	1,06
			joy-Monolíngues	7965,75	39,47	435,25	2,16
			joy-Multilíngue	29875	39,25	1477,00	1,94
			surprise-Monolíngues	863,00	4,28	611,00	3,03
			surprise-Multilíngue	3131	4,11	2296,00	3,02
			disgust-Monolíngues	4928,25	24,43	237,25	1,18
			disgust-Multilíngue	18598	24,44	763,00	1,00

Fonte: a autora