WILLIAN AUGUSTO DIAS DOS REIS

Um Método de Identificação de Emoções Baseado na Mineração de Padrões Sequenciais

CURITIBA

2017

WILLIAN AUGUSTO DIAS DOS REIS

Um Método de Identificação de Emoções Baseado na Mineração de Padrões Sequenciais

Dissertação de Mestrado apresentado ao Programa de Pós-Graduação em Informática da Pontifícia Universidade Católica do Paraná como requisito parcial para obtenção do título de Mestre em Informática.

Área de Concentração: *Ciência da Computação*

Orientador: Prof. Dr. Emerson Cabrera Paraiso

CURITIBA

2017

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

AS	*Análise de Sentimentos*
IHC	*Interação Humano-Computador*
FP-Growth	*Frequent Pattern Growth*
FreeSpan	*Frequent pattern-projected Sequential pattern mining*
PrefixSpan	*Prefix-projected Sequential pattern mining*
GSP	*General Problem Solver*
DNA	*Deoxyribonucleic acid*
RNA	*Ribonucleic acid*
TF-IDF	*Term Frequency - Inverse Document Frequency*
MPF	*Mineração de Padrões Frequentes*
MPS	*Mineração de Padrões Sequenciais*

Resumo

As emoções fazem parte da vida do ser humano e podem ser expressas e percebidas por meio de gestos e expressões, da fala e por meio da escrita. Existe uma área de pesquisa que busca estudar as emoções no ambiente computacional: a Computação Afetiva. Entre as pesquisas realizadas em Computação Afetiva encontra-se a identificação automática de emoções em diferentes mídias, como os textos, conhecida como Análise de Sentimentos. As pesquisas em Análise de Sentimentos se iniciaram, e têm crescido, em decorrência da grande quantidade de dados que são disponibilizados diariamente na internet. Em uma imersão neste grande volume de dados é possível observar padrões, sejam comportamentais, sentimentais ou de escrita que podem ser usados em vários tipos de aplicações. Como uma subárea da mineração de dados, a mineração de padrões sequenciais obtém conhecimento através da análise dos dados identificando padrões. Com base nesse contexto, este trabalho apresenta um método de identificação de emoções em textos utilizando mineração de padrões sequenciais. O trabalho utiliza uma abordagem onde é possível identificar padrões que ocorrem com frequência na rotulação de emoções de textos e que podem ser aplicados na correção de problemas de classificação. O trabalho contém uma etapa de treinamento onde analisa-se um corpus rotulado com emoções e identifica padrões nestas rotulações. Na etapa de treinamento foram utilizados conceitos da biologia computacional, como alinhamento de sequencias, para identificar regiões de maior similaridade entre as sequencias formadas pelas rotulações e assim, pelo parâmetro de suporte mínimo, encontrar regiões que se repetiam nas sequencias. As regiões acima de um determinado suporte mínimo são consideradas padrões. Na etapa de teste um corpus não rotulado é submetido a um método de classificação. O método disponibiliza as probabilidades estimadas para cada instância classificada. Neste caso cada instância é um texto. Através das probabilidades estimadas de cada emoção para uma instancia é possível identificar momentos de confusão que o método teve ao classificar as instâncias. Os padrões encontrados na etapa de treinamento são utilizados então para tentar resolver estas confusões. Ao final deste processo é possível perceber que padrões de rotulação ajustaram pontos de confusão de outro classificador. Ao ser avaliado, o método obteve uma taxa de acerto de 65,1% ao identificar as seis emoções básicas em textos.

Palavras-Chave: Análise Sentimentos, Mineração de Padrões Sequenciais, Classificação, Padrão.

Abstract

Emotions are part of the life of the human being and can be expressed and perceived through gestures and expressions, speech and through writing. There is a research area that seeks to study the emotions in the computational environment: Affective Computing. Among the research carried out in Affective Computing is the automatic identification of emotions in different media, such as texts, known as Sentiment Analysis. The researches in Sentiment Analysis began, and have grown, due to the large amount of data that is made available daily on the internet. In an immersion in this large volume of data it is possible to observe patterns, be they behavioral, sentimental or writing that can be used in various types of applications. As a subarea of data mining, sequence pattern mining obtains knowledge by analyzing the data by identifying patterns. Based on this context, this work presents a method of identifying emotions in texts using sequence patterns mining. The work uses an approach where it is possible to identify patterns that occur frequently in the labeling of emotions of texts and that can be applied in the correction of classification problems. The work contains a training step where an emotion-labeled corpus analyzed and identifies patterns in these labeling. In the training phase, concepts of computational biology, such as sequence alignment, were used to identify regions of greater similarity between the transactions formed by the labeling and thus, by the parameter of minimum support, to find regions that were repeated in all the transactions. Regions above a given minimum support are considered patterns.

In the test step an unlabeled corpus is subjected to a classification method. The method provides the estimated probabilities for each classified instance. In this case each instance is a text. Through the estimated probabilities of each emotion for an instance it is possible to identify moments of confusion that the method had when classifying the instances. The patterns found in the training step are then used to try to resolve these confusions. At the end of this process it is possible to notice that labeling patterns have adjusted the confounding points of another classifier.

The method was evaluated through a corpus of English-language children's tales. When evaluated, the method obtained a hit rate of 65.1% by identifying the six basic emotions in texts.

Keywords: *Sentiment Analysis, Sequence Pattern Mining, Classification, Pattern*.

Capítulo 1

Introdução

A interação entre homem e computador, como tecnologia e área de pesquisa, tem crescido desde os anos 80 (BOOTH, 1995), (MYERS, 1998). O ápice da interação homem-computador, afirma-se ser quando um computador tem a capacidade de interagir com o usuário de forma natural, semelhante ao modo como acontece na interação entre humanos (SEBE; LEW; HUANG, 2004). Para tornar as interfaces computacionais mais "amigáveis", pesquisadores de diferentes áreas estudam a aplicação das Emoções.

As emoções podem ser expressas e percebidas de diversas formas. Entre elas estão os gestos e expressões, da fala e por meio da escrita. A Análise de Sentimentos (AS) é o campo que analisa opiniões, atitudes, sentimentos, avaliações e emoções das pessoas em diversas mídias, dentre elas os textos, de forma computacional. A AS é uma área multidisciplinar, envolvendo a Psicologia, a Mineração de Textos, passando pelo Reconhecimentos de Padrões até chegar na Interação Humano-Computador.

A web ajudou a impulsionar a pesquisa em AS. Comentários deixados por consumidores em sites que comercializam diferentes produtos, criaram uma subárea da AS, conhecida como Mineração de Opiniões onde algoritmos tentam identificar se os usuários estão falando bem ou mal de um produto qualquer. As pesquisas atuais estão focadas em identificar polaridade das emoções em textos, ou seja, identificar se os textos são positivos ou negativos. Alguns procuram inserir ainda a classe *neutro*, ou seja, ausência de emoção.

A relação entre emoções e métodos computacionais que as identificam, fez surgir uma área de pesquisa conhecida como Computação Afetiva (PICARD, 1995). Nesta área o principal objetivo é fazer com que a emoção existente entre as pessoas, também esteja presente na relação homem e computador.

A Mineração de Padrões é uma das tarefas no ramo de pesquisa da mineração de dados e consiste na descoberta de padrões interessantes, úteis e inesperados em um banco de dados. Este campo de pesquisa teve seu início nos anos 90 com o seminário de (AGRAWAL, SRIKANT, 1995). Esse artigo

introduziu o algoritmo *Apriori*, projetado para criar conjuntos de itens frequentes, que são grupos de itens que aparecem juntos em um banco de dados de transações de clientes, por exemplo compras em um supermercado.

Embora o padrão de mineração tenha se tornado muito popular devido às suas aplicações em muitos domínios, várias técnicas de mineração de padrões, como *frequent itemset mining* e mineração de regras de associação destinam-se a analisar dados, onde a ordem sequencial de eventos não é levada em consideração.

Assim, se essas técnicas de mineração de padrões forem aplicadas em dados com tempo ou informações de pedidos sequenciais, essas informações serão ignoradas. Isso pode resultar na falha em descobrir padrões importantes nos dados ou em padrões que podem não ser úteis porque ignoram a relação sequencial entre eventos ou elementos.

Para resolver este problema, foi proposta a tarefa de Mineração de Padrões Sequenciais (MPS). É uma solução proeminente para analisar dados sequenciais. Consiste em descobrir subsequências interessantes em um conjunto de sequências, onde a interação de uma subsequência pode ser medida em termos de vários critérios, tais como a frequência de ocorrência, o comprimento e o benefício.

Nesta pesquisa nos interessamos em unir a Computação Afetiva e os conceitos da MPS. Como identificado ao analisar o estado da arte, a maior parte dos pesquisadores realizam suas pesquisas interessados em identificar a polaridade dos textos. Neste trabalho o interesse é a identificação das seis emoções básicas (EKMAN, 1992), o que torna o problema mais complexo.

1.1 Motivação

Determinar as emoções em textos pode ser considerado um problema de classificação, onde: D é um conjunto de documentos e d um documento, onde $d \in D$; k é o número de classes e E o conjunto de rótulos das classes $E = \{ emo_1 \ emo_2, ..., emo_k\}$, onde emo_1 denota o caso especial de neutralidade ou ausência de emoção. O objetivo é determinar uma função, $f : d \rightarrow emo_1$, sendo que esta função mapeia um documento d em um dos elementos de E (DOSCIATTI, 2015).

Neste contexto um documento $d \in D$ e este conjunto D pode ser chamado de corpus. O corpus é o conjunto de dados que são submetidos à fase de treinamento de um método de classificação para extrair as informações dos textos e poder gerar parâmetros para a fase de classificação. Neste corpus cada documento d é rotulado com uma emoção E. Nesta pesquisa estudamos a hipótese de que hajam padrões que se repetem frequentemente em corpus com estas características. O corpus de contos

infantis, escrito na língua inglesa, e apresentando em (ALM; ROTH; SPROAT, 2005), contém textos em uma ordem cronológica que é característica de um conto infantil e, consequentemente, rótulos ordenados. Estes rótulos poderão ser utilizados no processo de identificação de padrões.

Neste trabalho é proposto um método para melhorar a identificação automática das emoções presentes em textos utilizando padrões extraídos de corpus utilizando a MPS. As emoções a serem identificadas nos textos se referem às seis emoções básicas propostas por (EKMAN, 1992) sendo elas *alegria*, *tristeza*, *raiva*, *medo*, *repugnância* e *surpresa*, além da classe *neutro*. Para a avaliação do método proposto foi obtido um *corpus* de textos de contos infantis na língua inglesa (ALM; ROTH; SPROAT, 2005).

O método aqui proposto possui algumas características que o difere dos demais trabalhos já desenvolvidos na área de AS: 1) utiliza padrões sequenciais para identificar emoções (EKMAN, 1992); 2) não utiliza recursos linguísticos (de um idioma qualquer); 3) é independente de idioma e não está vinculado a uma taxonomia específica de emoções, mesmo que neste trabalho seja utilizada a de Ekman (EKMAN, 1992).

O método proposto nesta pesquisa possui duas fases. Na primeira fase, de treinamento, ocorre a extração de padrões e o resultado é armazenado com as informações de suporte de cada padrão encontrado e o próprio padrão. Na segunda fase, de classificação, os padrões encontrados na fase anterior são utilizados para melhorar a classificação de um método de identificação de emoções já existente. Neste trabalho foi selecionado o método desenvolvido por Dosciatti (DOSCIATTI, 2015).

Na fase de treinamento foi proposto um novo algoritmo para extração de padrões. Os algoritmos da literatura que realizam a extração de sequencias buscam nas sequencias estes padrões, porém não se preocupam com a ordem dos *itens* dentro de uma transação na sequência. Alguns algoritmos, como *PrefixSpan* e *GSP*, que tratam dados ordenados, analisam a ordem das transações em uma sequência um banco de dados e não a ordenação dos itens em uma transação. Por isso, para atender a necessidade de identificar padrões nos itens de cada transação ordenada nas sequencias foi proposto o algoritmo *FP-Align*.

O *FP-Align* se utiliza de conceitos da biologia molecular, como alinhamento de sequencias, para identificar regiões de semelhança entre as transações e calcular o suporte mínimo destas regiões. O *FP-Align*, através de experimentos com a base de contos infantis, foi comparado com os algoritmos *FP-Growth* e *PrefixSpan*, onde foram analisadas as medidas de precisão, cobertura, F1 e acurácia.

1.2 Objetivos

O objetivo deste trabalho é desenvolver um método que utilize padrões sequenciais para a identificação das seis emoções básicas (Ekman et al., 1992) e *neutro* em textos.

Os objetivos específicos desta pesquisa compreendem:

- Desenvolver um algoritmo capaz de identificar padrões sequenciais utilizando algoritmos de alinhamento de sequencias;
- Comparar o algoritmo de identificação de padrões sequenciais com os algoritmos *FP-Growth* e *PrefixSpan*;
- Avaliar o método proposto através de experimentos.

1.3 Hipóteses de Trabalho

Neste trabalho são identificadas duas hipóteses a serem validadas. A primeira hipótese afirma que é possível extrair padrões sequenciais a partir de textos rotulados com emoções. A segunda hipótese afirma que é possível melhorar a identificação de emoções em textos utilizando padrões sequenciais obtidos através da rotulação.

1.4 Contribuição Científica

As contribuições desta pesquisa se enquadram no eixo cientifico. A principal contribuição deste trabalho se refere à disponibilização de um método capaz de melhorar a identificação de emoções em textos, utilizando padrões sequencias.

1.5 Organização do Documento

O Capítulo 2 apresenta a fundamentação teórica relacionada com a Computação Afetiva. Este capítulo inicia conceituando emoções e apresentando os principais modelos de emoções que surgiram a partir dos diversos estudos realizados por pesquisadores de diferentes áreas. Na sequência apresenta a definição de Computação Afetiva e de AS e mostra os conceitos relacionados à AS.

O Capítulo 3 apresenta as principais abordagens e conceitos relacionados Mineração de Padrões Sequenciais. Em seguida apresenta o conceito *Apriori* e algoritmos que implementam esta técnica como o algoritmo *FP-Growth*. Também apresenta os conceitos de transação, *itemset sequência e suporte mínimo*. Outros algoritmos de mineração de padrões sequenciais são

apresentados: GSP e *PrefixSpan*.

O Capítulo 4 apresenta o conceito de padrões aplicado na área da biologia computacional. Em seguida é apresentado o conceito de alinhamento de sequencias, seus tipos e aplicações e os algoritmos que implementam este conceito. No final é descrito o algoritmo *Needleman-Wusch* e como é implementado.

O Capítulo 5 apresenta o estado da arte. Este capítulo é dividido em três seções principais, sendo que a primeira seção deste capítulo trata de trabalhos de AS, desenvolvidos para textos, utilizando Mineração de Padrões Sequenciais, que identifiquem categorias de emoções em textos. A segunda seção é destinada à apresentação de trabalhos de AS em textos desenvolvidos para diversos idiomas que estejam relacionados, mesmo que indiretamente, com o método de identificação de emoções que está sendo proposto neste trabalho e a última seção trata de trabalhos que são trabalhos relacionados ao *FP-Align*.

O Capitulo 6 apresenta o método de pesquisa deste trabalho. Inicialmente é apresentada a caracterização do trabalho e a estratégia de pesquisa. O capítulo segue detalhando cada etapa da estratégia de pesquisa do trabalho desde a exploração do objeto de pesquisa até a avaliação do método proposto.

O Capítulo 7 apresenta o método. Este capítulo é dividido em quatro seções principais, sendo a que primeira apresenta os pressupostos do método. A segunda apresenta uma visão geral do método, mostrando de forma resumida as fases do método. A terceira seção apresenta a fase de treinamento do método, onde se extrai padrões de um banco de transações. A quarta apresenta a fase de classificação do método, aplicando os padrões para, melhorar a identificação de emoções.

O Capítulo 8 apresenta os resultados do método. Este Capítulo é dividido em três seções principais, sendo a primeira detalhar a implementação do método, seguido da segunda seção onde é realizada a descrição dos resultados produzidos pela Fase de Treinamento. Na terceira seção é detalhado os resultados produzidos pela Fase de Classificação e por último os resultados são avaliados e analisados com testes estatísticos.

Capítulo 2

Computação Afetiva

Este capítulo apresenta e conceitua-se as emoções e seu estudo em diferentes áreas como Filosofia, Psicologia e outras, culminando nos conceitos de Análise de Sentimentos. Também se define a Computação Afetiva, que estuda como os computadores podem ter a capacidade de se expressar emocionalmente com um humano, de forma simular à relação Humano-Humano. A última seção do capítulo se destina aos conceitos relacionados à Análise de Sentimentos, campo da Computação Afetiva destinado a análise de emoções, sentimentos e opiniões presentes principalmente em textos.

2.1. As Emoções

As emoções tem sido objeto de estudo em diversas áreas, tais como: a Psicologia, Neurociência, Filosofia e a Inteligência Artificial. A natureza das emoções é subjetiva, e há divergência dos pesquisadores quanto a sua origem. Também há divergência na terminologia, pois é utilizado para descrever uma enorme gama de estados cognitivos e fisiológicos. Por estas razões não existe unanimidade sobre sua definição (MICHAEL S. GAZZANIGA; T; HEATHERTON, 2005). Segundo (FEHR; RUSSELL, 1984), todas as pessoas sabem o que é emoção, até pedirem para defini-la.

A noção precisa do que se conhece por emoção é algo ainda tão incompleto quanto o conhecimento acerca de sua importância (ROMAN, 2007). Existem definições na literatura, porém são estabelecidas pela ótica da área de conhecimento que a defini. Quando se trata emoções na área da Psicologia, as emoções podem ser vistas como respostas sistêmicas que ocorrem quando ações altamente motivadas são proteladas ou inibidas (LANG, 1995), ainda

que estas ações realmente não tenham ocorrido (ROMAN, 2007). Segundo (LANG, 1995), as emoções dizem respeito à execução de algo importante ao organismo.

Na ótica de Charles Darwin (DARWIN, 1872), as emoções são estados emocionais expressos em forma de expressão facial é uma forma de comunicação compreensível e eficiente para as pessoas, independente de país ou povo, e ainda complementa que são adaptativas em todas as formas de vida.

Estas tentativas de definição para o termo emoção têm ainda maior relevância quando entra em discussão outro assunto: o sentimento. No escopo deste trabalho, é importante entender a sutil distinção entre emoção e sentimento. As emoções são entendidas como movimentos ou ações do corpo, "públicas" no sentido em que são visíveis para terceiros a olho nu (DAMASIO, 2003). Para o neurocientista Damásio em (DAMASIO, 2003), os sentimentos são "privados" do organismo em cujo cérebro ocorrem, ocultos para o público, escondidos de todos menos daquele que "tem o sentimento".

Uma vez apresentada a distinção entre emoção e sentimento, é possível citar como uma emoção pode afetar atividades diárias. O *humor*, por exemplo, que atinge a parte cognitiva, é utilizada em ambientes educacionais como forma de atrair e motivar a atenção, desenvolver sentimentos afetivos para o conteúdo ensinado e promover uma experiência de aprendizado mais prazerosa. Esta emoção possui um papel mais visível no dia a dia, além de consciente (NIJHOLT, 2003).

Para a área de ciências cognitivas, o estudo de afetividade resultou, ao longo dos anos, diferentes teorias cientificas, cada uma procurando explicar o fenômeno diversificado de emoções. Essas teorias originaram três principais modelos de emoções: os modelos discretos, também chamados de categóricos; os modelos dimensionais; e os modelos baseados na teoria *Appraisal*.

O modelo de emoção discreta teve sua origem em resultados de pesquisa afetiva realizados em experimentos com animais. Nos experimentos se estimulou as vias neurais e observou que o comportamento subsequente, ou inverso, ao induzir comportamentos e medir a atividade neural dos animais. Foi possível então construir uma taxonomia das emoções básicas (PANKSEPP, 1998).

Pode-se citar como principal vantagem dos modelos discretos a comprovação, por meio de experimentos psicofísicos, que a percepção das emoções pelos seres humanos é discreta,

dessa forma, estes modelos permitem facilmente associar as emoções com expressões faciais que as representam (LIBRALON, 2014).

O modelo de emoções básicas, segundo (LIBRALON, 2014), dentre todos existentes no modelo discreto, é o proposto pelo psicólogo Paul Ekman (EKMAN, 1992), que afirma a existência de seis emoções básicas: *alegria, tristeza, raiva, medo, repugnância* e *surpresa.* As seis emoções básicas de (EKMAN, 1992) são representadas por seis expressões faciais universais, pois são compreensíveis pelas pessoas em diferentes culturas e localidades. A expressão facial das seis emoções básicas é mostrada na Figura 2.1.

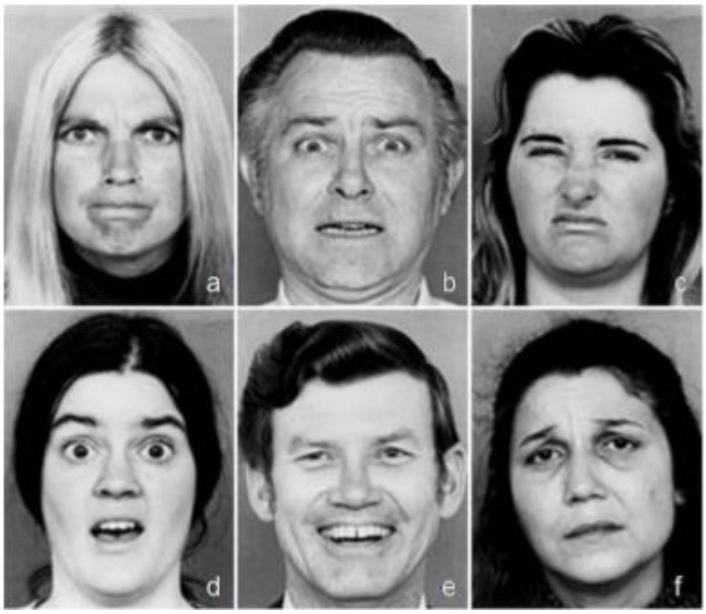

Figura 2.1. Expressão facial das seis emoções básicas de Ekman. (a) Raiva; (b) Medo; (c) Repugnância; (d) surpresa; (e) Alegria; (f) Tristeza. Extraído de: (EKMAN; FRIESEN, 1976)

Segundo Ortony e colegas (ORTONY; TURNER, 1990), além do modelo de emoções de Ekman, existem outros pesquisadores que fizeram suas categorizações. Dentre eles podemos listar alguns: (MCDOUGALL, 1926) composto pelas emoções *raiva*, *repugnância*, *alegria*, *medo* e *espanto*; (WATSON, 1930) composto pelas emoções *medo, amor* e *raiva*; (ARNOLD, 1960) composto pelas emoções raiva, *aversão*, *coragem*, *tristeza*, *desejo*, *desespero*, *medo*, *ódio*, *esperança*, *amor* e *tristeza* e (MOWRER, 1960) composto pelas emoções *dor* e *prazer*.

Segundo (EKMAN, 1993), não são todas as emoções que são acompanhadas de expressões faciais. Por isso, em (PANKSEPP, 1998) e (KAGAN, 2003), é descrito que os comportamentos afetivos não são suficientes e necessários para caracterizar todas as emoções.

Para exemplificar, a *ansiedade* pode ser sentida sem qualquer alteração evidente no comportamento.

Os modelos de emoções dimensionais, são fruto de outras teorias de emoções, e procuram descrever como existe relação entre as emoções, diferenciando-as de acordo com duas caracteristicas (modelos 2D) ou três características (modelos 3D). Se apresenta como o modelo mais difundido dentre os modelos dimensionais o proposto pelo psicólogo James A. Russel (RUSSELL, 1980). No modelo circunflexo de afeto, as emoções estão relacionadas entre si, expostas em circulo e com duas dimensões: valência (sentimentos agradáveis versus desagradáveis) e ativação (desperto versus sonolento).

Quando se trata do modelo de (RUSSELL, 1980), os pesquisadores que utilizam este modelo sugerem que cada experiencia afetiva é consequencia de uma combinação linear dessas duas dimensões, que é então, interpretado como representando uma emoção particular. Na Figura 2.2 é apresentado o modelo de (RUSSELL, 1980).

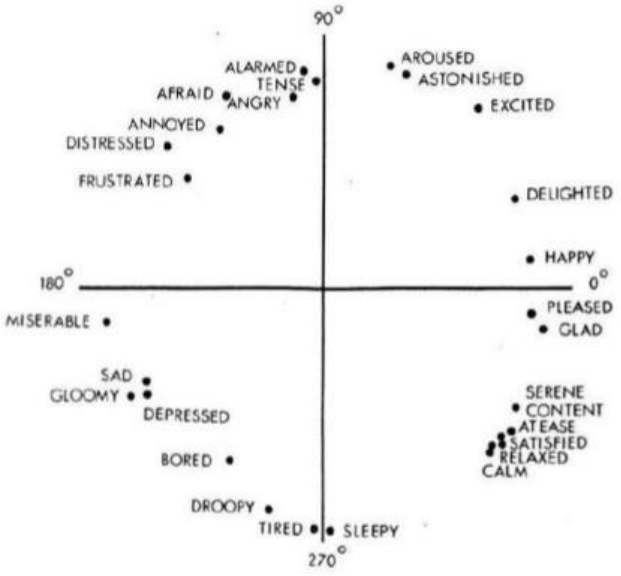

Figura 2.2 Modelo Circunflexo de afeto.
Extraído de: (RUSSELL, 1980)

Em 1999, o modelo de (RUSSEL, 1980) foi refinado por (WATSON et al., 1999). Para esta nova interpretação do modelo, a proximidade ou a distância entre as emoções representadas na circunferência pressupõem a semelhança ou a diferença entre as emoções. O modelo define que relação entre as emoções ocorre a partir do grau que as separa. As emoções perdem seu relacionamento positivo na medida que a distância entre elas se próxima a 90 graus. Por sua

vez, aos 180 graus de afastamento, os estados afetivos devem estar negativamente relacionados. Na Figura 2.3 o modelo (WATSON et al., 1999) é apresentado.

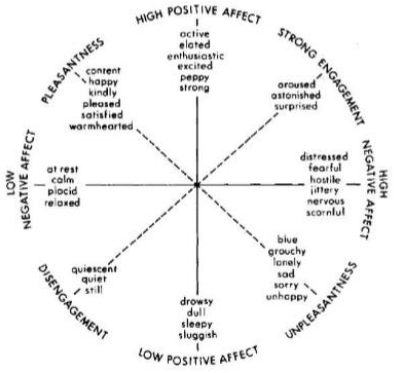

Figura 2.3 A estrutura bidimensional de afeto.
Extraído de: (WATSON et al., 1999)

O psicólogo Robert *Plutchik* propôs um modelo de dimensões em 3D (PLUTCHIK, 1980). Este modelo tem por característica principal permitir visualizar as relações existentes entre as emoções. O modelo se assemelha a um cone, onde a dimensão vertical representa a intensidade, e o círculo representa o grau de similaridade entre as emoções. Este modelo propõe oito emoções primarias sendo *alegria, tristeza, raiva, medo, repugnância, surpresa, expectativa* e *confiança*. Estas emoções primárias representam setores para indicar que há oito dimensões. As emoções em espaço em branco são combinações de duas emoções primarias. Este modelo é representado na Figura 2.4.

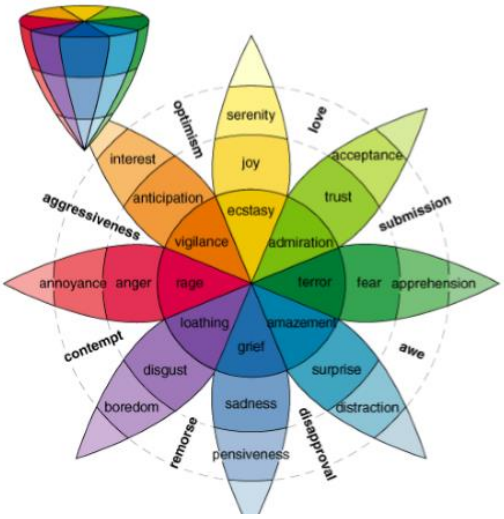

Figura 2.4. Modelo de emoções de Plutchik.
Extraído de: (PLUTCHIK, 1980)

O modelo de emoções PAD (iniciais de *Pleasure, Arousal, Dominance*) (MEHRABIAN, 1995) é outro exemplo de modelos 3D. A orientação para este modelo são as coordenadas nas três dimensões, pois as emoções constituem pontos em um espaço tridimensional. Neste modelo as três dimensões são: *satisfação, alerta* e *dominância.* A primeira dimensão, *satisfação*, distingue estados afetivos positivos e negativos, por exemplo, *exaltação, conforto, segurança versus aborrecimento, ansiedade, raiva*. A segunda dimensão, *alerta,* se refere ao estado de alerta e/ou de atividade física. A terceira e última *dominância* se refere ao nível de controle sobre pessoas ou situações.

O último modelo apresentado neste capítulo é o modelo baseado na teoria *Appraisal*, que consideram as emoções como resultado de avaliações subjetivas de um determinado evento ou situação que está acontecendo em um determinado momento (LAZARUS, 1991).

Neste modelo, a teoria mais conhecida é o de OCC (ORTONY; CLORE; COLLINS, 1988). Neste modelo as emoções são desenvolvidas como resultado das cognições e interpretações, sendo consideradas como reações com valência (positivo, negativo ou neutro) para eventos, agentes ou objetos. O modelo consiste em 22 emoções, sendo 11 positivas e 11 negativas. O conjunto de emoções é composto por: *satisfação - remorso, gratidão - raiva, orgulho - vergonha, admiração - censura, alegria - angústia, feliz* por - *ressentimento, soberba - pena, esperança - medo, satisfação - medo confirmado, confiança - desapontamento, amor - ódio*. Em (SMITH; LAZARUS, 1993), (ROSEMAN, 1996) e (SCHERER; SCHORR; JOHNSTONE, 2001) são propostos outros modelos na teoria *Appraisal.*

Os modelos de emoções apresentados, são modelos que categorizaram as emoções, porém mesmo assim ainda pode ser difícil para uma pessoa identificar as emoções expressas em outra. A dificuldade, segundo (ROLLS, 1998), (ORTONY, 2002), (DAVIDSON, 2003) e (ORTONY; NORMAN; REVELLE, 2004), ocorre pelo fato de humanos terem diferenças individuais, experiências diferentes vividas e outros fatores mais genéricos.

Na área das Ciências da Computação, este estudo de emoções tem atraído a atenção de diversos pesquisadores, sobretudo quando se trata da interação humano-máquina (GROSSMAN; FRIEDER, 2004). Computacionalmente, as emoções se referem à capacidade de uma máquina armazenar, processar, construir e manter um modelo emocional do usuário (LIBRALON, 2014). A área que trata a identificação de emoções em textos chama-se Analise

de Sentimentos (AS), que é uma subárea da Computação Afetiva. Ambas são definidas na próxima seção.

2.2. Computação Afetiva e Análise de Sentimentos

Na computação o estudo de emoções se relaciona com a Interação Humano-Computador (IHC). A IHC visa relacionar da melhor maneira possível usuários e computadores (BOOTH, 1995). A Computação Afetiva é uma interseção entre IHC e ciências cognitivas, e cada vez mais, devido ao uso intenso de computadores na comunicação entre pessoas, pesquisadores vem trabalhando no assunto.

A pesquisadora Rosalind Picard (PICARD, 1995), é uma referência nesta área e define que a habilidade de manipular emoções é fundamental, uma vez que, computadores com maiores habilidades afetivas, interagindo com maior qualidade com humanos, aumenta a possibilidade de aplicações lúdicas, cientificas e educativas. Para (PICARD, 1995), o afeto é algo intrínseco à natureza humana, e as pessoas naturalmente utilizam para interagir entre humanos, e assim também quando interagem com computadores. Se computadores apresentam reciprocidade neste afeto, e com qualidade, o processo de relação entre Homem-Máquina se torna mais natural.

Segundo Rosalind Picard (PICARD, 1995), o termo Computação Afetiva instiga aos pesquisadores da área a darem "habilidades emocionais" aos computadores. Para a habilidade de computadores interagirem de forma natural com humanos, segundo a autora, eles precisam ser dotados com as habilidades de reconhecer e expressar emoções. Neste contexto, trabalhos desenvolvidos para categorizar estados afetivos a partir de documentos escritos em linguagem natural, se enquadram em uma subárea da Computação Afetiva que se chama Análise de Sentimentos. Nas décadas de 1970 e 1980, segundo (BALAHUR; HERMIDA; MONTOYO, 2012), já existiam trabalhos com o intuito de identificar automaticamente emoções em texto, porém foi a partir de (PICARD, 1995) que pesquisas nesta área tomaram maiores proporções.

A Análise de Sentimentos (ou *Sentiment Analysis* em inglês) é o campo que estuda e analisa opiniões, sentimentos, avaliações, atitudes e emoções das pessoas a favor das entidades e seus atributos expressos em texto escrito (LIU, 2015). Este termo, apareceu primeiramente no trabalho de (NASUKAWA, 2003), porém é importante salientar que há muitos nomes e tarefas que são agrupados sob o conceito de AS, como por exemplo, *Mineração de Opiniões, Análise*

de Opiniões, Extração de Opiniões, Mineração de Emoções, Análise de Subjetividade, Análise de Emoções e *Mineração de Avaliação.* Academicamente se utiliza comumente dois termos, Análise de Sentimentos e Mineração de Opiniões. O termo Mineração de Opiniões apareceu pela primeira vez em (DAVE et al., 2003).

Para melhor compreensão e diferenciação entre os dois termos, em (MUNEZERO et al., 2014) se defini cada uma delas: Sentimentos são construções parcialmente sociais das emoções que se desenvolvem ao longo do tempo e são duradouras. As opiniões são interpretações pessoais sobre informações que podem ou não podem ser cobrados emocionalmente.

Uma premissa para se categorizar emoções, segundo (LIU, 2015), é saber se as emoções e sentimentos humanos podem ser expressos em linguagem natural e como eles podem ser reconhecidos. Segundo (LIU, 2015), a gramática e as expressões léxicas, estão contidas nas principais formas do ser humano se expressar: a fala e escrita. Porém, a emoção na forma não verbal pode ser expressa de outras formas, por exemplo, entonação, expressão facial, movimentos corporais, sinais biofísicos, gestos e postura. E na escrita: pontuação especial, a capitalização de todas as letras de uma palavra, *emoticons*, alongamento de palavras, dentre outros, são frequentemente utilizados para transmitir emoções, especialmente nos meios de comunicação social (DOSCIATTI, 2015).

A área de Processamento de Linguagem Natural (PLN), que estuda os problemas da geração e compreensão automática de línguas humanas naturais também se interessou nos estudos de emoções. Por esta área trabalhar com textos, a investigação de detectar afeto centrou-se na captura de palavras que contém emoções baseadas nos três modelos de emoções (vistos na Seção 2.1): modelos discretos, modelos dimensionais e modelos baseados na teoria *Appraisal*. Segundo (MUNEZERO et al., 2014), estes modelos são comumente utilizados em métodos de identificação de emoções em texto. Outra tentativa da área é detectar emoções em textos utilizando recursos lexicais como, por exemplo, o *WordNetAffect* (STRAPPARAVA; VALITUTTI, 2004) e o *SentiWordNet* (ESULI; SEBASTIANI, 2006). Estes recursos permitem identificar a presença de palavras com teor emocional nos textos.

Para (LIU, 2012), as crescentes pesquisas em PLN e AS sobre emoções impactam diversas áreas como por exemplo, gestão, política e ciências sociais e econômicas, onde as opiniões das pessoas estão presentes. Na AS, os métodos computacionais utilizam das informações disponibilizadas diariamente na web para identificar emoções, sentimento e opiniões de usuários expressos por meio de textos.

2.3. Análise de Sentimentos em Texto

Na Análise de Sentimentos, normalmente a identificação de textos ocorre através de *corpus* rotulados por anotadores de corpora. O grau de concordância entre os anotadores ao rotular um corpus é um coeficiente importante para descrever o quão é difícil identificar emoções em textos. A medida mais comum utilizada para medir esta concordância é a *Kappa* (CARLETTA, 1996), onde um valor menor do que 0.4 pode indicar uma anotação na qual não se pode confiar; se estiver entre 0.4 e 0.75, a anotação é satisfatória; e, se for maior do que 0.75, é muito boa. Para comprovar o fato de a Análise de Sentimentos ser uma tarefa complexa, observando que o próprio ser humano tem dificuldades em identificar emoções em textos, diversos corpora foram rotuladoss e obtidos o *Kappa* da anotação delas. O *corpus* de contos infantis desenvolvido por (ALM; ROTH; SPROAT, 2005) é um exemplo, o *corpus* de blogs desenvolvido por (AMAN; SZPAKOWICZ, 2007) é outro e o *corpus* de notícias desenvolvido por (STRAPPARAVA et al., 2007) e ainda o *corpus* de notícias para análise de sentimentos desenvolvido por (DOSCIATTI et al., 2015). Estes corpora são relevantes, sobretudo porque fazem parte do desenvolvimento de diversos métodos na AS. O grau de concordância entre os anotadores humanos, nestes corpora, entretanto, no melhor dos casos, pode chegar em torno de 0,75 (DOSCIATTI, 2015). Podemos concluir que, com este grau de concordância, mesmo humanos analisando emoções em texto tem dificuldade, o que prova ser uma tarefa difícil também para o computador.

As dificuldades encontradas na AS em textos são diversos, porém podemos elencar algumas: documentos não estruturados não facilitam o trabalho de identificação de emoções, além de estarem em formatos diferentes; o uso da linguagem informal, como gírias, dificulta a tarefa de identificação; em geral os métodos são sensíveis ao domínio e/ou ao idioma e a subjetividade com a qual estão expressas as opiniões e emoções em textos dificultam sua identificação.

Para (LIU, 2015), a AS pode ser aplicada em três níveis: em nível de documento, em nível de sentença e em nível de aspecto ou entidade. O nível de documento procura detectar opiniões, emoções e sentimentos como um todo, normalmente com uma única emoção para todo o documento. Um exemplo para este nível pode ser encontrado no trabalho de (MORAES; VALIATI; GAVIÃO NETO, 2013) e (DOSCIATTI et al., 2013).

O nível de sentença, devido à subjetividade, pode ser tratado de duas formas: Sentenças

que expressam informação factual (sentenças objetivas) de sentenças que expressam ponto de vista subjetivo (sentenças subjetivas) (LIU, 2015). Neste nível, o sentimento, a opinião ou a emoção é determinada considerando cada sentença do texto. Um exemplo de identificação de emoções por subjetividade pode ser visto em (WIEBE; BRUCE; O'HARA, 1999).

O nível de aspecto ou nível descrito por (LIU, 2015), determina um alvo de sentimento para o texto. Para exemplificar este nível, (LIU, 2012) utiliza o texto "A qualidade das chamadas do iPhone é boa, mas a vida útil da bateria é curta". Nesta frase são avaliados dois aspectos, a qualidade da chamada e a vida útil da bateria do produto. O primeiro aspecto, "qualidade da chamada", é positivo, porém o segundo, "vida útil da bateria" é negativo.

Outra característica a ser entendida na AS é o fato de poder ser realizada a partir de duas perspectivas. A primeira perspectiva é do ponto de vista do autor, que expressa a emoção e o sentimento nas palavras do texto. A segunda perspectiva é a dos leitores que leem o texto, que interpretam as emoções e sentimentos contidos nele. Importante considerar que vários leitores podem identificar emoções distintas a partir do mesmo texto lido. Em (LIU, 2012), o Texto: "O preço da habitação caiu, isso é ruim para a economia.", pode-se notar que o autor faz referência ao impacto negativo que o baixo preço da habitação está causando na economia. Para os vendedores que leem este texto, de fato é negativo, mas para os compradores, isto poderia significar uma boa notícia.

Para os autores (BALAHUR et al., 2010) ainda pode-se dizer que tem mais um ponto de vista: o que se refere ao sentimento expressamente explicito, e não o que está subentendido. O *corpus* que os autores utilizaram é do domínio de notícias jornalísticas e afirmam que para o autor e o ponto de vista do leitor, os textos são puramente informativos e os fatos são interpretáveis pela emoção que transmitem.

2.4. Conclusão

Neste capítulo foram apresentados conceitos como Computação Afetiva e a subárea que identifica emoções em mídias: Análise de Sentimentos. A principal ênfase foi dada ao estudo de emoções e, apesar de existirem diversas definições, muitos pontos de vista e inúmeros estudos desenvolvidos sobre o assunto, ainda não se chegou a um conceito definido de emoções.

Como a Análise de Sentimentos tem foco na identificação de emoções, sentimentos e opiniões, têm sido aplicados suas técnicas em outros tipos de corpora, como por exemplo, fala e imagens.

Como o foco deste trabalho de pesquisa é desenvolver um método de Análise de Sentimentos em textos, utilizando MPS, no próximo capítulo é apresentado os principais conceitos envolvidos nesta mineração.

Capítulo 3

Mineração de Padrões Sequenciais

Neste capítulo são apresentados os principais conceitos relacionados à Mineração de Padrões Sequenciais. Os conceitos como transação, sequencia, itens e suas características, são definidos. Algumas medidas, como por exemplo, Suporte Mínimo, serão apresentadas. Além disso, também são apresentados algoritmos que implementam a mineração de padrões sequenciais.

3.1. Conceitos

A Mineração de Dados é um ramo de pesquisa na computação que teve seu início anos 80. O que fez da Mineração de Dados possível foi a preocupação de grandes organizações, e seus profissionais, com o grande volume de dados estocados dentro da empresa, sobretudo porque estavam inutilizados. A mineração de dados consiste em extrair informações de dados armazenados em bancos de dados para compreender os dados e/ou tomar decisões. Algumas das tarefas de mineração de dados mais fundamentais são: agrupamentos, classificação, análise de *outliers* e mineração de padrões (AGRAWAL, 2015), (HAN;PEI;KAMBER, 2011).

Também chamada de *Data Mining*, no seu início a área consistia em extrair informações de enormes volumes de dados da maneira mais automática possível. Com a sua evolução, a Mineração de Dados começou a fazer análises dos dados após sua extração, buscando informações interessantes, úteis e inesperadas.

A Mineração de Padrões é uma das tarefas no ramo de pesquisa da mineração de dados e consiste na descoberta de padrões interessantes, úteis e inesperados em um banco de dados. Este campo de pesquisa teve seu início nos anos 90 com o seminário de (AGRAWAL,

SRIKANT, 1995). Esse artigo introduziu o algoritmo *Apriori*, projetado para criar conjuntos de itens frequentes, que são grupos de itens que aparecem juntos em um banco de dados de transações de clientes, por exemplo compras em um supermercado.

Os padrões, assim como as emoções, são alvo de interesse da Ciência da Computação pois através deles é possível extrair conhecimento sobre fatos, pessoas, ações, etc. A subárea que trata a relação padrões e computação é a Mineração de Padrões Frequentes (MPF).

Os padrões frequentes são conjuntos de itens frequentes, subsequências ou subestruturas que aparecem em um conjunto de dados, com frequência superior ou igual a um limite especificado pelo usuário (AGRAWAL, 2014).

Embora o padrão de mineração tenha se tornado muito popular devido às suas aplicações em muitos domínios, várias técnicas de mineração de padrões, como *frequent itemset mining* (AGRAWAL;SRIKANT, 1994), (HAN et al, 2004), (ZAKI, 2000), (PEI et al, 2001), (UNO;KIYOMI;ARIMURA, 2004) e mineração de regras de associação (AGRAWAL;SSRIKANT, 1994) destinam-se a analisar dados, onde a ordem sequencial de eventos não é levada em consideração.

Assim, se essas técnicas de mineração de padrões forem aplicadas em dados com tempo ou informações de pedidos sequenciais, essas informações serão ignoradas. Isso pode resultar na falha em descobrir padrões importantes nos dados ou em padrões que podem não ser úteis porque ignoram a relação sequencial entre eventos ou elementos. Em muitos domínios, a ordenação de eventos ou elementos é importante. Por exemplo, para analisar textos, muitas vezes é relevante considerar a ordem das palavras em sentenças (POKOU;FOURNIER-VIGER;MOGHRABI, 2016). Na detecção de intrusão de rede, a ordem dos eventos também é importante (PRAMONO, 2014).

Para resolver este problema, foi proposta a tarefa de Mineração de Padrões Sequenciais (MPS). É uma solução proeminente para analisar dados sequenciais (AGRAWAL;SRIKANT, 1995), (SRIKANT;AGRAWAL, 1996), (ZAKI, 2001), (ASERVATHAM;OSMANI;VIENNET, 2006), (HAN et al, 2000), (PEI et al, 2004), (AYRES et Al, 2002), (GOUDA;HASSAAN;ZAKI, 2010), (FOURNIER-VIGER et al, 2014), (YANG;KITSUREGAWA, 2005), (FOURNIER-VIGER et al, 2014A), (FOURNIER-VIGER et al, 2014B), (FOURNIER-VIGER;GOMARIZ;GUENICHE, 2013), (FOURNIER-VIGER;WU;TSENG, 2013), (FOURNIER-VIGER et al., 2008), (SALVEMINI et al, 2011), (MABROUKEH;EZEIFE, 2010). Consiste em descobrir subsequências interessantes em um

conjunto de sequências, onde a interação de uma subsequência pode ser medida em termos de vários critérios, tais como a frequência de ocorrência, o comprimento e o benefício.

A MPS tem inúmeras aplicações práticas devido ao fato de que os dados são naturalmente codificados como sequencias de símbolos em muitos campos, tais como bioinformática (WANG;HAN;LI, 2007), *e-learning* (FOURNIER-VIGER et al., 2008), (ZIEBARTH;CHOUNTA;HOPPE,2015) análise de cestas de mercado (SRIKANT;AGRAWAL, 1996), análise de texto (POKOU; FOURNIER-VIGER;MOGHRABI, 2016), redução de energia em *smarthomes* (SCHWEIZERet al, 2015), análise de fluxo nos cliques de páginas web (FOURNIER-VIGER;GUENICHE;TSENG, 2012), etc.

A MPS é um tópico de pesquisa muito ativo, onde vários trabalhos apresentam novos algoritmos e aplicações a cada ano, incluindo inúmeras extensões de mineração de padrões sequenciais para necessidades específicas. Por isso, pode ser difícil para os recém-chegados a este campo ter uma visão geral. Para abordar esta questão, foram publicas pesquisas que tratam sobre o assunto em (MABROUKEH;EZEIFE, 2010) e (FOURNIER-VIGER et al, 2017).

O problema da MPS foi proposto por *Agrawal* e *Srikant* (SRIKANT;AGRAWAL, 1996), como o problema de mineração de subsequências interessantes em um conjunto de sequências. Embora, originalmente foi projetado para ser aplicado a sequências, também pode ser aplicado em séries temporais após a conversão de séries temporais em sequências usando técnicas de discretização (LIN et al, 2007). Neste trabalho o interesse está na aplicação em sequencias, por isso para saber mais sobre mineração de padrões sequenciais em series temporais consultar.

Para entender os padrões sequenciais apresentamos o exemplo clássico da cesta de supermercado (SRIKANT;AGRAWAL, 1996). Suponha que um gerente de um supermercado esteja interessado em conhecer os hábitos de compra de seus clientes como, por exemplo, quais os produtos os clientes costumam comprar juntos e em qual ordem toda vez que vêm ao supermercado.

Conhecer a resposta a esta questão pode ser útil. Ele poderá planejar melhor os catálogos do supermercado, os folhetos de promoções de produtos, as campanhas de publicidade, além de organizar melhor a localização dos produtos nas prateleiras do supermercado colocando próximos os itens frequentemente comprados juntos para encorajar os clientes a comprar tais produtos conjuntamente.

Para isto, se dispõe de um conjunto de dados, que é o banco de dados de compras efetuadas pelos clientes. A cada compra de um cliente, são registrados neste banco todos os itens comprados. Para facilitar, na Tabela 3.1, são apresentados artigos do supermercado e a identificação de cada item.

Tabela 3.1: Itens do supermercado.

Id	Item
a	Pão
b	Leite
c	Açúcar
d	Manteiga
e	Fralda
f	Suco
g	Iogurte

Seja I = {i_1, i_2, ..., i_n} o conjunto de todos os itens do supermercado da Tabela 3.1. Um *itemset* ou conjunto de itens é um subconjunto não vazio de I. Um *itemset* com *k* elementos é chamado de *k-itemset*. Cada *itemset* comprado pelo cliente em uma única compra é chamado de *Transação.* Repare que, embora uma transação e um *itemset* sejam a mesma coisa (conjunto de itens), chamamos de transação somente aqueles *itemsets* que estão registrados no banco de dados como sendo a compra total feita por algum cliente.

Na MPS, além de definir o que é uma transação, faz necessário entender o que é uma sequência. Sem perda de generalidade, suponha que exista uma ordem total nos itens $\prec$, como uma ordem lexicográfica (por exemplo, $\prec$ b $\prec$ c $\prec$ d $\prec$ e $\prec$ f $\prec$ g). Uma sequência é uma lista ordenada de *itemsets* ou transações S = (I_1, I_2, I_3, ... I_n) tal que $I_k \subseteq I$ ($1 \leq k \leq n$). Por exemplo, considere a sequência ({a, b}, {c}, {f, g}, {g}, {e}) representando cinco transações feitas por um cliente em uma loja de varejo. Esta sequência indica que um cliente comprou itens *a* e *b* ao mesmo tempo, então comprou o item *c*, depois comprou itens *f* e *g* ao mesmo tempo, então comprou *g, e* finalmente comprou *e*.

Outra definição, uma sequência S_a = {A_1, A_2, ..., N_a} é dito de comprimento k ou uma sequência k se contiver k itens, ou em outras palavras se k = | A_1 | + | A_2 | + · · · + | A_n |. Por exemplo, a sequência ({a, b}, {c}, {f, g}, {g}, {e}) é uma sequência de tamanho 7.

Sendo assim, seja *D* um banco de dados de sequencias (BDS), uma tabela de duas colunas, a primeira correspondente ao atributo *SID* (identificador da sequência) e o segundo correspondente à sequência propriamente dita. Os elementos de *D* são chamados de sequencias e são representados na Tabela 3.2.

Tabela 3.2: Sequencias de compras em um supermercado.

SID	Sequencias
101	({a, b}, {c}, {f, g}, {g}, {e})
102	({a, b}, {c}, {f, g}, {g}, {e})
103	({a}, {b}, {f}, {e})
104	({b}, {f, g})

Repare que o que identifica uma sequência é o identificador SID e não o identificador do cliente, pois neste caso não será avaliado o cliente e sim o comportamento geral. Em outra análise, é possível adicionar o cliente para identificar padrões específicos dele.

Várias medidas podem ser usadas para avaliar o quão interessante é uma subsequência em MPS. No problema original da MPS, a medida de suporte é usada. O suporte (ou suporte absoluto) de uma sequência S_a, é definido como o número de sequências que contém S_a, e é denotado por $sup(S_a)$. Em outras palavras, $sup(S_a) = |\{s|s \text{ v } sa \wedge s \in BDS\}|$. Por exemplo, o suporte da sequência ({b}, {f, g}) no banco de dados da Tabela 3.2 é 2 porque esta sequência aparece em duas sequências: 101 e 104. Esta definição, chamada também de suporte relativo, é denotada por $relSup\ (S_a) = sup\ (S_a)$ / | SDB |, que é o número de sequências contendo S_a dividido pelo número de sequências no banco de dados. Por exemplo, o suporte relativo da subsequência ({b}, {f, g}) é 0,5.

Sobre uma sequência estar contida em outra. Chamamos este caso de subsequências. Para uma sequência ser subsequência de outra é preciso respeitar uma regra: a sequência $S_a = (A_1, A_2, ..., A_n)$ é dito estar contido em outra sequência $S_b = (B_1, B_2, ..., B_n)$ se e somente se houver números inteiros $1 \leq i_1 < i_2 < ... < em \leq m$ tal que $A_1 \subseteq B_{i1}$, $A_2 \subseteq B_{i2}$, ..., $A_n \subseteq B_{in}$ (denotado como $S_a \subseteq S_b$). Por isso, a sequência ({b}, {f, g}) está contida nas seqüências101 e 104, enquanto a sequência ({b}, {g}, {f} não está.

Suponha que o gerente decide que uma subsequência que aparece em pelo menos 2 de todas as sequencias registradas seja considerado um padrão. Por exemplo, se o banco de dados

de que você dispõe é o ilustrado na Tabela 3.2, então o *itemset* {a,b} é considerado padrão, pois aparece em 2 sequencias. O *suporte mínimo* de um *itemset* definido para este caso é 2.

Na Tabela 3.3 estão contabilizados os suportes de diversas subsequências encontradas no banco de dados de sequencias D.

Na próxima seção serão apresentadas em detalhes as características de algumas técnicas que tratam problemas de MPS, em especial métodos de origem *Apriori*.

3.2. Técnicas para Mineração de Padrões Sequenciais

Para a MPS, uma tarefa importante é encontrar padrões em um banco de dados de transações. Para (PEI et al., 2001), é uma atividade desafiadora, pois a busca gera um número combinatoriamente explosivo de padrões possíveis.

Tabela 3.3: Suporte Mínimo de cada sequência em um banco de dados.

Padrão	Suporte
({a})	3
({a},{g})	2
({a},{g},{e})	2
({a},{f})	3
({a},{f},{e})	2
({a},{c})	2
({a},{c},{f})	2
({a},{c},{e})	2
({a},{b})	2
({a},{b},{f})	2
({a},{b},{e})	2
({a} ,{e})	3
({a, b})	2
({b})	4
({b},{g})	3
({b},{g},{e})	2

({b},{f})	4
({b},{f, g})	2
({b},{f},{e})	2
({b},{e})	3
({c})	2
({c},{f})	2
({c},{e})	2
({e})	3
({f})	4
({f, g})	2
({f},{e})	2
({g})	3
({g},{e})	2

Pesquisadores da área vem procurando contribuir com formas de mineração eficazes, ou seja, evitando que sejam gerados um número explosivo de combinações para encontrar padrões sequenciais (PEI et al., 2001). Neste site[1] está organizado um repositório com algumas implementações e que podem ser avaliadas por pesquisadores.

As técnicas para mineração de padrões sequenciais iniciam com o *kernel* da propriedade *Apriori,* que defini*: se um padrão com k itens não é ;, nenhum de seus super padrões com (k+1) ou mais itens pode ser frequente.*

O algoritmo *Apriori* é um dos algoritmos mais reconhecidos para MPS. O Algoritmo aplica busca em profundidade e gera *itemsets* de *k* elementos a partir de *itemsets* de *k* - 1 elementos. Para este algoritmo padrões não frequentes são eliminados.

No Algoritmo 3.1 é possível observar que toda a base de dados é explorada e os *itemsets* frequentes são encontrados a partir de *itemsets* candidatos.

[1] http://fimi.ua.ac.be/

Algoritmo 3.1: Algoritmo Apriori.

```
F1 ←{Conjuntos de itens freqüentes de tamanho 1} /* Na
    primeira passagem k = 1                          */
1 para k = 2; Fk−1 != vazio; k++ faça
  |  /* Na segunda passagem k = 2                     */
2 |  Ck ← apriori-gen (Fk−1) /* Novos candidatos     */
3 |  para todo transação t ∈ T faça
4 |  |  Ct ← subconjunto (Ck, t) /* Candidatos contidos
  |  |      em t                                     */
5 |  |  para todo candidato c ∈ Ct faça
6 |  |  |  c.contagem++
7 |  |  fim
8 |  |  Fk ← {c ∈ Ck|c.contagem ≥ MinSup}
9 |  fim
10 fim
11 Resposta F ← Reunião de todos os Fk
```

No Algoritmo 3.1, duas variáveis são importantes de serem detalhadas: F_k e C_k. A variável F_k é o itemset frequente de tamanho k que contempla o suporte mínimo estabelecido pelo usuário. Cada instancia deste conjunto tem dois atributos. O primeiro atributo é o *itemset* e segundo o contador de suporte. A variável C_k descreve o *itemset* candidatos de tamanho k. Cada instancia deste conjunto tem também dois campos que são o *itemset* e o contador de suporte.

O algoritmo *Apriori* faz uso de duas sub-rotinas chamadas: *apriori-gen* e *subconjunto*. A sub-rotina *apriori-gen* é implementada para gerar o *itemset* candidatos que são o conjunto composto pelos valores correspondentes ao suporte de cada item. Neste conjunto são considerados todos os itens, independente se estão acima ou abaixo do valor de suporte mínimo. A sub-rotina *subconjunto* implementa a extração de regras de associação, informação posterior à identificação de padrões. Em função das regras de associação, este algoritmo calcula além do suporte mínimo outra medida que é a confiança, da qual não apresentaremos neste momento.

O algoritmo trabalha sobre uma base de sequencias em busca de padrões sequenciais, isto quer dizer que são aqueles que possuem um suporte maior ou igual ao suporte mínimo. Por isso, como entrada para o algoritmo são necessários o valor de suporte, a confiança e um arquivo contendo as sequencias.

O algoritmo, na primeira passagem, conta o valor de suporte para cada item individual (*itemset* de $k = 1$) e todos aqueles que satisfazem o suporte mínimo são selecionados,

constituindo o *itemset* de *1* item frequente (F_1).

Na segunda iteração, *itemsets* candidatos de tamanho *k=2* são gerados pela união dos *itemsets* de *k=1* com os novos. Este processamento é realizado pela sub-rotina *apriori-gen*. São calculados os suportes para estes novos *itemsets* e assim são encontrados os *itemsets* de tamanho *k*=2. Estas iterações então em repetição constante até que um *itemset* de tamanho $k = n$ seja vazio.

Ao final destas iterações o resultado é uma tabela com todos os *itemsets* de tamanho *k=n* que respeitem o suporte mínimo dado como entrada no algoritmo.

Percebeu-se com novas pesquisas que os algoritmos para mineração de padrões sequencias poderiam ser mais performáticos comparado com o *Apriori*.

Segundo (PEI et al., 2001b), diversos métodos foram desenvolvidos para melhor atender a exploração de padrões sequenciais em uma base de dados. Estes algoritmos podem ser divididos em duas categorias: 1) abordagem de geração e teste de candidatos, como o GSP *(General Problem Solver);* 2) algoritmo de crescimento de padrões, que tem como um de seus representantes o *FP-Growth* (HAN; PEI; YIN, 1999).

Como vimos, algoritmos *Apriori* executam múltiplos passos, e esta estrutura é aplicada sobre os dados. Para todos os passos, se inicia por um *itemset* semente que são chamados de *itemset* candidatos. O elemento suporte mínimo, dentro deste tipo de algoritmo é calculado e ao final são determinados os *itemsets* que compõem a semente para o próximo passo (SRIKANT; AGRAWAL, 1996). O elemento suporte é um filtro para evitar a "explosão" de todas possibilidades combinatórias entre os *itemsets* para encontrar padrões de repetição.

O algoritmo *GSP* (*Generalized Sequential Pattern*), segundo (SRIKANT; AGRAWAL, 1996), escala linearmente de acordo com o número de *itemsets* e possui propriedades escaláveis respeitando o número de sequencias e o número de itens por sequência. Por se tratar de um algoritmo que busca subsequências, é importante destacar que os *itemsets* são gerados levando em consideração a ordem dos itens, pois *itemsets* nestas condições são chamados de subsequências.

O algoritmo *GSP* é executado passo-a-passo e a ordem de execução é dada por: o algoritmo encontra todos os *itemset* de tamanho *k=1*. Destes, os mais frequentes são separados e é formado um conjunto de *itemset* candidatos de tamanho *k=2*. O próximo passo é realizado para recolher o suporte dessas sequências. Assim segue a execução para gerar *itemsets* candidatos de k=*3*. O processo é repetido até que nenhuma sequência frequente seja encontrada.

Por se tratar de um algoritmo baseado em *Apriori*, os passos são parecidos com o algoritmo *Apriori* descrito no Algoritmo 1, porém é relevante lembrar que este algoritmo busca subsequências, o que quer dizer que um *itemset* {a, b, c} é diferente de um *itemset* {c, b, a}.

Outro exemplo de algoritmo é o SPADE (ZAKI, 2001). Este algoritmo é baseado na execução *Apriori* e também é aplicado para encontrar subsequências.

Nesta primeira categoria de métodos, que é a geração de candidatos se atinge bons resultados de desempenho devido a geração de candidatos ser reduzida por filtros como suporte mínimo. No entanto, quando o suporte é baixo ou tamanho dos padrões gerados em muito grande, há um aumento de custos de processamento não-triviais, independentes de técnicas utilizadas para a implementação, que são detalhadas em (HAN; PEI; YIN, 1999).

A segunda categoria de métodos, que é chamada de crescimento de padrões, foi desenvolvida para sanar problemas na primeira categoria. Suas características são:

- A análise de contagem de frequência de conjuntos de dados relevantes é a forma utilizada para evitar a geração de candidatos, porque o método preserva os agrupamentos essenciais dos elementos originais.
- A busca é feita em espaços reduzidos, uma vez que se utiliza métodos de divisão e conquista para isto. O conjunto de dados e o conjunto de padrões a ser examinado em cada passo é particionado.
- Devido a esta divisão e aliado à capacidade de crescimento da memória principal, torna-se possível processar o maior número de dados em uma memória principal.

Os algoritmos que mais se destacaram nesta categoria são: FP-Growth *(**F**requent **P**attern - **Growth**)*) (HAN; PEI; YIN, 1999); FreeSpan *(**Fre**qu**e**nt pattern-projected **S**equential **pa**ttern mining)* (HAN et al, 2000) e PrefixSpan (***Prefix**-projected **S**equential **pa**ttern mining)*.

O FP-Growth (HAN; PEI; YIN, 1999) é um método escalável e com um bom desempenho para a MPS, sejam eles de diversos tamanhos, curtos ou longos. Ele utiliza uma estrutura chamada *FP-Tree*, baseada no crescimento de fragmentos de padrões, que armazena informação quantitativa sobre padrões frequentes de forma comprimida.

O algoritmo FreeSpan utiliza uma técnica onde os itens frequentes são utilizados para projetar recursivamente bancos de dados de sequências e em um conjunto menor de bancos de dados e crescer os fragmentos de padrões em cada banco de dados projetado (HAN et al, 2000*).*

No algoritmo PrefixSpan, se explora a projeção de prefixos nos *itemsets*. A ideia principal

é que, ao invés de projetar *itemsets* de bancos de dados considerando-se todas as ocorrências possíveis de padrões, a projeção seja baseada apenas em prefixos frequentes porque qualquer padrão crescente pode sempre ser encontrada pelo crescimento de um prefixo crescente (PEI et al, 2001).

Os algoritmos originários da teoria de crescimento de padrões são, no geral, mais eficientes e escaláveis do que outros métodos da MPS. Isto ocorre devido a alguns fatores: 1) adoção de estratégia de divisão e conquista; 2) integração do uso de memória principal ao uso do disco para algoritmos de projeção de bancos de dados; 3) uso das propriedades *Apriori*, de forma correta, evitando a geração de um grande número de candidatos (HAN; PEI; YIN, 1999). No domínio de algoritmos *Apriori*, os que utilizam crescimento de padrões são mais aceitos pelos pesquisadores. Contudo, alguns trabalhos demonstram que algoritmos *Apriori* são inapropriados em bases que contêm padrões muito longos (BAYARDO, 1998).

3.3. Conclusão

Neste capítulo foram abordados conceitos fundamentais que orientam este trabalho. Na seção 3.1 foram apresentados os conceitos de mineração de dados. Na seção 3.2, foram apresentados conceitos de mineração de padrões frequentes e mineração de padrões sequenciais, assim como a medida de suporte mínimo que indica os padrões de interesse em uma base de dados e outras medidas adicionais. E na seção 3.3 métodos que já trabalham com mineração de padrões sequenciais.

Capítulo 4

Análise Comparativa de Sequencias

Neste capítulo são apresentados os principais conceitos relacionados à Análise comparativa de sequencias. Estes conceitos complementam a fundamentação teórica deste trabalho para compreender como o método proposto analisa as transações (textos) submetidas a ele, para extrair os padrões de interesse utilizando a abordagem de comparação de sequência.

Sendo assim, a seguir, nas seções 4.1, 4.2 e 4.3 são apresentados os conceitos da biologia e como são encontrados padrões na análise biológica, assim como, o alinhamento de sequencias, e algoritmos que implementam o alinhamento de sequencias, respectivamente.

4.1. Padrões e Biologia

Buscar padrões não se limita a um determinado domínio. Áreas como a Biologia Molecular fazem parte deste conjunto de áreas onde os padrões sequenciais são importantes para o conhecimento.

Segundo (BOGUSKI, 1998), o uso de ferramentas computacionais na área de Biologia Molecular já estavam presentes anos antes do sequenciamento de DNA ser comumente utilizado nos laboratórios de pesquisa da área.

Para esta pesquisa, entender alguns trabalhos da área da Biologia Molecular é importante. Algoritmos utilizados no conceito de *alinhamento de sequencias,* serão aprimorados no método de identificação de emoções, para encontrar padrões sequenciais. Há uma proximidade na estrutura de *itemset* da Biologia Molecular com as deste trabalho, variando, neste caso, os itens que compõem os *itemsets*.

Como descrito no capito de MPS, esta área atua na busca de subsequências, que são *itemsets* gerados levando em consideração a ordem dos itens.

A partir do momento que se identificou que DNA, RNA (*Ribonucleic acid*) e proteínas são portadores de informações do sistema biológico de seres vivos (HAGEN, 2000), as pesquisas foram direcionadas a extrair informações destas sequencias com métodos adequados.

Para exemplificar o uso de sequencias como meio de extrair conhecimento na Biologia Molecular, pode-se citar os padrões que regulam a função de genes no DNA, ou padrões que são comuns em membros de uma mesma família de proteínas.

As sequencias de DNA são a fonte das informações e padrões extraídos. Estas sequências compõem bancos de dados, e estes bancos, muitos deles são públicos, onde se pode descarregar o arquivo e trabalhar. Um destes bancos é o *PubMed*[2]. Estas sequencias são compostas pelos itens I = {A, C, G, T}. As letras que compõem este conjunto de itens são abreviações de quatro nucleotídeos: Adenina; Citosina, Guanina e Timina. Outros bancos são compostos por proteínas, e para este caso, o conjunto de itens é composto por vinte aminoácidos.

Quando se trata de problemas biologia os pesquisadores identificam alguns tipos de problemas que se tornaram objeto de investigação. São eles: Descobrir padrões significantes, Reconhecimento de padrões e Classificação.

Para um problema de *descobrir padrões significantes*, uma classe de padrão pela qual o pesquisador tem interesse é definida e uma busca no banco de dados de sequencias é feita procurando descobrir os padrões desta classe. Neste problema, o que filtra os padrões, assim como nos métodos *Apriori* é uma propriedade chamada suporte mínimo. O suporte de um padrão é medido pelo número de sequencias nas quais o padrão ocorre, seja um número absoluto ou um percentual de todo o conjunto de sequencias.

Problemas de *Reconhecimento de Padrões* estão inclusos na área de Biologia Molecular e que também procuram padrões em sequencias. Porém, neste caso, os padrões são conhecidos, seja ele, por sua função biológica ou por seu maleficio ao sistema biológico de alguns seres vivos. Por esta razão, programas para reconhecimento de padrões buscam estes padrões em função do seu conhecimento prévio sobre ele. Isto facilita enormemente, pois as características do padrão podem informar como encontrá-lo.

Nos *Problemas de Classificação* a principal característica é o fato de utilizar padrões de sequencias, que ocorrem em uma determinada classe, para identificar elementos não classificados. Por exemplo, classificar proteínas em sua respectiva família. Dada uma proteína desconhecida, pode-se classifica-la como membro ou não de uma família, baseando-se pelo fato

de conter ou não os padrões pertinentes a alguma família.

Um complemento para MPS, aplicados nas Ciências Biológicas pode ser encontrado no *survey* de (FLORATOS, 1999). Este autor elenca alguns algoritmos importantes que procuram resolver os tipos de problemas elencados. Um destes algoritmos é o TEIRESIAS. Em (BREJOVA et al., 2000), o leitor vai encontrar endereços de diversos bancos de dados de proteínas e instruções para carregá-los. Uma discussão detalhada dos tipos de problema em mineração de padrões sequenciais em biologia pode ser encontrada em (BRAZMA et al., 1998).

Todos os problemas mencionados procuram extrair características das sequencias analisadas para, posteriormente, se utilizar destas características para alguma finalidade, seja classificação, reconhecimento de padrões ou a descoberta de padrões significantes. Uma forma utilizada para extrair características de sequencias biológicas é o alinhamento de sequencias.

O alinhamento de sequencias é uma forma de organizar sequencias para identificar regiões similares entre as sequencias. Estas regiões similares podem representar que as sequencias tem relações funcionais biológicas, estruturais ou até que sejam de um mesmo ascendente biológico. Quanto estas regiões similares aparecem frequentemente em diversas sequencias é possível compreende-las como padrões sequenciais. Na próxima seção o alinhamento de sequencias é descrito, assim como seus tipos e algoritmos.

4.2. Alinhamento de Sequencias

De forma simplificada, suponha que o alinhamento de sequencias seja colocar uma sequência em uma linha e outra sequência em outra linha abaixo dela e passar por cada coluna verificando as igualdades entre elas, sendo que cada coluna é composta por elementos que derivem de um conjunto comum.

Na prática, além de se colocar uma abaixo da outra é necessário introduzir *gaps* (espaços) dentro das sequencias, fazendo que as sequencias fiquem no mesmo comprimento e ainda admitindo substituições. Os *gaps* podem ser inseridos, representando uma remoção na sequência ou ainda uma inserção nas demais sequencias alinhadas. É importante lembrar que *gaps* podem estar no início, meio e fim das sequencias.

Para (SETUBAL; MEIDANIS, 1997), alinhar sequencias é inserir *gaps* em posições arbitrárias das sequencias conquistando a mesma medida de tamanho entre elas. Tendo o mesmo tamanho, é possível procurar uma correspondência entre os elementos das sequencias, ou seja,

as colunas se adequarem entre uma sequência e outra para apresentarem o maior número de igualdade entre elas, sem que nenhum *gap* seja alinhado com outro.

As aplicações na Biologia Computacional destes métodos têm posição de destaque em: comparação de sequencias utilizando alinhamento na construção de árvores evolutivas ("arvores filogenéticas"), predição de estrutura secundária de RNA e proteínas. Outras operações na Biologia Computacional têm como base o alinhamento de sequencias segundo (SETUBAL; MEIDANIS, 1997), como por exemplo, busca de similaridades entre sequencias biológicas (SELLERS, 1980), o emparelhamento de *strings* (HALL; DOWLING, 1980), a comparação de arquivos (HUNT; SZYMANSKY, 1977) e a pesquisa em textos com erros (WU;MANBER, 1992).

Cada aplicação necessita de um tipo de alinhamento. Existem dois tipos de alinhamentos: alinhamento global, que leva em consideração sequencias inteiras e alinhamento local, que considera, apenas, fragmentos de sequencias.

No problema: Dada duas sequencias X e Y, compostas pelos itens I = {A, T, G, C}, onde X = {G,A,A,G,G,A,T,T,A,G} e Y = {G,A,T,C,G,A,A,G}, o alinhamento global resultante para elas é apresentado na Tabela 4.1.

Tabela 4.1: Exemplo de alinhamento global entre as sequencias X e Y.

G	**A**	**A**	**-**	**G**	**G**	**A**	**T**	**T**	**A**	**G**
G	**A**	**T**	**C**	**G**	**G**	**A**	**-**	**-**	**A**	**G**

Para o alinhamento global, a obtenção do melhor alinhamento possível entre as sequencias é prioridade, desta forma se utiliza *gaps* o quanto for necessário. As sequencias que são melhores alinhadas são aquelas que já contém uma similaridade antes do alinhamento e/ou que contenham um tamanho aproximado entre si. No caso biológico, o emparelhamento de dois caracteres distintos indica a ocorrência de uma "mutação" em uma das sequencias. Aplicado em outro domínio pode ser um erro qualquer. O emparelhamento entre um caractere de uma sequência e um espaço da outra indica a inserção ou remoção de caracteres em uma das sequencias.

Para o alinhamento local é necessário outro exemplo. Dada duas sequencias X e Y, compostas pelos itens I = {A, T, G, C}, onde X = {A, A, G, A, C, G, G} e Y = {G, A, T, C, G, A, A, G}, o alinhamento local resultante para elas é apresentado na Tabela 4.2.

Tabela 4.2: Exemplo de alinhamento local entre as sequencias X e Y.

					A	**A**	**G**	**A**	**C**	**G**	G
G	**A**	**T**	**C**	**G**	**A**	**A**	**G**	**-**	**A**	**G**	

Como se observa na Tabela 4.2, o alinhamento local ao invés de ocorrer em toda a extensão das sequencias, somente processa uma subsequência das duas sequencias.

Este tipo de alinhamento é utilizado para comparar grandes sequencias com tamanhos diferentes ou que compartilham regiões conservadas entre elas. Para os casos onde as sequencias são pouco relacionadas este tipo favorece que se encontrem padrões entre as sequencias mais conservadas.

O alinhamento de sequência pode ser parametrizado com uma pontuação. Esta pontuação significa obter o melhor caminho para se chegar no alinhamento das sequencias. Para este cálculo se utiliza o recurso esquema de pontuação. A próxima seção é destinada a tratar sobre o assunto esquema de pontuação, que está presente nos modelos de alinhamento apresentados.

4.2.1. Esquema de Pontuação

O esquema de pontuação está presente em qualquer tipo de alinhamento que se utilizar. Este recurso é utilizado normalmente para identificar o alinhamento ótimo entre as sequencias, ou seja, a maior similaridade possível entre elas. A importância do esquema de pontuação é tão relevante durante o processo de alinhamento que influencia o resultado diretamente, uma vez que diferentes esquemas de pontuação levam a alinhamentos diferentes. Outra informação a respeito do esquema de pontuação é o fato da similaridade entre as sequencias ser definida pelo maior valor da pontuação do esquema.

O esquema de pontuação é dado pelas funções (p, g), onde a função p é utilizada para pontuar cada par de caracteres alinhados e g é utilizado para penalizar espaços inseridos, sendo, normalmente, menor que zero $(g < 0)$. Cada alinhamento possível gerado pelo esquema de pontuação é um numérico e o melhor alinhamento é o maior numérico.

Para exemplificar o uso do esquema de pontuação, inicialmente será utilizado um modelo genérico e posteriormente um modelo concreto da Biologia Computacional. Dada duas sequencias *X* e *Y* e o alinhamento *(X* e Y*),* se insere *p (a, b)* toda vez que um caractere *a* de *X** é alinhado com um caractere de *b* de *Y*,* e toda vez que um caractere *a* de *X** ou *b* de *Y** é alinhado com um caractere de espaço, é inserido *g* à pontuação. No final das possibilidades de alinhamento entre as duas sequencias é denotado um *score,* que soma todas as pontuações de alinhamento de *(X*, Y*)*. A similaridade entre duas sequencias é dada por:

$$sim(X,Y) = \max(X^*,Y^*) \in \beta score(X^*,Y^*) \quad (1)$$

onde β é o conjunto de todos os alinhamentos entre *X e Y.*

Na Tabela 4.3 temos um exemplo de alinhamento exposto com o esquema de pontuação. Neste caso, a pontuação foi 0, o que sugere ser um alinhamento não tão ótimo para estas duas sequencias.

Tabela 4.3 Exemplo de cálculo da pontuação entre duas sequencias. Valores utilizados são: 1 – match, -1 para mismatch e -2 para o alinhamento de um caractere com um espaço.

G	A	A	-	G	G	A	T	T	A	G
G	**A**	**T**	**C**	**G**	**G**	**A**	**-**	**-**	**A**	**G**
1	**1**	**-1**	**-2**	**1**	**1**	**1**	**-2**	**-2**	**1**	**1**
Pontuação: 0										

O esquema de pontuação utilizado neste exemplo é:

1. Somar 1 na pontuação quando ocorre um emparelhamento (*match*) entre caracteres das sequencias, ou seja, quando os caracteres são iguais e na mesma posição;
2. Somar -1 na pontuação quando os caracteres das sequencias na mesma posição são diferentes. Isto também é chamado de *mismatch*;
3. Somar -2 na pontuação quando um caractere é alinhado com um caractere de espaço.

Diante disto, na próxima seção são apresentados métodos capazes de realizar o

alinhamento de sequências.

4.3. Algoritmos de Alinhamento de Sequencias

Os tipos de alinhamentos apresentados possuem algoritmos que implementam suas funções, sobretudo o esquema de pontuação, utilizando programação dinâmica. Para alinhamentos locais, os pesquisadores Smith e Waterman em (SMITH;WATERMAN, 1981) criaram seu próprio algoritmo, que foi batizado com seus nomes. O alinhamento global ganhou um algoritmo, utilizando programação dinâmica, uma década antes do alinhamento local. O pesquisador Needleman, unido ao pesquisador Wusch, possibilitaram a obtenção do alinhamento ótimo para o alinhamento global.

Algoritmos utilizando programação dinâmica, segundo (SETUBAL; MEIDANIS, 1997) consistem em resolver uma instância de um problema a partir de instâncias menores. Esta técnica é bastante custosa computacionalmente, utilizando muita memória e processamento, porém a grande vantagem, aplicado no contexto de alinhamento de sequencias, é o fato de possibilitar os alinhamentos ótimos entre duas sequencias (GUSFIELD, 1997).

Ambos algoritmos são excelentes, uma vez que se conhece seu objetivo. Para o nosso problema de identificação de emoções, obter um maior grau de similaridade entre as sequencias com tamanhos semelhantes é um problema a ser resolvido. Por isso, o alinhamento global é o mais indicado com o algoritmo *Needleman-Wusch*.

4.3.1. Algoritmo Needleman–Wunsch

Nesta seção o objetivo é detalhar o algoritmo *Needleman-Wusch* utilizado no método proposto nesta pesquisa. Para ilustrar, se utilizará duas sequencias hipotéticas s e t. Estas sequências possuem comprimento m e n, respectivamente.

Inicialmente constrói-se uma matriz bidimensional A de tamanho $(m + 1)$ e $(n + 1)$, onde cada entrada (i, j) desta matriz contém a similaridade entre $s[1..i]$ e $t[1..j]$. As sequencias s e t são colocadas na margem esquerda e no topo da matriz, respectivamente.

Na matriz, o primeiro valor a ser preenchido é $A[0, 0] = 0$, pois quando se alinha duas sequencias vazias a pontuação é 0. Após isso, se preenche a primeira linha e a primeira coluna, as quais são inicializadas com múltiplos da penalidade por espaço (g). Desta forma, os valores calculados para a primeira linha são $A[0, j] = gj$, para $1 \leq j \leq n$ e para a primeira coluna são $A[i,$

0] = gi, para $1 \leq i \leq m$. Quando uma das sequencias (*s* ou *t*) é vazia, faz com que a pontuação do alinhamento de uma das sequencias com a sequência vazia seja *-2k*, onde *k* é um tamanho da sequência não vazia.

O restante da matriz é preenchido de cima para baixo e da esquerda para a direita. Salienta-se que, para obter-se o valor *(i, j)* da matriz, três valores são anteriormente calculados: (*i -1, j*), (*i -1, j -1*) e (*i, j - 1*). Existem, então, três possibilidades para obtenção do alinhamento entre *s* e *t* , sendo elas:

- *Alinhar s[1..i] com t[1..j - 1] e um espaço com t[j];*
- *Alinhar s[1..i - 1] com t[1..j -1] e s[i] com t[j];*
- *Alinhar s[1..i – 1] com t[1..j] e s[i] com um espaço.*

Para cada possibilidade está associada uma determinada pontuação e, fazendo *p (i, j),* ser o valor associado a um *match* ou *mismatch* e *g* o valor atribuído a um *gap,* estas possibilidades podem ser representadas pela equação 2.

$$A[i,j] = max \left\{ \begin{array}{c} A[i, j-1] + g \\ A[i-1, j-1] + p(i,j) \\ A[i-1, j] + g \end{array} \right\} \quad (2)$$

Esta etapa do algoritmo dá-se por encerrada quando todos os elementos da matriz foram calculados. Cada elemento da matriz, mantem um apontador direcionado para o elemento que o derivou (*[i, j- 1], [i- 1, j- 1] ou [i- 1, j]*).

A segunda parte, chamada *traceback,* tem o objetivo de encontrar o alinhamento propriamente dito entre as duas sequências. Inicia-se do elemento *(i, j)* da matriz, percorrem-se os apontadores até o elemento *(0, 0)* ser encontrado. Nesta operação três alternativas são possíveis:

- Se o apontador indicar que o elemento *(i, j)* derivou do elemento da esquerda *([i, j- 1]),* então um *gap* é inserido na sequência *s*;
- Se o apontador indicar que o elemento *(i, j)* derivou do elemento superior *([i- 1, j]),* então um *gap* é inserido na sequência *t*;
- Se o apontador indicar que o elemento *(i, j)* derivou do elemento da diagonal superior esquerda *([i- 1, j- 1])*, então nenhum *gap* é inserido, indicando que os elementos de *s* e *t* que estão sendo analisados, são iguais.

Desta forma, após o *traceback*, se obtém o maior valor possível para o alinhamento global.

4.4. Conclusão

Neste capítulo foram abordados conceitos fundamentais que orientam este trabalho. Na seção 4.1 foram apresentadas as formas com a biologia e os padrões se integram, o que é importante para melhor compreensão da técnica de extração de padrões deste método. Na seção 4.2, foi tratado sobre o alinhamento de sequencias e os tipos de alinhamentos possíveis para identificar similaridades entre sequencias. Na seção 4.3 foram apresentados algoritmos para alinhamentos globais e locais. O algoritmo *Needleman-Wunsch* é o algoritmo utilizado inicialmente neste projeto.

Capítulo 5

Estado da Arte

O objetivo principal deste capítulo é apresentar os trabalhos encontrados publicados na literatura sobre Análise de Sentimentos, desenvolvidos para textos, e que eventualmente tenham utilizado MPS, que identifiquem categorias de emoções em textos.

Dessa forma, a primeira seção deste capítulo trata deste assunto. A segunda seção é destinada à apresentação de trabalhos de Análise de Sentimentos em textos desenvolvidos para diversos idiomas que estejam relacionados, mesmo que indiretamente, com o método de identificação de emoções que está sendo proposto neste trabalho e a última seção trata de trabalhos que utilizaram algoritmos de MPS de forma relacionada com esta pesquisa.

A seleção dos trabalhos apresentados neste capítulo foi realizada por meio do acesso a diversos repositórios de pesquisa, dentre os quais estão: *SciencDirect*, *IEEExplore*, *CiteSeer*, *ACM Digital Library* e *Google Scholar*. Para dar início à busca foram utilizadas palavras-chave como: *Sentiment Analysis*, *Pattern Mining*, *Text Mining*, *Opinion Mining*, *Effect Mining, Análise de Sentimentos, Mineração de padrões sequenciais*. Outros trabalhos foram obtidos por meio da análise de citações e referências contidas em artigos, teses, livros e relatórios técnicos.

5.1. Identificação de Emoções em Textos e Mineração de Padrões Sequenciais

A pesquisa em identificação de emoções em textos é um tema atual na área de AS, porém poucos são os trabalhos desenvolvidos especificamente utilizando técnicas de MPS. Neste contexto, o trabalho de (AHMAD, 2013), propõe um framework que utiliza algoritmos de MPS, aplicados na AS.

No Trabalho de (AHMAD, 2013), o problema a ser resolvido está vinculado à opinião dos clientes com relação a um determinado produto. Segundo (AHMAD, 2013), a maioria das vezes, as avaliações dos produtos são tão longas que é impossível ler. Nestes casos, os clientes,

como os fabricantes, não podem ter uma informação precisa com relação à opinião dos clientes, uma vez que os clientes podem gravitar em torno dos comentários e ler somente alguns para formar sua opinião.

A abordagem implementada utiliza técnicas para reunir as informações extraídas de várias fontes de dados, em uma estrutura semântica e fornecer uma ferramenta de visualização que pode ajudar os usuários em vários níveis de complexidade. O objetivo foi desenvolver um projeto com uma nova forma de Identificação de Candidato e Geração de Padrões de Sequência (CI-FPG) para extrair as características frequentes dos documentos. A CI-FPG utiliza dois passos para aplicar este conceito: Na primeira etapa, ele usa o Stanford Parser, e gera uma árvore de dependência, além de utilizar sistemas de linguística e semântica na análise de texto para identificar as características. Na segunda etapa, utiliza o algoritmo de crescimento FP-Growth para gerar padrões de sequência a partir das características extraídas.

Os experimentos foram realizados com uma base formada por comentários de clientes sobre dois produtos: câmeras e hotéis. No total são 1260 registros sobre os produtos, sendo 1025 sobre câmeras e 235 sobre hotéis, onde, após avaliações foram extraídas 10 características de sobre cada produto.

Para cada documento, se uma característica apareceu nele o numeral 1 foi computado, para o contrário foi computado 0. Uma lista parcial de documentos, juntamente com as características extraídas de tal documento é apresentada na Tabela 5.1.

Tabela 5.1 Lista de documentos e caracteristicas extraídas. Adaptado de (AHMAD, 2013).

Nro. Documento	Característica Extraída
Documento 1	F1, F2, F3
Documento 2	F2, F3
Documento 3	F1, F3, F5
Documento 4	F2, F5, F7
Documento 5	F4, F6, F7, F8
Documento 6	F2, F3, F5
Documento 7	F1, F2, F5
Documento 8	F2, F4, F5
Documento 9	F1, F2, F9
Documento 10	F2, F3, F4, F10

Para encontrar as características frequentes foi utilizado o FP- Growth por causa de sua vantagem sobre o Apriori. Na Tabela 5.2 é apresentado os padrões encontrados com o método.

Tabela 5.2: Padrões encontrados. Adaptado de (AHMAD, 2013).

Nro. Característica	Característica
F1	{F1}, {F2, F1}, {F3, F1}, {F5, F1}
F2	{F2}, {F3, F2}, {F4, F2}, {F5, F2}
F3	{F3}, {F5, F3}
F4	{F4}
F5	{F5}
F6	{}
F7	{F7}
F8	{}
F9	{}
F10	{}

Outro trabalho (ZHANG; JIA; ZHU; ZHOU; HAN, 2014), também utiliza algoritmos de MFS. Neste trabalho, se estuda o problema da evolução sentimento em um microblog. A análise da evolução sentimento, por vezes, contém informação mais valiosa do que a orientação estática que trata único sentimento em um texto. Os comerciantes podem usar isso para investigar a opinião pública da sua empresa e produtos. Governos também podem usar isso para obter um feedback crítico sobre problemas na política recém-lançados ou para monitorar a opinião pública sobre emergência, de modo a prever as tendências de desenvolvimento de eventos.

Para realizar plenamente análise da evolução sentimento, é necessário construir um modelo mais preciso e específico em vez de usando o modelo de métrica ternária tradicional (positivo, neutro e negativo).

Segundo os autores, a pesquisa apresentou um método eficaz, mas ainda simples, para resolver o problema. Em primeiro lugar, construiu um modelo multidimensional com estrutura e conceito de hierarquia para representar usuários e complicar seus sentimentos. Com base neste modelo, foram extraídos os sentimentos agregados das mensagens dos usuários, de modo a detectar padrões de frequência em sequencias de sentimento e executar análise de evolução. Os resultados experimentais sobre um conjunto que a abordagem pode resolver o problema evolução sentimento em um grande conjunto de dados e o modelo multidimensional pode efetivamente refletem as tendências de evolução.

Na Tabela 5.3, foram compilados exemplos de documentos de usuários onde foram encontrados sentimentos.

Tabela 5.3 Documentos de usuários onde foram encontrados sentimentos. Adaptado de (ZHANG; JIA; ZHU; ZHOU; HAN, 2014)

Doc. Id.	Usuário	Hora	Conteúdo	Sentimento
1	u_1	9:02	d_1	e_1
2	u_1	10:01	d_2	e_1, e_2
3	u_2	11:01	d_3	e_2
4	u_3	20:15	d_4	e_1, e_2, e_3
5	u_2	20:23	d_5	e_3, e_5
6	u_3	20:30	d_6	e_6
7	u_4	21:15	d_7	e_1, e_2
8	u_1	22:10	d_8	e_3, e_4
9	u_4	8:10	d_9	e_4
10	u_1	10:01	d_{10}	e_4

Para realizar a análise de sentimento, primeiro se extrai o sentimento para cada mensagem, e se agrega o vector sentimento para cada usuário. Em seguida, utiliza-se FPGrowth para encontrar os padrões de frequência de sentimentos (FCSP). Para cada FCPS, realizamos análise da evolução de sentimento de acordo com Kullback-Leibler. Para cada evolução do FCSP, usamos algoritmo de agrupamento Affinity Propagation para detectar a razão pela qual usuário mudou suas atitudes. O algoritmo FP-Growth foi usado para minerar todas as possibilidades de padrões de frequência de sentimentos de todos os usuários.

5.2. Identificação de Emoções em Textos

Dos trabalhos relacionados a esta pesquisa, que trabalham com textos identificando emoções, podemos destacar o trabalho de Dosciatti (DOSCIATTI, 2015), onde um método foi construído especialmente para textos escritos em Português Brasileiro. Neste método se utiliza uma abordagem de AM supervisionada para identificar textos. Outras características deste método são: completamente independente de recursos léxicos de emoções como dicionários, ontologias e thesaurus. O método opera basicamente em duas camadas: 1) a primeira camada usa um classificador SVM multiclasse e o conceito de rejeição para rejeitar os textos mais

complexos de serem classificados; 2) os textos rejeitados são encaminhados para serem classificados pelos classificadores binários. O método foi submetido a uma avaliação por meio de um corpus de notícias, anotado para classificar sentimentos em texto, em português. Ao ser avaliado com o corpus de notícias, o método obteve uma taxa de acerto de 65,5% ao identificar as seis emoções básicas, além de neutro, em textos. Ainda com o mesmo corpus, obteve 93% na taxa de acerto ao identificar a polaridade das emoções nos textos.

Outro trabalho que trata especificamente a classificação de emoções em textos na língua portuguesa é o de Nascimento e colegas (NASCIMENTO et al., 2012). Este método também não utiliza léxicos. Nesta pesquisa o autor visa avaliar a reação das pessoas em relação às notícias compartilhadas na mídia por meio da análise de publicações feitas no *Twitter.* As notícias escolhidas para este estudo estão inseridas nas categorias policiais, política e entretenimento. Para avaliar o desempenho do método, foi construído um corpus composto de 850 documentos, sendo 425 positivos e 425 negativos. Os textos foram anotados por três avaliadores.

Neste trabalho o objetivo foi concluir, a partir das emoções expressas nos *tweets*, se a população tem a opinião sobre um fato entre positivo ou negativo. Assim, o trabalho mostra um experimento de classificação supervisionada onde foram avaliados dois modelos estatísticos baseados em n-grama (*unigrama e octagrama*) e o classificador estatístico *Naive Bayes*.

A ferramenta apresentada em (EVANGELISTA; PADILHA, 2014), é desenvolvida para classificar comentários postados em redes sociais como positivo, negativo e neutro. A ferramenta faz uso do léxico SentiWordNet (ESULI; SEBASTIANI, 2006) e do algoritmo *Naive Bayes* para classificar os textos.

Em (MALHEIROS, 2014) foi usado um conjunto de textos em Inglês (GO; BHAYANI; HUANG, 2009) por onde foi avaliado o desempenho de uma ferramenta que identifica a polaridade das emoções em textos de *Twitter* usando uma abordagem léxica e de AM. O conjunto de textos, composto por 800.000 *tweets* positivos e 800.000 *tweets* negativos, foram destinados ao treinamento e um conjunto de teste, composto de 177 *tweets* negativos e 182 *tweets* positivos, foi coletado e anotado manualmente.

Os pesquisadores (GHAZI; INKPEN; SZPAKOWICZ, 2010) utilizam uma abordagem de AM para construir um método de identificação de emoções que considera níveis de classificação, sendo que as categorias são parcialmente ordenadas, da mais genérica para mais específica. Foram utilizados dois conjuntos de dados textuais para avaliar o método

desenvolvido neste trabalho. O primeiro conjunto é composto 2.090 sentenças de blogs e faz parte do corpus desenvolvido por (AMAN; SZPAKOWICZ, 2007) e o segundo, é composto de 1.207 sentenças de contos infantis e faz parte do corpus desenvolvido por (ALM; ROTH; SPROAT, 2005). Daremos destaque a este trabalho nos próximos parágrafos visto que utilizaremos a mesma base de dados em nossos experimentos.

Os resultados são mostrados na Tabela 5.4, que foram obtidos com a classificação em níveis e com a classificação plana para o conjunto de dados de blogs. No primeiro nível, a classificação determina se uma instância é Emocional ou Não Emocional. No segundo nível considera todas as instâncias que foram classificadas como emocional no primeiro nível e as classifica em uma das seis emoções básicas alegria, tristeza, raiva, medo, repugnância e surpresa.

Tabela 5.4 Dois níveis de classificação no conjunto de dados de (AMAN; SZPAKOWICZ, 2007). Fonte: Adaptado de (GHAZI; INKPEN; SZPAKOWICZ, 2010)

Nível	**Categoria**	**Dois níveis de classificação**			**Dois níveis de classificação**		
		Precisão	**Precisão**	**Cobertura**	**F1**	**Cobertura**	**F1**
1º nível	Emocional	0,88	0,85	0,86	---	---	---
	Não Emocional	0,88	0,81	0,84	0,54	0,87	0,67
2º nível	Alegria	0,59	0,95	0,71	0,74	0,60	0,66
	Tristeza	0,77	0,49	0,60	0,69	0,42	0,52
	Medo	0,91	0,49	0,63	0,82	0,49	0,62
	Surpresa	0,75	0,32	0,45	0,64	0,27	0,38
	Repugnância	0,66	0,35	0,45	0,68	0,31	0,43
	Raiva	0,72	0,33	0,46	0,67	0,26	0,38
Acurácia		**68,3%**			**61,7%**		

Na Tabela 5.5 é apresentado o resultado da classificação de sete classes, realizada em três níveis. No primeiro nível classifica se a instância é Emocional ou Não Emocional. No segundo nível, as instâncias classificadas como emocionais no primeiro nível são definidas em sua polaridade. O terceiro nível, tem como premissa que as instâncias da classe alegria têm polaridade positiva e *tristeza*, *raiva*, *repugnância*, *medo* e *surpresa* têm polaridade negativa. Assim, as instâncias negativas do segundo nível são classificadas em cinco classes de emoção no terceiro nível.

Tabela 5.5 Três níveis de classificação no conjunto de dados de (AMAN; SZPAKOWICZ, 2007). Fonte: Adaptado de (GHAZI; INKPEN; SZPAKOWICZ, 2010)

Nível	**Categoria**	**Três níveis de classificação**		
		Precisão	**Cobertura**	**F1**
1º nível	Emocional	0,88	0,85	0,86
	Não Emocional	0,88	0,81	0,84
2º nível	Positivo	0,89	0,65	0,75
	Negativo	0,79	0,94	0,86
3º nível	Tristeza	0,63	0,54	0,59
	Medo	0,88	0,52	0,65
	Surpresa	0,79	0,37	0,50
	Repugnância	0,42	0,38	0,40
	Raiva	0,38	0,71	0,49
Acurácia		**65,5%**		

Na Tabela 5.6 são apresentados os resultados obtidos no experimento executado com o conjunto de dados de contos infantis de (ALM; ROTH; SPROAT, 2005). O experimento, se embasou no pressuposto de que a classe alegria é positiva e as quatro classes restantes são negativas. O primeiro nível, determina se uma instância possui uma polaridade positiva ou negativa. O segundo nível considera todas as instâncias que foram classificadas como negativa e são então classificadas em uma das quatro classes negativa: *tristeza*, *medo*, *surpresa* e *raiva*.

Tabela 5.6 Dois níveis de classificação no conjunto de dados de (ALM; ROTH; SPROAT, 2005). Fonte: Adaptado de (GHAZI; INKPEN; SZPAKOWICZ, 2010)

Nível	**Categoria**	**Dois níveis de classificação**			**Dois níveis de classificação**		
		Precisão	**Precisão**	**Cobertura**	**F1**	**Cobertura**	**F1**
1º nível	Negativo	0,81	0,93	0,87	---	---	---
	Positivo	0,84	0,64	0,72	0,56	0,86	0,68
	Tristeza	0,65	0,68	0,66	0,67	0,53	0,59
	Medo	0,59	0,40	0,47	0,59	0,38	0,46
	Surpresa	0,45	0,21	0,29	0,35	0,10	0,16
	Raiva	0,49	0,73	0,59	0,54	0,43	0,48
Acurácia		**59,1%**			**57,4%**		

Com os dois experimentos com classificação plana, mostram que em ambos os casos a Precisão da abordagem de dois níveis é significativamente melhor do que a Precisão da classificação plana. Um aspecto interessante dos resultados na Tabela 5.2 é a Precisão da classe Não Emocional. Esta classe aumenta a precisão enquanto a Cobertura diminui. Esta relação também ocorre em outras experiências e acontece com as classes que normalmente dominavam na classificação plana, mas que não dominam mais na classificação em níveis. O comportamento dos classificadores é de observar que tendem a dar prioridade a uma classe dominante, de modo que mais instâncias são classificadas nesta classe, dessa forma a classificação alcança uma baixa Precisão e uma alta Cobertura. No trabalho de (GHAZI; INKPEN; SZPAKOWICZ, 2010) os experimentos demostram que na classificação em níveis tende a produzir maior valor de Precisão e F1 do que a classificação plana.

O trabalho de (TURKMENOGLU; TANTUG, 2014), desenvolveu dois métodos de AS para o idioma Turco. Um baseado em léxico e outro baseado em AM supervisionada, para identificar a polaridade das emoções em dois corpora diferentes, textos de Twitter e textos de comentários de filmes. Neste trabalho o corpus é de *tweets* e é composto de 5.900 *tweets*, extraídos de seis páginas de marcas populares e anotados manualmente com as categorias *positivo*, *negativo* e *neutro*. O número médio em cada documento é de 14 palavras. No *corpus* de comentários de filmes é composto por 20.244 textos. O número médio de 39 palavras por documento. A fonte destes textos é um site popular que permite aos usuários emitir comentários sobre os filmes e selecionar uma categoria de uma a cinco estrelas.

Assim, os comentários que possuem uma classificação superior a quatro estrelas são considerados como positivos e os que possuem uma classificação inferior a 2,5 são negativos, e os textos pertencentes à escala restante são descartados.

Devido a inexistência de léxico de palavras afetivas para o idioma Turco, os autores traduziram um léxico básico do Inglês, composto por 2.547 palavras, para o idioma Turco. E ainda, houve um árduo trabalho linguístico no pré-processamento para lidar com as características da língua Turca no método baseado em léxico.

A abordagem de AM utilizada é a unigramas e bigramas com um ranqueamento dos pesos *TF-IDF (Term Frequency - Inverse Document Frequency)* para selecionar as melhores características. Os experimentos foram executados com os classificadores *Suporte Vector Machine*, *Naive Bayes* e *J48*, que foram avaliados por meio de Validação Cruzada com dez

partes. Na Tabela 5.7 são apresentados os melhores resultados obtidos com os experimentos mostrados no trabalho.

Tabela 5.7: Melhor acurácia obtida em cada corpus. Fonte: Adaptada de (TURKMENOGLU; TANTUG, 2014)

Twitter				**Filmes**			
SVM	**NB**	**J48**	**Léxico**	**SVM**	**NB**	**J48**	**Léxico**
85,0%	84,3%	81,0%	75,2%	89,5%	89,5%	83,0%	79,0%

Os resultados mostrados na Tabela 5.7 é possível observar que, embora a diferença de acurácia entre os classificadores não seja tão significativa, o SVM obteve a maior taxa de acerto no geral. Outro aspecto que se observa é que a abordagem de AM superou a abordagem léxica. O trabalho não fornece outras métricas de desempenho por classe, o que permitiria uma análise mais completa.

5.3. Trabalhos relacionados com FP-Align.

Alguns trabalhos são importantes para o estado da arte mesmo não relacionados à AS. Um exemplo é o trabalho (KATOH, STANDLEY, 2016) que apresenta um novo recurso do programa de alinhamento MAFFT. Neste caso o MAFFT faz alinhamentos múltiplos entre as sequências, ou seja, alinha várias sequências entre si (WALLACE, 2005). Convencional, MAFFT é altamente sensível no alinhamento de regiões, chamadas conservadas em homólogos remotos, mas o risco de *over-alignment* (alinhando segmentos não relacionados) (RC; SJOLANDER, 2004) foi recentemente tornando-se maior, como de baixa qualidade ou sequências ruidosas que estão aumentando em bases de dados de sequências de proteínas, devido, por exemplo, a erros de sequenciamento e dificuldades de previsão de genes (RC; SJOLANDER, 2004). O novo recurso procura suprimir o *over-aligment*.

Na pesquisa de (KATOH, STANDLEY, 2016) utiliza o alinhamento de sequências de programação dinâmica (NEEDLEMAN; WUNSCH, 1970). Outras ferramentas, por exemplo BLOSUM (HENIKOFF; HENIKOFF, 1992), GCB (GONNET et al., 1992), PAM (DAYHOFF et al., 1978), JTT (JONES et al., 1992), também utilizam (NEEDLEMAN; WUNSCH, 1970). MAFFT usa BLOSUM62 por padrão. Programação Dinâmica dá o ótimo alinhamento de duas sequências, maximizando a pontuação de alinhamento, definido como a soma das pontuações de pares alinhados e custos de gap.

Outro trabalho que utiliza o algoritmo *Needleman-Wusch* como base é (MUHAMAD; AHMAD; MURAD, 2015), onde os algoritmos propostos e avaliados têm como objetivo reduzir as lacunas no alinhamento de sequências, assim como o comprimento das sequências alinhadas sem comprometer a qualidade ou exatidão dos resultados.

5.4. Conclusão

Neste capítulo foram apresentados trabalhos relacionados com a pesquisa que está sendo proposta. Na primeira seção, buscou-se por trabalhos de MPS, aplicados no contexto de analise de sentimentos. Entretanto, a maioria dos trabalhos encontrados na literatura são utilizam algoritmos derivados do *Apriori.* Na segunda seção, foi realizada uma pesquisa de artigos AS produzidos para identificar categorias de emoções em texto e na terceira seção, buscou-se trabalhos relacionados ao algoritmo *Needleman-Wusch*, aplicamos no contexto das Ciências Computacionais.

Capítulo 6

Procedimentos Metodológicos

Neste capítulo são detalhados os procedimentos metodológicos para a construção e avaliação do método. No capítulo ainda se define a forma de trabalho e a estratégia, assim como as etapas utilizadas pelo pesquisador para estruturar seu trabalho e alcançar os objetivos previamente definidos.

6.1 Caracterização da Pesquisa

Com relação aos objetivos, a pesquisa foi classificada como experimental, pois realiza análises das situações experimentais que são flexíveis no sentido de que muitos e variados aspectos da teoria podem ser testados (KERLINGER, 1980). Quanto ao procedimento técnico, nenhuma das classificações previstas em Gil (2002) foi considerada inteiramente adequada a este trabalho. Desta maneira, esta pesquisa foi estruturada combinando diversos procedimentos técnicos, tais como, pesquisa exploratória e experimento, os quais serão detalhados de acordo com a etapa da pesquisa apresentada.

6.2 Estratégia da pesquisa

Como estratégia desta pesquisa, foi definida uma estrutura inspirada em (VALASCKI, 2017) própria combinando vários métodos e procedimentos técnicos para atingir os objetivos iniciais. Como pode ser visto na Figura 6.1 esta estrutura é dividida em cinco etapas principais: Exploração do objeto de pesquisa, Avaliação sobre o uso de MPS na identificação de emoções,

Proposta do método de identificação de emoções utilizando MPS, Desenvolvimento do método proposto e Avaliação do método proposto. Todas as etapas serão detalhadas a seguir.

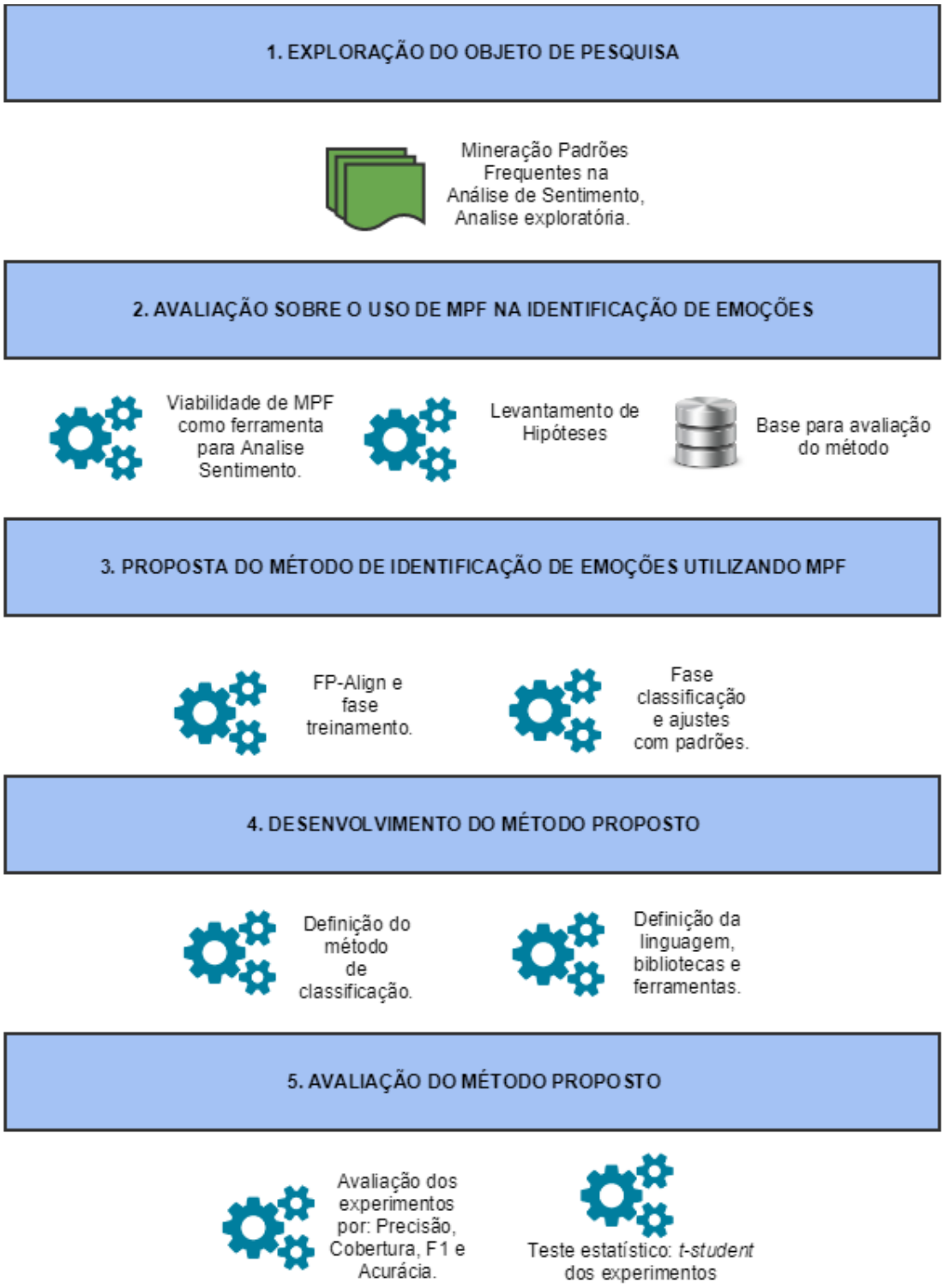

Figura 6.1: Estrutura de pesquisa. (fonte: autor).

6.2.1 Exploração do objeto de pesquisa

Para Clemente e colegas (CLEMENTE et al, 2007) uma pesquisa pode ser considerada de cunho exploratória, quando esta envolver levantamento bibliográfico, entrevistas com pessoas que tiveram, ou têm, experiências práticas com o problema pesquisado e análise de exemplos que estimulem a compreensão. As pesquisas exploratórias visam proporcionar uma visão geral de um determinado fato, do tipo aproximativo.

Nesta etapa foi realizada uma pesquisa exploratória nas seguintes bases: *SciencDirect*, *IEEExplore*, *CiteSeer*, *ACM Digital Library* e *Google Scholar*. Para dar início à busca de artigos nestas bases foram utilizadas palavras-chave como: *Sentiment Analysis*, Sequence *Pattern Mining*, *Text Mining*, *Opinion Mining*, *Effect Mining, Análise de Sentimentos, Mineração de padrões sequenciais*. Outros trabalhos foram obtidos por meio da análise de citações e referências contidas em artigos, teses, livros e relatórios técnicos.

Os artigos foram distribuídos em três categorias: Identificação de Emoções em Textos e Mineração de Padrões Sequenciais, Identificação de Emoções em Textos, Trabalhos de Mineração de Padrões Sequenciais. Estas categorias representam a escala de aproximação dos artigos com esta pesquisa, sendo que na primeira categoria estão os trabalhos mais relacionados e na última categoria os menos relacionados.

6.2.2 Avaliação sobre o uso de MPS na Identificação de Emoções

Foi possível observar na pesquisa exploratória que a utilização de MPS na identificação de emoções ainda não é suficientemente explorada. No âmbito desta pesquisa, foi possível observar que a natureza dos padrões sequenciais encontrados nos artigos é diferente dos propostos nesta pesquisa. No geral os artigos buscam padrões textuais como: palavras que aparecem sempre juntas, características que aparecem sempre juntas, etc.

As hipóteses de trabalho apresentadas no Capítulo 1 foram construídas em função do que foi encontrado na pesquisa exploratória. Para fortalecer a viabilidade da pesquisa, foi encontrada também uma base de contos infantis com a característica de rotulações de textos cronológica que posteriormente foi usada para avaliar o método.

Em (ALM; ROTH; SPROAT, 2005), um *corpus* com 15302 textos extraídos de 185

contos infantis, foi anotado em um nível de sentença por dois anotadores. Cada texto foi anotado com as seguintes emoções: *raiva, repugnância, medo, alegria, tristeza, surpresa positiva* ou *surpresa negativa*. O grau de concordância Kappa entre os anotadores neste *corpus* ficou entre 0,24 e 0,51. O *corpus* é em inglês e na Tabela 6.1 é possível observar a distribuição dos textos por autor e emoção.

Tabela 6.1: Distribuição dos textos do corpus (ALM; ROTH; SPROAT, 2005) por emoção e autor.

Emoção	Potter	HC Andersen	Grimms	TOTAL
Angry (A)	87	166	477	730
Disgusted (D)	40	272	151	463
Fearful (F)	102	151	444	697
Happy (H)	119	829	662	1610
Neutral (N)	1440	5832	2867	10139
Sad (Sa)	45	390	396	831
Positive Surprised (Su+)	29	137	163	329
Negative Surprised (Su-)	84	219	200	503
TOTAL	1946	7996	5360	15302

O *corpus* foi usado no trabalho de (GHAZI; INKPEN; SZPAKOWICZ, 2010), criando um subconjunto do desenvolvido por (ALM, 2008). O subconjunto é composto por 169 contos e foi anotado por dois anotadores com as emoções básicas de (EKMAN, 1992).

O *corpus* usado no trabalho de (GHAZI; INKPEN; SZPAKOWICZ, 2010), e nos experimentos deste método, é composto por textos que tiveram total concordância pelos dois anotadores durante o processo de anotação. O *corpus* contém 1.207 sentenças, sendo divididas em 169 documentos, que foram anotadas com cinco emoções, sendo elas: *alegria*, *tristeza*, *raiva-repugnância*, *medo* e *surpresa*. O *corpus* foi originalmente anotado com as emoções básicas de (EKMAN, 1992) porém, na versão usada pelos autores (GHAZI; INKPEN; SZPAKOWICZ, 2010), as classes *raiva* e *repugnância* foram unidas na classe *raiva-repugnância.*

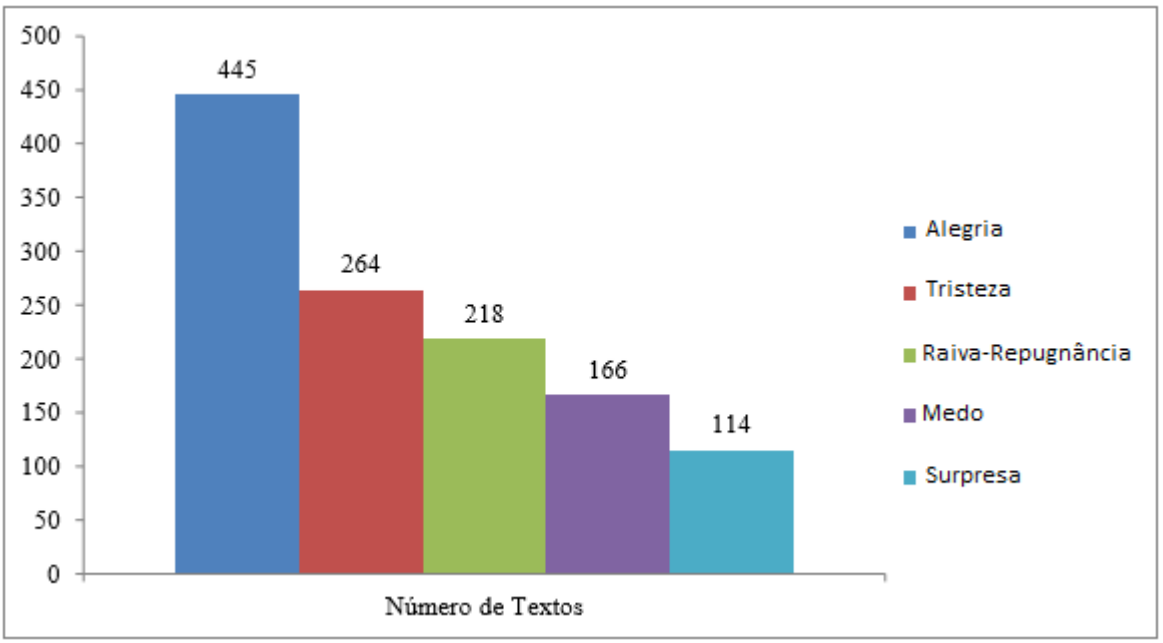

Figura 6.2: Distribuição dos textos do subconjunto do corpus de contos infantis por classes.

Este *corpus* é composto por contos infantis de três autores diferentes: *Grimms, H.C. Andersen e Potter.* Os contos foram misturados para compor um único *corpus*. Este processo é importante, pois procura ir ao encontro com um dos objetivos do trabalho que é identificar padrões que se repetem na rotulação de todos os textos. Seria possível identificar padrões de um único autor e classificar se o conto infantil é ou não é daquele autor, porém não está no foco deste trabalho.

6.2.3 Proposta do método de identificação de emoções utilizando MPS

Ao ser realizada a pesquisa exploratória de bibliografia foi constatado que na área de identificações de emoções em texto os pesquisadores não utilizam as rotulações do *corpus* como transação para a classificação. Nesta constatação reafirmou a motivação deste trabalho que é de identificar padrões sequenciais em rotulações de uma base escrita cronologicamente e reutilizá-los na fase de testes para ajustar confusões de classificação.

Na fase de treinamento foi desenvolvido um algoritmo chamado FP-Align que utiliza conceitos da biologia computacional para a extração de padrões.

Na fase de classificação, o método desenvolvido por (DOSCIATTI, 2015) foi escolhido para realizar a classificação dos textos em uma primeira etapa. O método escolhido para a etapa de classificação é o proposto por (DOSCIATTI, 2015), por ser capaz de identificar

as emoções básicas. O método (DOSCIATTI, 2015) também exporta as Probabilidades Estimadas (*PEs*) calculadas durante o processo de classificação para cada instancia e utiliza em suas rotinas o classificador SVM. O SVM foi projetado originalmente para a solução de problemas de classificação contendo apenas duas classes, entretanto, muitos problemas de classificação são multi-classe. Este fato não inviabiliza o uso do SVM e algumas técnicas são propostas para estendê-lo para essa finalidade. É esperado que ao final desta fase a classificação inicial realizada por (DOSCIATTI, 2015) tenha sido modificada e melhorada com os padrões encontrados na fase de treinamento.

6.2.4. Desenvolvimento do método proposto

O método foi implementado na linguagem *Python,* na versão 2.7. O algoritmo de *MPS* foi implementado pelo próprio autor e validado com dados disponibilizados na *Web: http://pt.slideshare.net/mcastrosouza/algoritmo-needlemanwunsch*.

Para que o método pudesse ser testado e avaliado, foi construída uma ferramenta de Análise de Sentimentos, com um ambiente configurável via arquivo e sua execução através de *scripts.*

O método de (DOSCIATTI, 2015) foi desenvolvido utilizando a plataforma de desenvolvimento *Intellij IDEA* e o algoritmo *SVM*, implementado por (CHANG; LIN, 2011) e incorporado ao *Weka* (HALL et al., 2009) por (EL-MANZALAWY; HONAVAR, 2005). Para os experimentos desta pesquisa o método (DOSCIATTI, 2015) foi executado utilizando método de testes *holdout,* considerando que 70% da base de dados foi para treinamento e 30% para testes.

6.2.5. Avaliação do método proposto

Uma forma de verificar o desempenho de um classificador é o uso dos dados de uma matriz de confusão. A matriz de confusão ilustra o número de classificações corretas e incorretas de cada classe. As linhas da matriz representam as classes verdadeiras e as colunas, as classes preditas pelo classificador. A diagonal principal representa os acertos do classificador, enquanto os outros elementos correspondem aos erros cometidos nas suas predições. Por meio da matriz de confusão, tem-se medidas quantitativas de quais classes o algoritmo de aprendizado tem

maior dificuldade em acertar (FACELI et al., 2011). Para um problema binário, onde usualmente uma classe é chamada de positiva e a outra de negativa, tem-se a matriz de confusão ilustrada na Tabela 6.2.

Tabela 6.2: Matriz de confusão para um problema binário. Fonte: (KOHAVI; PROVOST, 1998)

		Classe Predita	
		Positiva	Negativa
Classe Real	Positiva	VP	FN
	Negativa	FP	VN

- *VP* (Verdadeiro Positivo) é o número de exemplos da classe positiva classificados corretamente.
- *VN* (Verdadeiro Negativo) é o número de exemplos da classe negativa classificados corretamente.
- *FP* (Falso Positivo) é o número de exemplos cuja classe verdadeira é negativa, mas que foram classificados incorretamente como sendo da classe positiva.
- *FN* (Falso Negativo) é o número de exemplos cuja classe verdadeira é positiva, mas que foram classificados incorretamente como sendo da classe negativa.

A partir da matriz de confusão, uma série de medidas quantitativas de desempenho pode ser derivada, dentre elas: Precisão, Cobertura e F1, cujas equações estão sendo apresentadas neste capítulo.

Taxa de acerto ou acurácia: é o número de classificações corretas, ou seja, a soma dos verdadeiros positivos e verdadeiros negativos, dividido pelo número total de classificações (Equação 3).

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} \tag{3}$$

Precisão: é o número de verdadeiros positivos dividido pela soma dos verdadeiros positivos e dos falsos positivos (Equação 4)

$$Precisão = \frac{VP}{VP + FP} \tag{4}$$

Cobertura: é o número de verdadeiros positivos, dividido pelos verdadeiros positivos e falsos negativos (Equação 5)

$$Cobertura = \frac{VP}{VP + FN} \tag{5}$$

F1: é a média harmônica das medidas de Precisão e de Cobertura (Equação 6):

$$F1 = \frac{2.Precisão.Cobertura}{Precisão + Cobertura} \tag{6}$$

Quando se trata de um problema de classificação multi-classe, é importante observar as medidas de desempenho por classe e a classificação global que geralmente avaliada é por meio de média ponderada calculada em função do número de amostras de cada classe.

Para as análises estatísticas dos resultados foi utilizado o teste *t* de *Student*. O teste *t* pode ser utilizado para avaliar se há diferença significativa entre as médias de duas amostras. O teste também é caracterizado por ser um teste de hipótese que usa conceitos estatísticos para rejeitar ou não uma hipótese nula.

O teste *t* é uma distribuição de probabilidade teórica. É simétrica, campaniforme, e semelhante à curva normal padrão, porém com caudas mais largas, ou seja, uma simulação da *t* de *Student* pode gerar valores mais extremos que uma simulação da normal. O único parâmetro *v* que a define e caracteriza a sua forma é o número de *graus de liberdade*. Quanto maior for esse parâmetro, mais próxima da normal ela será.

No teste *t* existem pressupostos para as amostras que devem ser respeitados:

- **As duas amostras devem ter uma distribuição normal**. Para garantir esta

distribuição foi realizado o teste estatístico de Shapiro-Wilk (SHARIPO; WILK, 1965) e análise de histograma das amostras.

- **As duas amostras devem ter a mesma variância**. Neste caso, para garantir este pressuposto foi utilizado o teste *F* de Jhonson e colegas (JHONSON et al, 1995) onde é possível observar o desvio padrão amostral das amostras.
- **As duas amostras devem ser independentes.**

Os termos unicaudal e bicaudal são termos ligados ao teste de hipótese nula e alternativa. É comum pesquisadores definirem a hipótese nula como sendo médias iguais para as duas amostras. Caso a hipótese alternativa seja a média de uma amostra maior que a média da outra amostra, então temos um teste unicaudal. Caso a hipótese alternativa seja apenas médias diferentes, se tem um teste bicaudal.

O resultado *p-value* é a saída do teste t, onde é verificado se a hipótese nula é rejeitada ou não. Na Tabela 6.3 é possível identificar as faixas de valores do p-value e sua respectiva interpretação.

Tabela 6.3: Faixas e interpretações do p-value.

p-value	**Interpretação**
P < 0.01	Evidencia muito forte contra a hipótese nula
0.01 <= P < 0.05	Evidencia moderada contra a hipótese nula.
0.05 <= P < 0.10	Evidencia sugestiva contra a hipótese nula.
0.10 <= P	Pouca ou nenhuma evidencia real contra a hipótese nula.

As interpretações dos resultados dos experimentos deste trabalho também foram realizadas em função do domínio e características do problema de pesquisa.

Outro aspecto relevante para a forma como o método será avaliado é o fato de que será usado o *holdout*, considerando que 70% da base de dados foi para treinamento e 30% para testes. Para maior credibilidade do método, para cada conjunto de parâmetros serão executados 10 experimentos com recortes diferentes na base, mas seguindo as regas já citadas do *holdout*. As medidas que avaliarão o método refletirão a média destes 10 experimentos para cada

conjunto de parâmetros.

6.3 Conclusão

Neste capítulo foram apresentados os procedimentos metodológicos para a construção deste trabalho. Inicialmente foi exposta a forma como o problema foi identificado, utilizando a pesquisa exploratória e análises iniciais. Depois, analisada a viabilidade do uso de MPS para a identificação de emoções, porém com a hipótese de utilizar rotulações de bases rotuladas para melhorar a classificação de outro método. A proposta e desenvolvimento do método foram brevemente descritas e serão melhores abordadas nos próximos capítulos. E no fim a forma como o método será avaliado.

No próximo capítulo será apresentado o método com suas fases e etapas.

Capítulo 7

Um Método de Identificação de Emoções Baseado na Mineração de Padrões Sequenciais

Neste capítulo apresentamos o método proposto neste trabalho. O método de identificação de emoções foi desenvolvido em uma abordagem de aprendizagem de máquina supervisionada.

O método é fundamentado no conceito de Mineração de Padrões Sequenciais, aplicando conceitos como: *sequencias, subsequências* e padrões durante sua execução.

Diante dos diversos estudos de AS existentes na literatura, percebeu-se que existe uma lacuna a ser preenchida. Em abordagens tradicionais de classificação de emoções em textos, na fase de treinamento, são identificadas características numéricas no texto que, submetidas a um modelo matemático, possibilitam a identificação de uma emoção em um novo texto na fase de teste. Estas abordagens procuram diversas características nos textos que correspondam a emoção que será usada na classificação, porém não "observam" as próprias rotulações que já existem em uma base, que muitas vezes foram rotuladas por especialistas humanos. Este método utiliza estas rotulações em busca de padrões que serão utilizados na classificação de novos textos. Um corpus de contos infantis rotulados com emoções auxilia no entendimento do método. Um conto infantil, por ser uma história, tem começo, meio e fim. Segue uma ordem na escrita pensada pelo seu autor. As rotulações de emoções em um conto infantil também seguem esta ordem e possibilitam que as mesmas emoções encadeadas apareçam em outros contos rotulados. O princípio do uso dos padrões das rotulações possibilita que o método os utilize para ajustes de classificação. Por consequência, o método apresentado neste trabalho não utiliza algoritmos tradicionais da literatura que fazem extração de características do próprio texto. Em

abordagens comuns de AS, na fase de treinamento, a extração de características é realizada utilizando técnicas como, por exemplo, *TF-IDF ou Bag-of-Words,* lematizador, etc. Neste método a fase de treinamento somente se utiliza das rotulações dos textos para extrair padrões. Por isso, o método pode ser aplicado em bases de textos de qualquer idioma.

Este capítulo está dividido em cinco seções. A Seção 7.1 se refere aos pressupostos do método, a Seção 7.2 trata dos parâmetros que o método possui, a Seção 7.3 fornece uma visão geral do método, a seção 7.4 apresenta o módulo de extração de padrões sequenciais, a Seção 7.5 apresenta o módulo de Classificação e a seção 7.6 a conclusão.

7.1. Pressupostos do Método

O método proposto apoia-se nas probabilidades estimadas (TAN; STEINBACH; KUMAR, 2009) para a fase de classificação. Os próximos parágrafos descrevem rapidamente como elas são obtidas.

Tradicionalmente, a tarefa de classificação envolve um conjunto de dados de treinamento. Cada instância do conjunto de treinamento contém um "valor alvo" que se refere à classe ou rótulo. Problemas desta natureza podem conter mais de duas classes. Uma estratégia para solucionar problemas que contenham mais de duas classes é combinar os classificadores gerados em subproblemas binários, sendo essa estratégia conhecida como decomposição (LORENA, 2006). Em abordagens de decomposição uma determinada instância é classificada pela combinação das previsões feitas pelos classificadores binários. Um esquema de votação é geralmente empregado para combinar as previsões, onde a classe que recebe o maior número de votos é atribuída à instância. Uma alternativa à estratégia de votação utilizada é transformar as saídas dos classificadores binários em estimativas de probabilidades e então atribuir a instância à classe que possui um valor mais alto de probabilidade estimada (TAN; STEINBACH; KUMAR, 2009).

As probabilidades estimadas seguem a escala de *Platt* ou calibração *Platt.* O método foi proposto por Jhon Platt (PLATT, 1999) no contexto do SVM, substituindo um método anterior proposto por Vapnik (VAPNIK, 1995). A escala Platt funciona ajustando um modelo de regressão logística às pontuações de um classificador. No estudo de Platt ele trata do SVM em que o score é um número de -1 a 1 e este score é transformado em classificação usando a função

SIGN. (equação 7)

$$y = sign\ (f\ (x)) \tag{7}$$

Para alguns tipos de problema é conveniente obter uma probabilidade $P\ (y = 1 \mid x)$, ou seja, uma classificação que não só retorna uma resposta, mas também um grau de certeza sobre a resposta. Alguns modelos de classificação não fornecem tal probabilidade, ou dão estimativas de probabilidades ruins.

7.2. Visão Geral do Método

O método funciona em duas fases, treinamento e classificação, e cada fase contém etapas com entradas e saídas. A Figura 7.1 apresenta uma visão geral do funcionamento do método.

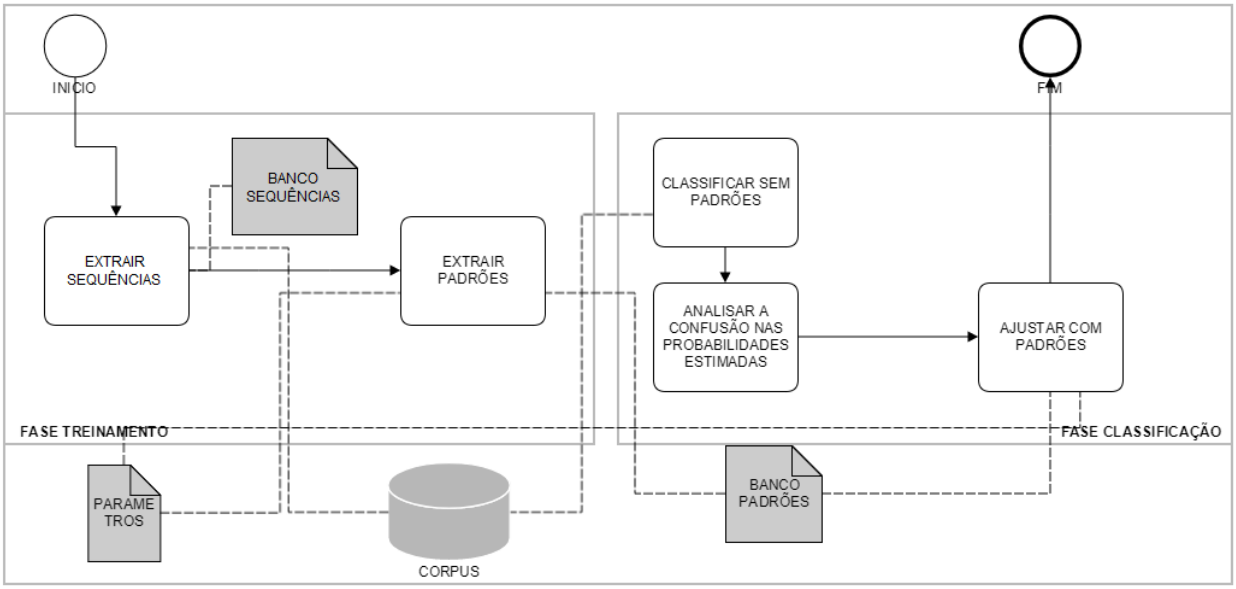

Figura 7.1: Visão Geral do método (fonte: o autor).

A fase de treinamento tem como objetivo extrair a lista de padrões a partir de uma base rotulada. Para atingir este objetivo realizadas duas etapas: Extrair Sequências e Extrair Padrões.

A fase de classificação tem como entrada a lista de padrões gerada na fase de treinamento. Esta fase é composta por três etapas: Classificar sem Padrões, Analisar a confusão nas probabilidades estimadas e Ajustar com Padrões. Nesta fase as probabilidades estimadas (TAN; STEINBACH; KUMAR, 2009) são importantes para identificar erros de classificação,

com o intuito de aplicar padrões e identificar as emoções.

7.3 Fase de treinamento

Esta seção detalha a fase de *treinamento* e as etapas que a compõem. Na Figura 7.2, é fornecido uma visão mais detalhada da fase de treinamento. O objetivo desta fase é transformar toda a base de entrada em uma lista de sequencias e extrair padrões utilizando algoritmos para esta finalidade como: *FP-Growth*, *PrefixSpan* e *GSP*. Ao final os padrões sequenciais são armazenados em uma lista para serem utilizados posteriormente.

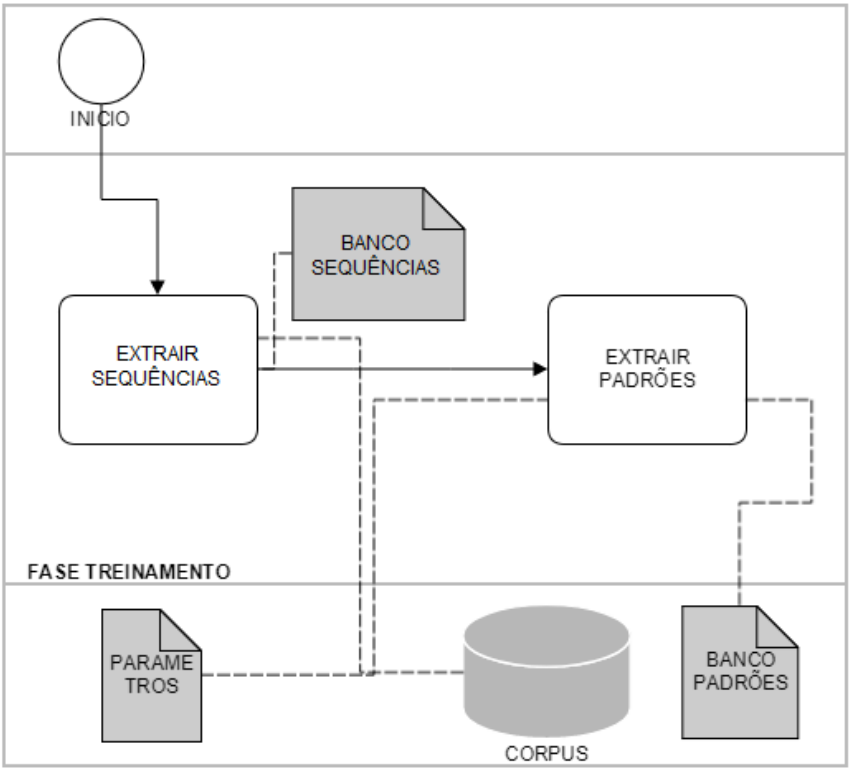

Figura 7.2: Visão detalhada da Fase de treinamento.

Esta fase é composta por dois principais módulos, sendo eles: *Extrair Sequências* e *Extrair Padrões*.

Para exemplificar o funcionamento do método, será utilizado um corpus de contos infantis (Alm et al. 2005), onde as sentenças são rotuladas com as emoções básicas de (EKMAN, 1992) e mais o *neutro*. Nesta base, cada conto infantil é um arquivo que contém textos rotulados com o conjunto de itens *E.* A Tabela 7.1 é um conjunto de textos de um conto infantil presente na base.

Tabela 7.1: Exemplo de Conto Infantil na língua inglesa

Sentença	Emoção
He has been one year with God	Triste
It was night, and quite still in the great town.	Medo
And as they flew the angel related the history.	Neutro
Understand!	Surpresa
Go along with you," said the farrier.	Neutro
Do you call this beautiful?	Repugnância
Why, there is not even a dung-heap.	Repugnância

Estes contos infantis são inicialmente processados convertendo-os de texto + rotulação para uma sequência. Nas próximas subseções são descritas as etapas do método utilizando esta base de contos infantis para exemplificar cada etapa do método, suas entradas e saídas.

7.3.1. Extrair Sequências

Esta seção tem o objetivo de detalhar a etapa de Extrair Sequências. Será apresentada a definição de sequência implementado no método e exemplificado por textos extraídos da base (ALM et al. 2005).

Uma sequência *S* pode ser definida formalmente como um ou vários conjuntos de itens: $S = <s_1, s_2, \ldots, s_k>$. O número de itens em *S* é o tamanho da sequência e é denotada por |t|. O $i^{ésimo}$ item na sequência é representado por s_i.

Neste contexto, o conjunto de itens definido para as sequências é o das emoções básicas, mais a classe *neutro* (nenhuma emoção básica identificada). O conjunto de itens *E* é definido: *E = {Raiva; Medo; Repugnância; Surpresa; Alegria; Tristeza, Neutro}*. Uma sequência hipotética formada a partir do alfabeto *E* poderia ser: *S = ({Raiva}, {Medo}, {Raiva}, {Medo}, {Repugnância}, {Surpresa}, {Surpresa}, {Surpresa}).*

Cada sentença presente em um conto infantil torna-se um item de uma sequência *S* criada a partir do arquivo de conto infantil. Na Tabela 7.2 contém uma sequência formada a partir do conto infantil da Tabela 7.1.

Tabela 7.2: Exemplo de sequência extraída de um documento do corpus de contos Infantis

Sentença01	Sentença02	Sentença03	Sentença04	Sentença05	Sentença06	Sentença07
Triste	Medo	Neutro	Surpresa	Neutro	Repugnância	Repugnância

Este processo de conversão de um conto infantil em sequência é realizado em todo *corpus*, gerando uma base de sequências. Na Tabela 7.3 é possível visualizar um exemplo de base de sequências. A coluna SID se refere ao identificador de sequência e a coluna sequência se refere às sequências formadas pelas emoções dos rótulos dos contos infantis.

Tabela 7.3: Visualização de uma base de dados de sequências utilizando contos infantis.

SID	Sequências
1	({N},{N},{N},{F},{F},{N},{N},{H},{H},{H},{N},{N},{N})
2	({Su},{Su},{N},{N},{F},{F},{N},{N},{N},{N})
3	({Sa},{Sa},{Sa},{N},{N},{N},{N},{N},{D},{D},{N},{N},{N},{N})
4	({A},{A},{A},{A},{N},{N},{N},{N},{D},{D},{D},{N},{N},{N})
5	({N},{N},{A},{N},{A},{H},{H},{N},{F})
6	({H},{H},{H},{N},{N},{N},{N},{N},{N},{N},{N},{F},{F})
7	({A},{N},{N},{N},{H},{N},{H},{N},{N})

É importante salientar que o método não impõe uma taxonomia de emoções particular sendo possível, inclusive, a utilização de polaridade (classe positiva e classe negativa).

7.3.2. Extrair Padrões

Nesta etapa do método o objetivo é identificar padrões no banco de sequências da etapa anterior. Estes algoritmos são responsáveis por gerar o banco de padrões. A saída deste processo é um arquivo conforme Tabela 7.4.

Tabela 7.4: Exemplo de banco de padrões (onde: N-Neutro, T-Tristeza, A-Alegria, S-Surpresa, M-Medo)

	Padrão									Suporte Mínimo
1º	N	N	N	T	N	N	N			36
2º	A	A	N	N	S	S				10
3º	M	M	M	N	A	A	A	N	N	8
4º	N	A	S	S	S					7
5º	T	T	T	N	M	M	M			5

Este arquivo é filtrado por parâmetros previamente configurados. O filtrar neste caso é para diminuir a quantidade de padrões em função do seu tamanho e suporte mínimo. O parâmetro *Tamanho de Padrões* reduz a quantidade de padrões do banco de padrões por tamanho, já o parâmetro *Padrões Distintos* remove os padrões contidos em outros e o parâmetro *Suporte Mínimo* filtra os padrões pelo seu suporte.

Para extrair padrões em função da ordenação dos itemsets e respeitar regiões de similaridade que agregam conhecimento sobre a rotulação, por exemplo dos contos infantis para saber se foram escritos de forma padronizada, foi implementado um algoritmo: FP-Align.

Como apresentado anteriormente, o método tem um algoritmo próprio para extrair padrões, que é descrito na próxima seção.

7.3.3 Algoritmo FP-Align

Conforme mencionado anteriormente, o corpus descrito em (Alm et al. 2005) será utilizado para exemplificar o algoritmo FP-Align. Para facilitar a escrita, serão utilizadas abreviações para as emoções básicas e neutro: N – Neutro; Tr – Tristeza; A – Alegria; Su – Surpresa; Rep – Repugnância; M – Medo; Ra – Raiva. De acordo com as abreviações o conjunto de itens para as sequências será: E = {N; Tr; A; Su; Rep; M; Ra}. O algoritmo processa uma base de sequências, por isso, para explicar o algoritmo, será utilizada a Tabela 7.3 que representa uma base desta natureza, extraída do corpus.

O algoritmo FP-Align, utilizado na extração de padrões, é uma das contribuições desta pesquisa. A partir de um banco de dados de sequências, o algoritmo se utiliza do conceito de alinhamento de sequencias para alinhar todas as sequências e encontrar padrões. Este algoritmo é uma alternativa aos algoritmos tradicionalmente utilizados como *Fp-Growth, PrefixSpan, GSP* nesta pesquisa.

Devido à natureza do problema, onde as sequências tem tamanhos diferentes e a busca

por padrões é feita em toda a extensão das sequências será utilizado o algoritmo *Needleman-Wusch* (NEEDLEMAN;WUNSCH, 1970) como fundamento para os alinhamentos feitos pelo FP-Align. O FP-Align, realiza alinhamentos de todas as todas as sequências entre si, conforme ilustrado na Figura 7.3, e o resultado de cada alinhamento é processado para encontrar regiões de similaridade entre as sequências e verificar se estas regiões podem se tornar um padrão, de acordo com o suporte mínimo.

Um alinhamento entre duas sequências nos garante encontrar a maior similaridade entre elas. As regiões de similaridade encontradas no alinhamento serão utilizadas para determinar os padrões entre todas as sequências.

Este processo de alinhamento é feito por toda a base, realizando alinhamento entre todas as sequências. Na Figura 7.3 é ilustrado este processo, onde a *Sequência 1* é alinhada com a *Sequência 2*, a *Sequência 3* ... *Sequência 5*. A *Sequência 2* é alinhada com a *Sequência 3* até a *Sequência 5*. E assim segue até que todas as sequências estejam alinhadas umas com as outras. No algoritmo, alinhamentos entre a *Sequência 1* e *Sequência 1* não é realizado, pois a região de similaridade que seria encontrada é a transação toda, o que seria um ruído.

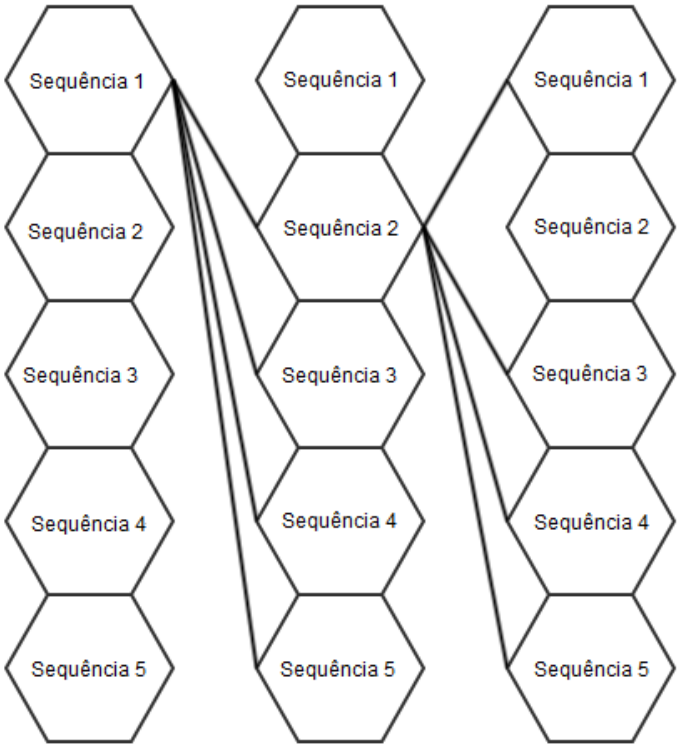

Figura 7.3: Alinhamento entre todas as sequências.

Para (SETUBAL; MEIDANIS, 1997), alinhar sequências ordenadas é inserir *gaps* em posições arbitrárias das transações conquistando a mesma medida de tamanho entre elas.

Na Tabela 7.5, é possível visualizar duas sequências sendo alinhadas para obter-se a mesma medida de tamanho entre elas. Este exemplo já recupera sequências de contos infantis.

Tabela 7.5: Duas sequências alinhadas com a maior similaridade entre elas.

M	**N**	**N**	**-**	**A**	**A**	**N**	**N**	**A**	**M**	**Tr**	**-**	**Tr**	**N**	**N**
-	N	N	A	-	-	N	N	A	-	-	N	Tr	N	N

O próximo passo do algoritmo é encontrar padrões nos alinhamentos realizados. As regiões de similaridade do alinhamento que são idênticas entre as duas sequências são consideradas padrões. No algoritmo, para identificar os padrões, utiliza-se o recurso de *matching,* (que contenha a mesma emoção na mesma posição entre os vetores sequenciais) e *mismatching* (o inverso do matching). Todas as regiões que contem *matchs* em transação e que, anterior e posterior a esta região contenha um *mismatch* é um padrão.

A Tabela 7.6 apresenta um exemplo de alinhamento e processamento o processamento de *matching*/*mismatching* para identificar as regiões de similaridade. Neste exemplo, *match* recebe 1 e *mismatch* recebe 0. Em vermelho estão selecionadas as regiões de similaridade no alinhamento.

Tabela 7.6: Alinhamento e busca de padrões.

M	N	N	-	A	A	N	N	A	M	Tr	-	Tr	N	N
-	N	N	A	-	-	N	N	A	-	-	N	Tr	N	N
0	1	1	0	0	0	1	1	1	0	0	0	1	1	1

Estas regiões são armazenadas em um *Banco Padrões* e contabilizadas com o seu respectivo suporte mínimo.

7.4. Fase de Classificação

A Figura 7.4 apresenta uma visão geral da fase de Classificação. Esta fase utiliza o *Banco Padrões* gerado na primeira fase para tratar possíveis erros encontrados nas probabilidades estimadas da *Classificação sem Padrões*.

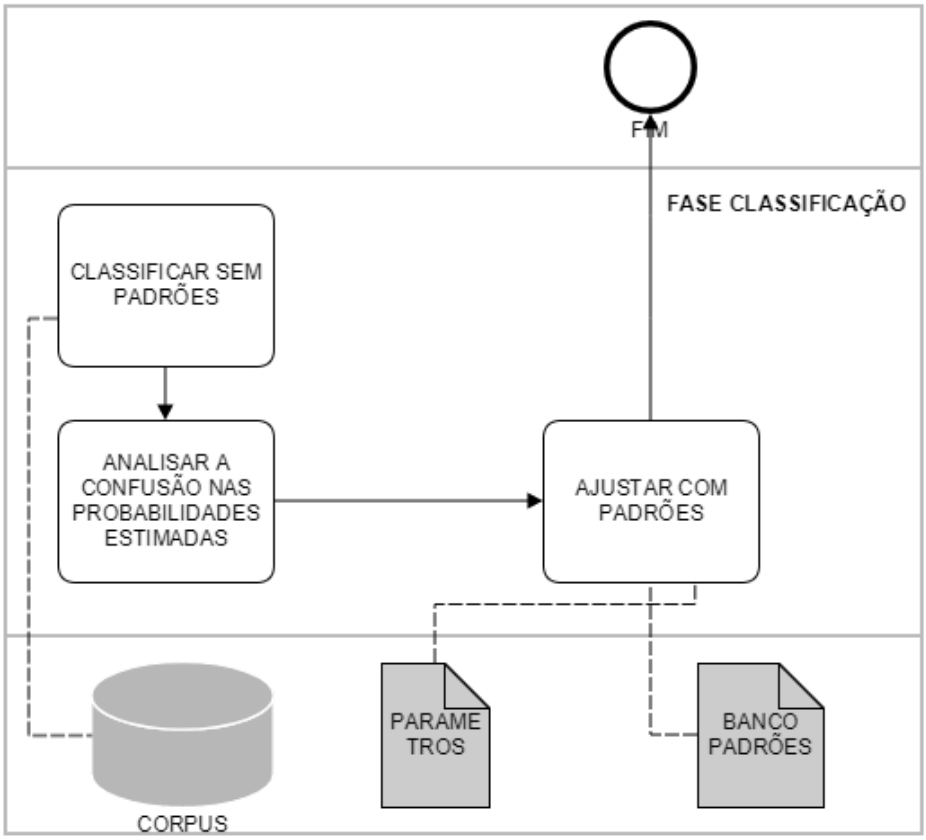

Figura 7.4: Visão detalhada da fase de classificação.

Esta fase é composta por três processos, sendo eles: Classificar sem Padrões, Analisar Confusão nas probabilidades estimadas e Ajustar com Padrões.

7.4.1. Classificar sem padrões

O processo de classificar sem Padrões consiste em aplicar um método de identificação de emoções capaz de indicar a emoção predominante em cada texto analisado. O método escolhido para esta etapa é o proposto por (DOSCIATTI, 2015), por ser capaz de identificar as emoções básicas. O método (DOSCIATTI, 2015) exporta as Probabilidades Estimadas (PEs) calculadas durante o processo de classificação, o que será útil na sequência da classificação utilizada neste novo método.

7.4.2. Analisar Confusão nas Probabilidades Estimadas

O processo *analisar confusão nas probabilidades estimadas* verifica como o processo anterior identificou as emoções nos textos do *corpus* submetido ao método de identificação de emoções escolhido. As *PEs* de cada instância são retornadas como um vetor, onde na primeira posição está o texto (instância) e nas outras posições a emoção e sua respectiva probabilidade estimada. O processo *analisar confusão nas probabilidades estimadas* reconstrói os contos infantis que foram submetidos ao processo anterior e extrai as sequências destes contos infantis. A saída deste processo é um banco de sequências gerado a partir dos resultados da classificação do processo *Classificar sem padrões*. A Tabela 7.7 ilustra como cada sequência é construída utilizando *PEs*.

Tabela 7.7: Uma sequência com as possíveis emoções para cada item da sequência. R-raiva, N-neutro, A-alegria, T-tristeza, S-surpresa, M-medo.

Classe/ instancia	**Valor da *PE* para cada instancia.**								
	1ª	2ª	3ª	4ª	5ª	6ª	7ª	8ª	9ª
1ª	*R (0.7)*	*M (0.6)*	*M (0.3)*	*A (0.4)*	*R (0.35)*	*R (0.9)*	*M (0.95)*	*T (0.8)*	*R (0.75)*
2ª	*N (0.2)*	*T (0.3)*	*T (0.25)*	*M (0.2)*	*N (0.25)*	*N (0.05)*	*S (0.05)*	A (0.1)	*M (0.15)*
3ª	*A (0.1)*	*N (0.1)*	*N (0.20)*	*S (0.1)*	*A (0.2)*	*A (0.05)*	*N (0.0)*	*S (0.1)*	*N (0.08)*
4ª	*T (0.0)*	*S (0.0)*	*S (0.22)*	*N (0.1)*	*M (0.2)*	*M (0.0)*	*A (0.0)*	*N (0.0)*	*A (0.02)*

O exemplo da Tabela 7.7 contêm um conto infantil com nove textos (instâncias). Para cada texto quatro classes são possíveis. Os valores das *PEs* estão ordenados de forma decrescente, sendo o *1ªClasse* a classe predominante. As probabilidades são distribuídas entre todas as possíveis classes somando cem por cento.

O próximo passo nesta etapa é encontrar pontos de confusão na classificação realizada no módulo *classificação sem padrões*. Para isto, são utilizadas as *PEs* geradas que retornam, em porcentagem, a probabilidade de uma classe estar associada a um texto. Exemplo conforme Tabela 7.7.

Para definir se uma instância tem classificação confusa, é utilizado o parâmetro *%Confusão.* Uma vez que cada instancia contém as PEs de todas as emoções para ela, confusão no contexto deste trabalho, significa que uma instância, não tem uma emoção com um percentual grande o suficiente para garantir que a classificação é correta. O parâmetro *%Confusão* é um valor flutuante, utilizado para encontrar as confusões da

classificação nas PEs. Para efeito de explicação, será utilizado o valor 0.15 no parâmetro de %Confusão e na seção de experimentos do capítulo 8, será explicado o parâmetro e o melhor valor encontrado para os experimentos com a base de contos infantis.

Na Figura 7.5 temos um exemplo de instância com as respectivas emoções e seus valores de probabilidades estimadas e uma análise de confusão.

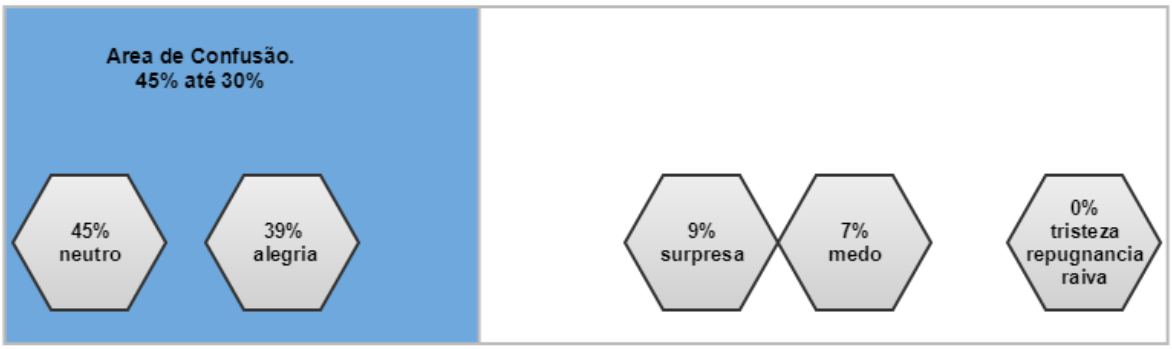

Figura 7.5: Análise de confusão: Encontro de duas classes confusas.

Para esta instancia, as possíveis emoções estão distribuídas probabilisticamente na linha de porcentagem e as classes mais prováveis são mais próximas possíveis de 100%, neste caso neutro. A partir do maior valor de percentual encontrado entre as classes, neste caso 45% é analisado uma área: área de confusão que, neste caso, vai de 45% até 30%. Caso tenha mais que uma classe nesta área, o método entende que a classificação é confusa e considera esta instancia como ponto de confusão.

Ainda no exemplo da Figura 7.5, a próxima emoção é a *alegria* com 0.39 percentuais. Esta emoção está acima de 30%, limite inferior da área de confusão. Neste caso, todas as classes que estiverem na área de confusão são consideradas para a próxima etapa do método para ajusta-las.

Na Figura 7.6 é ilustrado um caso onde não existe confusão na classificação do processo *Classificar Sem Padrões*.

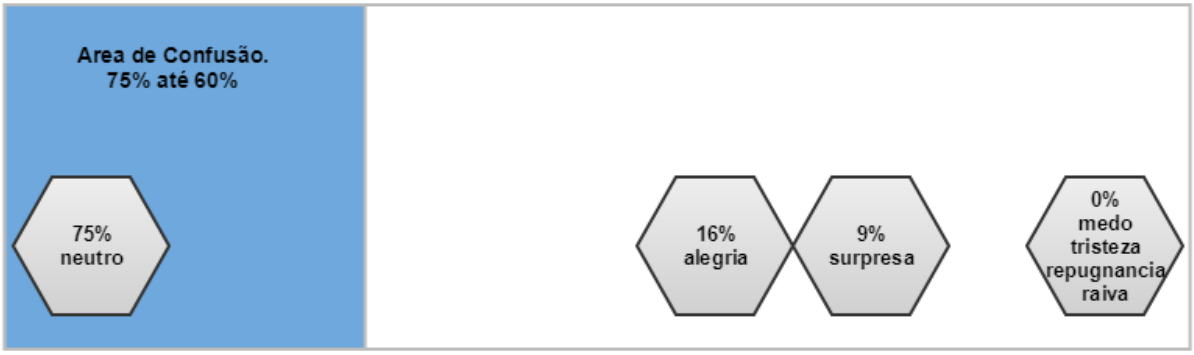

Figura 7.6: Análise de confusão. Sem confusão na classificação.

Neste caso, o parâmetro *%Confusão* também recebe o valor 0.15. A classe com maior percentual de probabilidade é a *neutro* com 0.75. A área de confusão neste caso tem seu início no percentual 0.75 e se estende até 0.60. Não há outra classe com probabilidade nesta área, por isto, esta instancia não tem confusão em sua classificação.

Todas as sequências nesta fase passam pelo processamento para buscar pontos de confusão. Uma vez que foram encontrados os pontos de confusão é possível aplicar os padrões da fase *de treinamento* para melhorar a classificação. Na próxima subseção é descrita como estas confusões são possíveis de serem corridas com os padrões da fase de *treinamento*.

7.4.3. Ajustar com Padrões

A terceira etapa *ajustar com padrões* aplica os padrões do *Banco Padrões* encontrados na primeira fase para corrigir as confusões da classificação do processo *classificação sem padrões*.

Uma sequência, neste contexto, é um conjunto de itens de emoções que representam textos ordenados. Cada item de uma sequência é uma instância por se tratar de um problema de classificação. Cada instância classificada até este momento será processada e analisada para identificar se existe confusão ou não em sua classificação. Neste caso, quando se coloca em evidencia uma sequência no processamento de análise de confusão, define-se que um ponto de confusão é uma instância da sequência que não foi classificada sem confusão. Em cada ponto de confusão de uma sequência as possíveis classes são identificadas em função da área de confusão pré-definida. Na Figura 7.7 é ilustrado uma sequência com dois pontos de confusão.

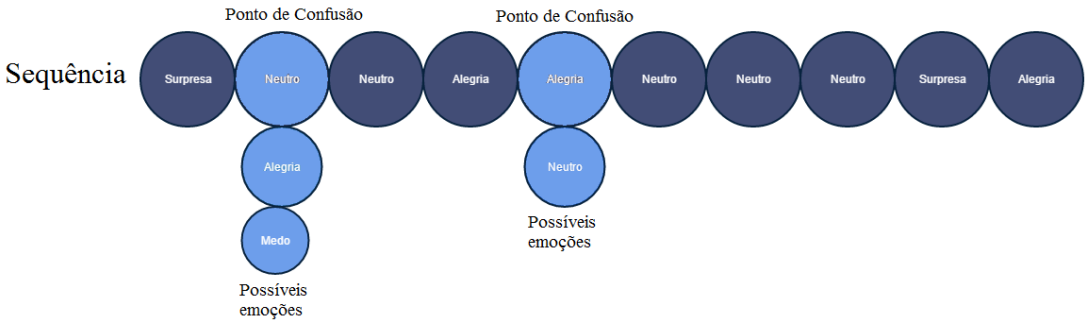

Figura 7.7: Sequência com dois pontos de confusão e suas possíveis emoções.

Os dois pontos de confusão citados na Figura 7.7 são os da 2ª instância e 5ª instância. Na confusão da 2ª instância desta sequência as classes que estão na área de confusão são: {*neutro, alegria, medo*}. Já na confusão da 5ª instância as classes que estão na área de confusão são: {*alegria, neutro*}

Em cada ponto de confusão deverá ser encaixado o melhor padrão para a melhor classe do ponto de confusão. Por isso, é realizada uma busca no *Banco Padrões*, identificando quais padrões possuem as classes que são possíveis naquele ponto de confusão. Na Figura 7.8 é possível visualizar as classes da área de confusão da 2ª instância da Figura 7.7. São possíveis três emoções para esta instância, e para cada emoção busca-se seus padrões no *Banco Padrões*. Esta busca de padrões no banco de padrões é realizada seguindo o critério de existência da emoção no padrão.

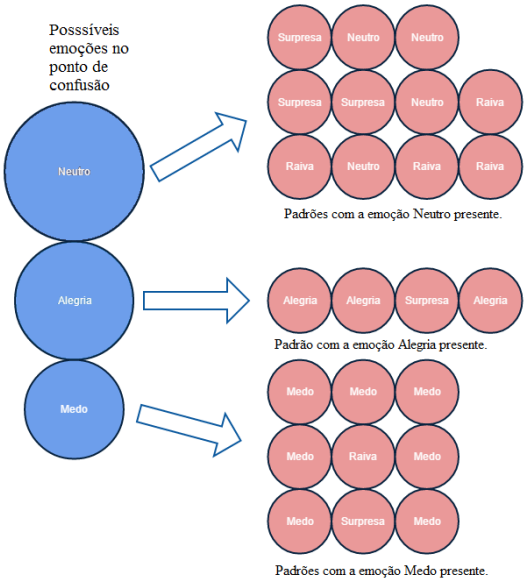

Figura 7.8: Emoções de um ponto de confusão com seus padrões filtrados do banco de padrões.

Identificados os padrões para cada emoção no ponto de confusão, inicia-se a aplicação destes padrões na sequência. De forma simples, o padrão é encaixado na posição da instância. Este encaixe leva em consideração a posição da emoção no padrão. Importante lembrar que a emoção cada emoção tem seu subconjunto de padrões.

A forma como o padrão será encaixado depende da quantidade de vezes que a emoção é encontrada no padrão. No mínimo o padrão contém uma vez a emoção que o filtrou do banco de padrões, porém há casos que a emoção aparece duas ou mais vezes. Sendo assim, todos os encaixes são testados.

Caso a emoção apareça mais de uma vez em um padrão, o padrão é encaixado de várias formas, de modo que todas as possibilidades de encaixe do padrão no ponto de confusão sejam exploradas. A Tabela 7.8 ilustra uma transação com um ponto de confusão e um padrão sendo encaixado de duas formas.

Tabela 7.8: Ponto de confusão e um padrão sendo encaixado como tentativa de melhoria de classificação.

Transação	**Triste**	**Medo**	**Neutro**	**Surpresa**	**Neutro**	**Repugnância**	**Repugnância**
Encaixe 01			Neutro	Surpresa	Neutro		
Encaixe 02	Neutro	Surpresa	Neutro				

Neste caso, a sequência tem sete itens e o ponto de confusão ocorreu na 3ª instância. Esta instância tem a emoção neutra como uma possível emoção que foi extraída da área de confusão. O padrão {neutro, surpresa e neutro} foi um padrão encontrado no banco de padrões. Este padrão contém a emoção neutro duas vezes na sua extensão, sendo assim deverá ser encaixada duas vezes. A primeira tentativa de encaixe a posição 1 do padrão é encaixado na instância. Neste caso é verificado se toda a extensão do padrão é totalmente igual às outras instâncias da sequência. Se todas as instâncias forem iguais, conforme o encaixe do padrão, este padrão é possível de ser utilizado. Na segunda tentativa a posição 3 foi encaixada na instância de confusão. Neste caso, as outras instâncias não são iguais ao padrão na sua extensão, então esta possibilidade é descartada.

Por definição, para considerar que um padrão se encaixa, toda sua extensão deve estar de acordo com a fração de sequência que ele está sendo comparado. Inclusive, para o cálculo de score da equação (3), somente são calculados padrões que tem este pré-requisito. No exemplo da Tabela 7.8, o padrão se encaixou na primeira tentativa.

Todos os padrões de todas as possíveis classes de um ponto de confusão são testados seguindo estas definições. Para escolher qual classe deverá ser usada naquele ponto de dúvida, é calculado um score por classe, dado pela seguinte equação:

$$Scoreclasse = \sum(TamanhoPadrao * SuporteMínimo) \quad (8)$$

Na equação (8) é realizado um somatório do produto entre o tamanho do padrão e seu suporte mínimo. Para cada emoção é computado um *score* e a emoção de maior *score* é selecionada. Em caso de empate é selecionada a classe com maior porcentagem na probabilidade estimada.

O método realiza o processamento em todos os pontos de confusão encontrados e em todas as sequências. Ao final deste processo espera-se que a sequência final contenha menos erros de classificação que a sequência inicial. É importante destacar que existe a possibilidade de inserção de erros na tentativa de melhorar a classificação, ou seja, "corrigir" erroneamente aquilo que já estava correto.

7.5 Conclusão

Neste capítulo é apresentado um método para melhorar a classificação de outros métodos através da Mineração de Padrões Sequenciais. O algoritmo *Fp-Align* possui um *kernel,* onde é utilizado o algoritmo *Needleman-Wusch* para identificar regiões de similaridade e analisar se estas regiões são padrões sequenciais.

O método apresentado neste capítulo possui algumas limitações: 1) possibilidade de inserção de maior quantidade de erros ao invés de acerto na fase de classificação; 2) Falta de bases para testes.

O próximo capítulo, que se refere ao procedimento metodológico, apresenta os detalhes de implementação, o modo de testar e avaliar o método e as métricas utilizadas para medir o desempenho do método.

Capítulo 8

Resultados Experimentais

Neste capítulo são descritos os resultados obtidos dos experimentos realizados. Os resultados foram obtidos através de experimentos realizados com base nas fases e etapas que compõem o método deste trabalho. A apresentação dos resultados obtidos é realizada do seguinte modo: primeiramente são mostrados os resultados produzidos pelas Fases de Treinamento e de Classificação do método proposto. Os resultados são comparados utilizando as medidas de precisão, cobertura, F1 e acurácia e testados estatisticamente com o teste *t-student*.

8.1. Implementação do Método

O método apresentado no Capítulo 7 foi implementado na linguagem *Python*[1], na versão 2.7. O algoritmo *Needleman-Wusch*[3], foi implementado pelo próprio autor e validado com dados disponibilizados na *Web: http://pt.slideshare.net/mcastrosouza/algoritmo-needlemanwunsch*.

Para que o método pudesse ser testado e avaliado, foi construída uma ferramenta, com um ambiente configurável via arquivo e sua execução através de *scripts*. Na *Figura 8.1* são apresentadas as configurações do método.

```
[NEEDLEMAN]
match = 1
mismatch = -2
gap = -3

[TREINAMENTO]
tamanhopadroes = 5
somentepadroesdistintos = 1
suporteminimo = 0.5
confusao = 0.15
```

Figura 8.1: Arquivo de Configuração do Método

Na seção *Treinamento* da Figura 8.1 é possível visualizar os parâmetros que o método possui. Na seção *Needleman* são configurados os valores para o cálculo do esquema de pontuação no alinhamento de sequencias. Neste caso foram utilizados os valores padrão: 1 para match, -2 para *mismatch* e 3 para gaps.

Para fazer uma avaliação do método a ferramenta requer um arquivo de treinamento e um arquivo de teste, ambos em formato .TXT. Este arquivo deve obedecer a um formato onde os textos devem vir em primeiro lugar, seguido de um espaço tabular e a respectiva rotulação do texto.

8.2. Experimento para construção de um *baseline*

Nesta seção são apresentados os resultados de experimentos realizados para compor um *baseline*. Estes experimentos têm como objetivo mostrar que o acaso não tem melhores resultados que o método proposto.

A base de contos infantis da seção 6.2.2 é uma base desbalanceada e por ter esta característica existem emoções mais predominantes na base do que outras. Partindo desta premissa, estes experimentos consistem em inserir, em todos os pontos de dúvida encontrados na fase de classificação, uma emoção predominante da base.

O primeiro experimento foi inserir a emoção Alegria nos pontos de dúvida e o resultado foi o da Tabela 8.1:

Tabela 8.1: Comparação de resultados entre o método (Dosciatti.2015) e o baseline Alegria.

Classe	(DOSCIATTI, 2015)			***Baseline*** (Alegria)		
	Precisão	Cobertura	F1	Precisão	Cobertura	F1

Happy	0.37	0.41	0,39	0.52	0.88	0.65
Sad	0.66	0.5	0,58	0.73	0.50	0.59
Angry-Disgusted	0.64	0.82	0,72	0.27	0.14	0.18
Fearful	0.69	0.60	0,64	0.76	0.4	0.52
Surprised	0.6	0.14	0,22	1.0	0.1	0.17
Acurária		**61,40%**			**56,0%**	

O resultado geral, medido pela acurácia, teve uma queda percentual considerável e como esperado, a cobertura da emoção Alegria aumentou, porém afetou todas as outras nesta medida e seus números caíram. A conclusão deste experimento mostra que, mesmo classificando a emoção com mais amostras na base o resultado não é satisfatório.

O segundo experimento foi inserir a emoção Tristeza nos pontos de dúvida, uma vez que ela é a emoção mais presente na base toda. O resultado foi o da Tabela 8.2:

Tabela 8.2: Comparação de resultados entre o método (Dosciatti.2015) e o baseline Tristeza.

Classe	(DOSCIATTI, 2015)			*Baseline* (Tristeza)		
	Precisão	Cobertura	F1	Precisão	Cobertura	F1
Happy	0.37	0.41	0,39	0.74	0.66	0.70
Sad	0.66	0.5	0,58	0.73	0.49	0.59
Angry-Disgusted	0.64	0.82	0,72	0.27	0.14	0.18
Fearful	0.69	0.60	0,64	0.26	0.77	0.39
Surprised	0.6	0.14	0,22	1.0	0.09	0.17
Acurária		**61,40%**			**52,0%**	

A acurácia deste experimento também demonstrou que mesmo utilizando outra emoção predominante na base para a classificação, os resultados não são satisfatórios. Mais uma vez, a

emoção que foi utilizada no experimento teve sua cobertura aumentada, e neste caso a precisão chegou a diminuir.

Estes experimentos mostraram a relevância do método no aspecto de encontrar o melhor padrão para o ponto de dúvida. Foi construída uma inteligência para que o ponto de dúvida seja corrigido de acordo com o melhor padrão da base. Os próximos experimentos seguirão esta inteligência e as mesmas medidas serão analisadas.

Nas próximas seções serão apresentados os resultados do experimento que obteve o melhor resultado, porém foram realizados mais de 250 experimentos modificando todos os parâmetros do método. Estes experimentos serviram para avaliar quanto os parâmetros influenciam nos resultados e estas analises serão apresentadas posterior ao melhor resultado obtido no método.

Outro detalhe importante é que os experimentos foram realizados com a forma *holdout* Para dar mais credibilidade método, para cada conjunto de parâmetros foram realizados 10 experimentos com diferentes subconjuntos da base. Isto quer dizer quer dizer que os resultados apresentados sempre serão a média dos resultados dos 10 experimentos para um conjunto de parâmetros.

8.3. Resultados produzidos pela Fase de treinamento

Nesta seção são apresentados os resultados que foram produzidos por cada umas das etapas que compõem a Fase de treinamento do método. O corpus utilizado neste momento é o subconjunto de Ghazi e colegas (GHAZI; INKPEN; SZPAKOWICZ, 2010) e as saídas de cada etapa correspondem a este corpus.

Extrair Sequências: Nesta etapa foram extraídas 130 sequencias, onde a origem foram contos inteiros processados, para compor o banco de transações. A saída deste processo é um arquivo de transações conforme a amostra da Tabela 8.3. Em todos os 10 experimentos holdout tiveram a mesma quantidade de sequencias, uma vez que é 70% da quantidade de contos da base de contos infantis.

Tabela 8.3: Saída da etapa de extrair padrões.

SID	Sequência
1	({Surprised}, {Sad}, {Sad}, {Happy}, {Happy}, {Angry-Disgusted}, {Angry-Disgusted}, {Happy}, {Surprised})

2	({Happy}, {Angry-Disgusted}, {Happy}, {Angry-Disgusted}, {Angry-Disgusted}, {Fearful}, {Happy})
3	({Sad}, {Sad}, {Sad}, {Fearful}, {Fearful}, {Angry-Disgusted}, {Surprised})
4	({Surprised}, {Happy}, {Angry-Disgusted}, {Angry-Disgusted}, {Angry-Disgusted}, {Angry-Disgusted}, {Happy}, {Happy})
5	({Sad}, {Fearful}, {Angry-Disgusted}, {Angry-Disgusted}, {Surprised}, {Happy}, {Happy}, {Surprised}, {Angry-Disgusted}, {Sad}, {Angry-Disgusted}, {Happy}, {Happy})
6	({Sad}, {Happy}, {Fearful}, {Fearful}, {Sad}, {Sad}, {Happy})
7	({Fearful}, {Happy}, {Happy}, {Sad}, {Angry-Disgusted}, {Happy})

A base é formada por sequencias de diversos tamanhos, por isso, para maior clareza segue um gráfico com a distribuição dos tamanhos das sequências. No eixo X da Figura 8.2 são plotados os tamanhos e no eixo Y a quantidade para cada tamanho de transação. Esta distribuição por tamanho é a média das sequencias encontradas nos subconjuntos usados nos 10 experimentos *holdouts* para os parâmetros usados.

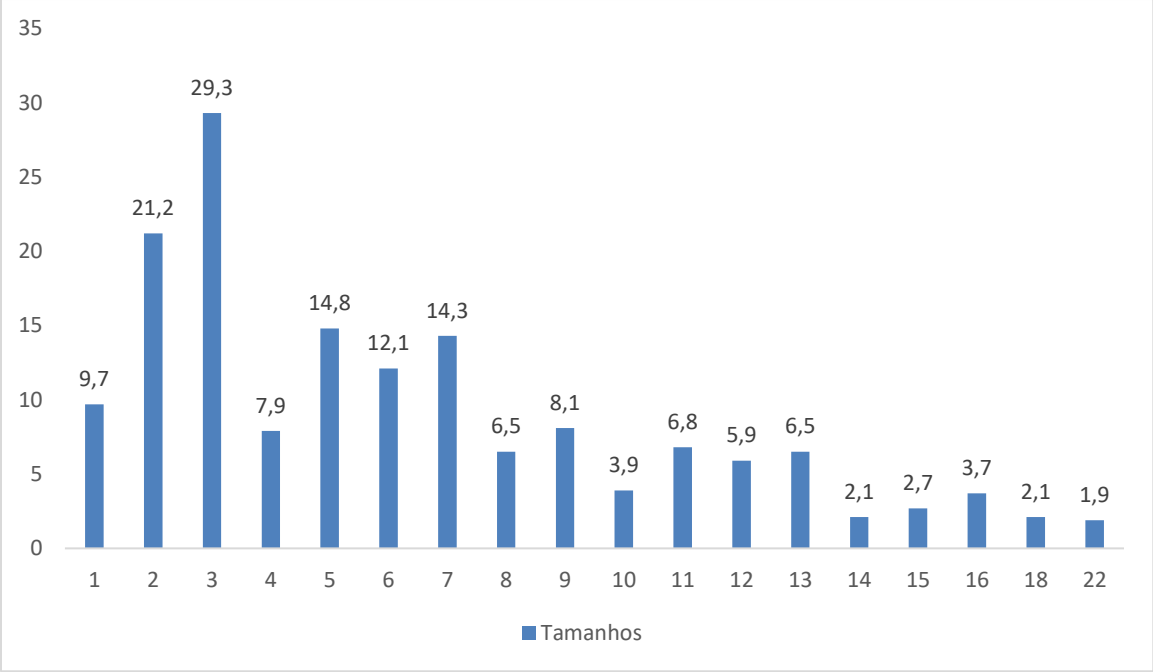

Figura 8.2: Distribuição da quantidade X tamanho de sequência.

Geralmente, as sequências são pequenas, o que dificulta encontrar padrões, uma vez que

o processamento do *FP-Align* utiliza o recurso de alinhamento. Este alinhamento é feito por similaridade e ainda definido um tamanho mínimo que esta região de similaridade deve ter para ser um padrão, o que reduz a quantidade de padrões.

Extrair Padrões: Os padrões que do banco de padrões desta etapa, foram extraídos usando os parâmetros suporte mínimo com valor 15, pois devido aos experimentos realizados foi o que melhor deu resultado, o que significa que os padrões que estão nesta lista apareceram em pelo menos 15-sequências do banco de sequencias. O tamanho mínimo para filtrar os padrões foi 3.

Nesta etapa, através do algoritmo FP-Align e o processamento realizado entre as sequências com o algoritmo *Needleman-Wusch* foi possível extrair a Banco de Padrões. Foram processadas 16.770 combinações de alinhamento entre as sequências com o intuito de extrair padrões. Este número de combinações é igual em todos experimentos porque a divisão da base sempre respeita usar 70% (130) das sequencias da base. Por considerarmos que não se alinha uma sequencia com ela mesma, o resultado é o produto de 130x129 alinhamentos.

No total foram em média 13 padrões extraídos com os seus respectivos suportes mínimos. Na Tabela 8.3 é possível visualizar o um conjunto de banco de padrões com seus respectivos suportes mínimos.

Tabela 8.3: Banco de Padrões extraída no experimento.

Padrão	**Suporte Mínimo.**
({Angry-Disgusted}, {Angry-Disgusted}, {Angry-Disgusted})	69
({Sad}, {Sad}, {Sad})	36
({Fearful}, {Fearful}, {Fearful})	31
({Happy}, {Sad}, {Happy})	22
({Sad}, {Sad}, {Happy})	21
({Surprised}, {Happy}, {Happy})	20
({Happy}, {Happy}, {Happy}, {Happy})	20
({Fearful}, {Sad}, {Sad})	20
({Sad}, {Happy}, {Happy})	18
({Happy}, {Happy}, {Surprised})	18

Como é possível observar na tabela 8.3, tem padrões que ocorrem em mais de 50% das sequencias do banco de sequencias, o que reforça a hipótese que existem padrões nas sequencias da base de dados. A próxima etapa se preocupa em usar estes padrões para melhorar a classificação de outro método.

8.4. Resultados produzidos pela Fase de Classificação

Esta seção descreve os resultados produzidos de cada etapa que compõem a Fase de Classificação do método proposto em relação ao experimento que foi realizado.

Classificar Sem Padrões: Nesta etapa a base foi submetida ao método (DOSCIATI, 2015) do qual foi extraído as medidas de desempenho de sua classificação como: acurácia, precisão, cobertura e F1. Estas medidas serão utilizadas para identificar as melhorias que este estudo trouxe aos resultados da classificação inicial. Também são utilizadas desta etapa as saídas das probabilidades estimadas que servem como meio para saber como o método (DOSCIATI, 2015) classificou os textos e como foram distribuídas as probabilidades para cada possível classe de cada instância na classificação. Também ao método (DOSCIATI, 2015) foi submetido os mesmos subconjuntos de base do *holdout*.

Analisar Confusão nas probabilidades estimadas: O arquivo de probabilidade estimada provê ao método proposto a classificação do método utilizado na etapa anterior e permite identificar confusões em função do parâmetro de confusão. Na Tabela 8.4 é descrito o número de confusões na classificação do método de Dosciatti.

Tabela 8.4: Número médio de confusões encontradas na etapa de análise de confusão nas probabilidades estimadas.

# Confusões	%Confusão
93,5	15%

Foram encontradas, na média, 93 confusões com um percentual da área de confusão de 0.15.

Ajuste com padrões: Os padrões encontrados no Banco de Padrões foram aplicados em cada ponto de confusão encontrado originando uma nova classificação, que se trata do melhor padrão aplicado e ajustado. O ajuste de padrões revelou algumas peculiaridades. Para instâncias onde o método de Dosciatti não tinha total convicção na sua classificação, com os padrões foi possível confirmar a emoção correta. Em outras instâncias o método corrigiu erros da classificação obtida por Dosciatti. Porém, em outros casos o método aqui proposto inseriu erros onde não existiam. Na Tabela 8.5 é possível observar estes números.

Tabela 8.5: Números de correções inseridas e erros inseridos.

# Confusões	Correções Inseridas	Erros Inseridos

93	13	4

Nota-se que o método surtiu o efeito esperado nesta base. Foram 13 inserções de acertos, na média, onde a classificação estava errada e 4 erros inseridos onde a classificação inicial estava correta.

Importante comparar os resultados das duas classificações considerando as medidas de Acurácia, precisão, cobertura e F1. Na Tabela 8.6 estão os resultados dos experimentos com o método de Dosciatti e o método proposto.

Tabela 8.6: Resultado do experimento do método de Dosciatti (DOSCIATTI, 2015) e o método proposto.

Classe	**(DOSCIATTI, 2015)**			**Método Proposto**		
	Precisão	Cobertura	F1	Precisão	Cobertura	F1
Happy	0.37	0.41	0,39	0.55	0.61	0,57
Sad	0.66	0.5	0,58	0. 79	0.55	0,61
Angry-Disgusted	0.64	0.82	0,72	0.68	0.81	0,75
Fearful	0.69	0.60	0,64	0.68	0.64	0,63
Surprised	0.6	0.14	0,22	0.73	0.13	0,24
Acurária	**61,40%**			**65.40 %**		

Entre as duas classificações é possível observar que o método de Dosciatti ficou abaixo no quesito precisão em quase todas as classes, deixando somente *Fearful* em uma situação muito semelhante do método deste trabalho. Na cobertura duas classes ficaram com mesmo valor: *Surprised* e *Angry-Disgusted*. Nas outras, este método obteve melhores resultados. Já na medida F1 o método deste trabalho teve resultados melhores em todas as classes. A medida geral de acurácia obteve uma variação de 3,7 comparado com (DOSCIATI, 2015).

A base também foi submetida ao método de classificação de Stanford nas mesmas condições que foram submetidas para (DOSCIATTI, 2015) e este método. O método está disponível em (https://nlp.stanford.edu/software/) e pode ser utilizado como forma de comparação para métodos de classificação de textos. O método foi desenvolvido pelo *Stanford*

NLP (Natural Language Processing) Group e fornece ferramentas de estatística para processamento de linguagem natural, *Deep Learning* para *NLP* e ferramentas de regras baseadas em linguagem natural para problemas de linguística computacional.

Os resultados deste experimento com o método de Stanford podem ser visualizados na Tabela 8.7 e comparados com outros métodos, como (GHAZI; INKPEN; SZPAKOWICZ, 2010), que também utilizaram a mesma base em suas pesquisas.

Tabela 8.7: Experimentos de métodos da literatura com a base de contos. Precisão (P), Cobertura (C) e F1 (F).

Classe	(GHAZI; INKPEN; SZPAKOWICZ, 2010)			Stanford			(DOSCIATTI, 2015)			Método Proposto		
	P	C	F	P	C	F	P	C	F	P	C	F
Happy	0.56	0.86	0.68	0.69	0.75	0.72	0.37	0.41	0.39	0.55	0.61	0,57
Sad	0.67	0.53	0.59	0.66	0.66	0.66	0.66	0.5	0.58	0. 79	0.55	0,61
Angry-Disgusted	0.54	0.43	0.48	0.47	0.54	0.5	0.64	0.82	0.72	0.68	0.81	0,75
Fearful	0.59	0.38	0.46	0.64	0.50	0.56	0.69	0.60	0.64	0.68	0.64	0,63
Surprised	0.35	0.10	0.16	0.36	0.23	0.28	0.6	0.14	0.22	0.73	0.13	0,24
Acurácia	**57,41%**			**62,30%**			**61,40%**			**65.40 %**		

Na Tabela 8.7 é feito a comparação dos resultados de todas as medidas (precisão, cobertura, F1 e acurácia) entre todos os métodos. Para cada medida foi colocado em evidência qual método obteve melhor resultado e para qual emoção foi o resultado.

8.5. Análise Estatística

Foi realizado o teste estatístico *t-student* para um comparativo entre as medidas de desempenho dos dois métodos: (DOSCIATTI, 2015) e o método proposto. A Tabela 8.8 apresenta o valor *p-value* calculado para cada medida de desempenho (precisão, cobertura e F1) e também o significado para cada *p-value*. A significância que será utilizada é de *0.05*.

Importante salientar que este e o resultado da média dos experimentos para o conjunto dos parâmetros que foram utilizados nos experimentos acima.

Tabela 8.8: p-value de cada medida de desempenho do experimento de (DOSCIATTI, 2015) e o método proposto.

Medida	***p – value***	**Significado**
Precisão	0.133342	Não rejeita a hipótese nula.
Cobertura	0.38918	Não rejeita a hipótese nula.
F1	0.35042	Não rejeita a hipótese nula.

Como é possível observar na Tabela 8.8, os *p-values* foram superiores a *0.05,* isto indica que não foi possível perceber melhorias na classificação estatisticamente. Porém, neste momento estamos comparando dois métodos que são complementares e é possível que a melhoria não seja o suficiente para comprovação estatística. Mais à frente serão utilizados outros métodos de classificação para melhor comparação com método proposto.

Foi realizado o teste estatístico *t-student* para um comparativo entre as medidas de desempenho dos dois métodos: (GHAZI; INKPEN; SZPAKOWICZ, 2010) e o método proposto. Também foi realizado o mesmo teste estatístico entre os métodos (GHAZI; INKPEN; SZPAKOWICZ, 2010) e (DOSCIATTI, 2015) para análise dos *p-values*. O objetivo desta comparação entre os métodos é identificar se os ajustes aplicados pelo método proposto na classificação de (DOSCIATTI, 2015) surtem efeito que podem ser comprovados estatisticamente.

A Tabela 8.9 apresenta o valor *p-value* calculado para cada medida de desempenho (precisão, cobertura e F1) e também o significado para cada *p-value*. A significância que será utilizada é de *0.05*.

Tabela 8.9: Comparação de p-values entre os experimentos da literatura que utilizou contos infantis.

MEDIDA	MÉTODO		*P - VALUE*	VARIAÇÃO	SIGNIFICADO
PRECISÃO	(GHAZI; INKPEN; SZPAKOWICZ, 2010)	(DOSCIATTI, 2015)	0.28	-0,23	Não rejeita a hipótese nula.
	(GHAZI; INKPEN; SZPAKOWICZ, 2010)	Método	0.05		Rejeita a hipótese nula.
COBERTURA	(GHAZI; INKPEN; SZPAKOWICZ, 2010)	(DOSCIATTI, 2015)	0.45	-0,09	Não rejeita a hipótese nula.
	(GHAZI; INKPEN; SZPAKOWICZ, 2010)	Método	0.36		Não rejeita a hipótese nula.
F1	(GHAZI; INKPEN; SZPAKOWICZ, 2010)	(DOSCIATTI, 2015)	0.41	-0,13	Não rejeita a hipótese nula.
	(GHAZI; INKPEN; SZPAKOWICZ, 2010)	Método	0.28		Não rejeita a hipótese nula.

A medida de desempenho precisão foi a que mais teve mudanças estatística entre os resultados. O *p-value* da precisão dos experimentos de (GHAZI; INKPEN; SZPAKOWICZ, 2010) X (DOSCIATTI, 2015) não era suficiente para rejeitar a hipótese nula. Uma vez aplicados os ajustes do método o p-value passou a rejeitar a hipótese nula do *t-student* concluindo que para a precisão os ajustes foram de uma boa significância. Para constar, a variação dos p-values foi de -0,23 negativo o que indica que o ajuste realmente melhorou a precisão. Para a medida de Cobertura e F1 o resultado não foi tão relevante quanto a precisão. Os *p-values* ainda não conseguiram chegar em um valor suficiente para comprovar que os ajustes foram significantes, porém ao destacar a variação dos p-values, tudo indica que houve melhorias importantes na classificação. Para a cobertura a variação foi de -0,1 e para a medida F1 foi de -0,14 negativo.

Embora o método disponibilize o algoritmo FP-Align para identificar padrões na fase de treinamentos, foram feitos experimentos com algoritmos existentes na literatura e descritos no Capitulo 3. Na fase de treinamento, na etapa de *Extrair Padrões* ao invés de utilizar o FP-Align para esta extração foi utilizado o *FP-Growth* e *PrefixSpan.* Devido ás características do método proposto, as saídas originais destes algoritmos foram adaptadas para que disponibilizassem o *Banco de Padrões* definidos pelo método.

Na Tabela 8.10 é possível visualizar os resultados de cada algoritmo que substituiu o *FP-Align*.

Tabela 8.10: Resultados dos experimentos com algoritmos similares ao FP-Align.

Classe	(DOSCIATTI, 2015)			Método (FP-Growth)			Método (PrefixSpan)			**Método (FP-Align)**		
	P	C	F	P	C	F	P	C	F	P	C	F
Happy	0.37	0.41	0,39	0.66	0.82	0.7	0.64	0.82	0.72	0.55	0.61	0,57
Sad	0.66	0.5	0,58	0.69	0.60	0.64	0.70	0.58	0.63	0. 79	0.55	0,61
Angry-Disgusted	0.64	0.82	0,72	0.41	0.5	0.45	0.39	0.45	0.42	0.68	0.81	0,75
Fearful	0.69	0.60	0,64	0.69	0.5	0.58	0.69	0.5	0.58	0.68	0.64	0,63
Surprised	0.6	0.14	0,22	0.75	0.14	0.23	0.75	0.14	0.23	0.73	0.13	0,24
Acurácia	**61,40%**			**62.7%**			**61.7%**			**65.40 %**		

Os experimentos foram realizados com os três tipos de algoritmos para extração de padrões. A Tabela 8.10 demostra que a acurácia da classificação original (DOSCIATTI, 2015) ainda permanece menor comparado com o método proposto, mesmo não utilizando o FP-Align.

Foi realizado o teste estatístico t-student para um comparativo entre as medidas de desempenho do método variando com as variações dos algoritmos de extração de padrões e o método de Dosciatti. Na Tabela 8.11 é possível visualizar os valores dos p-values de cada experimento.

Tabela 8.11: p-values dos experimentos de algoritmos semelhantes ao FP-Align.

MEDIDA	**MÉTODO**		***P – VALUE***
PRECISÃO	(DOSCIATTI,2015	METODO (FP-GROWTH)	0.29
	(DOSCIATTI,2015	METODO (PrefixSpan)	0.32

	(DOSCIATTI,2015	METODO (FP-Align)	0.13
COBERTURA	(DOSCIATTI,2015	METODO (FP-GROWTH)	0.45
	(DOSCIATTI,2015	METODO (PrefixSpan)	0.49
	(DOSCIATTI,2015	METODO (FP-Align)	0.39
F1	(DOSCIATTI,2015	METODO (FP-GROWTH)	0.47
	(DOSCIATTI,2015	METODO (PrefixSpan)	0.48
	(DOSCIATTI,2015	METODO (FP-Align)	0.35

Os p-values de cada experimento tem valores próximos um dos outros, porém é possível observar que para todas as medidas o FP-Align conseguiu ter p-values melhores que os outros algoritmos. Isto significa que para esta base os ajustes realizados pelo FP-Align têm maior significado estatístico que comparado com os outros algoritmos.

Outro aspecto dos resultados que é possível identificar são os números de correções realizadas por cada método e algoritmo de extração de padrão. Para relembrar, é possível que o método proposto neste trabalho insira erros de classificação, uma vez que são aplicados padrões nos pontos de confusão e estes padrões ajustam as transações. Na Figura 8.4 é possível observar a quantidade de correções e erros de cada (método + algoritmo).

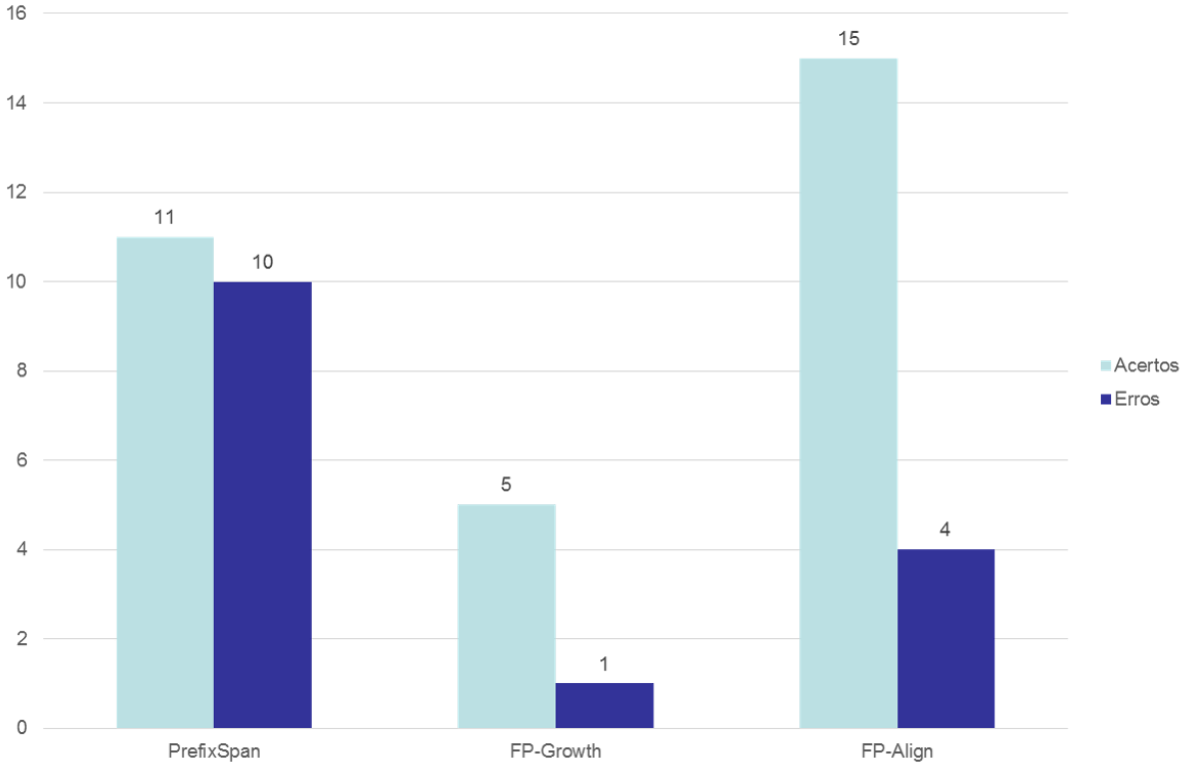

Figura 8.4: Quantidade de erros e acertos inseridos na classificação.

A quantidade de correções para todos os algoritmos é superior a inclusão de erros e isto indica que o método alcançou seu objetivo. *PrefixSpan* e *FP-Align* obtiveram resultados de correção mais próximos entre si que o *FP-Growth*. O efeito colateral do método está expresso nos números de erros inseridos. O método *PrefixSpan* foi o algoritmo que mais inseriu erros na classificação de (DOSCIATTI, 2015). O método *FP-Align* obteve melhor acurácia perante os outros métodos não só porque fez o maior número de correções, mas também porque não inseriu novos erros de classificação como *PrefixSpan*.

8.6. Análise dos parâmetros

O ajuste dos parâmetros foi de forma manual, o que quer dizer que não foram utilizadas técnicas de *hyperparameter* ou algoritmos da natureza *gridsearch* para identificar os melhores valores para os parâmetros. Porém, os experimentos realizados seguiram critérios que poderão ser explicados a seguir.

Por se tratar de vários parâmetros, é importante lembrar que eles foram testados simultaneamente com muitas combinações, o que implica em ser uma relação N x N entre eles. E o produto desta relação N x N é a quantidade de experimentos necessários para chegar nos valores com maior precisão, cobertura, f1 e acurácia.

Os parâmetros utilizados no método foram criados para, primeiramente, atender aos critérios dos algoritmos de MPS. O parâmetro Suporte Mínimo, que já foi explicado em seções anteriores, foi utilizado em um intervalo de 3 a 60 % , o que quer dizer, que a cada experimento os padrões estariam em pelo menos 3 % das sequencias do banco de sequencias e em outros experimentos no mínimo 60 % das sequencias. Para evitar um número ainda exaustivo de experimentos, o parâmetro foi inserido no método somando sempre 4 no número anterior.

Na Figura 8.5 é possível visualizar como o parâmetro foi modificado e os resultados da acurácia sendo modificada com relação ás mudanças.

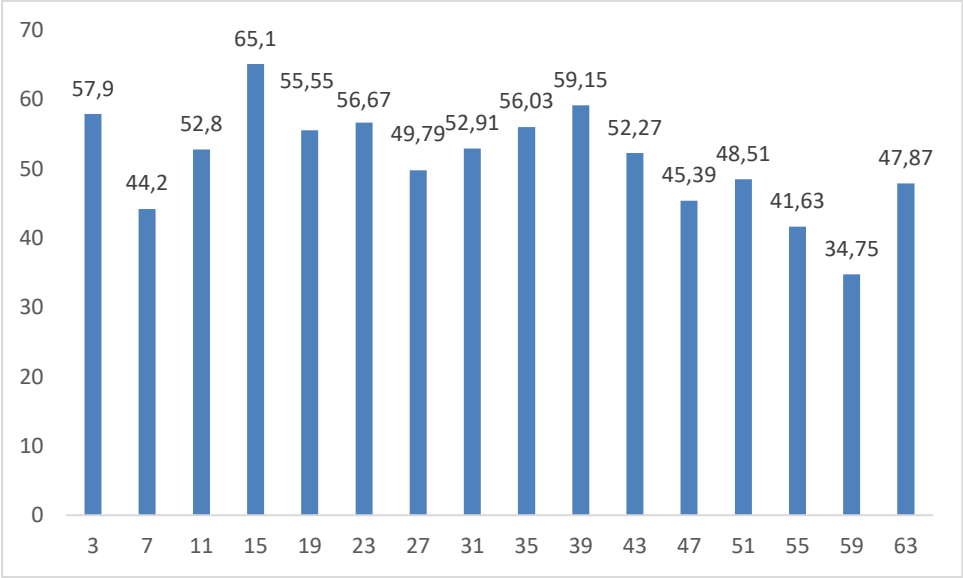

Figura 8.5: Acurácia por Suporte Mínimo.

Pelos valores da Figura 8.5 os resultados de acurácia não seguem uma regra específica quando modificado o parâmetro Suporte Mínimo. Ao calcular a correlação entre estas duas séries (acurácia e suporte mínimo) o valor foi de -0,57 o que dizer ter uma correção moderada entre elas.

O parâmetro de confusão é utilizado neste método na parte de classificação. Ele também é possível ser testado em intervalo. Nos experimentos ele foi usado com valor 0.05 até 0.4 pulando de 0,05 em 0,05. Na Figura 8.6 é possível visualizar a acurácia e as modificações de valor deste parâmetro.

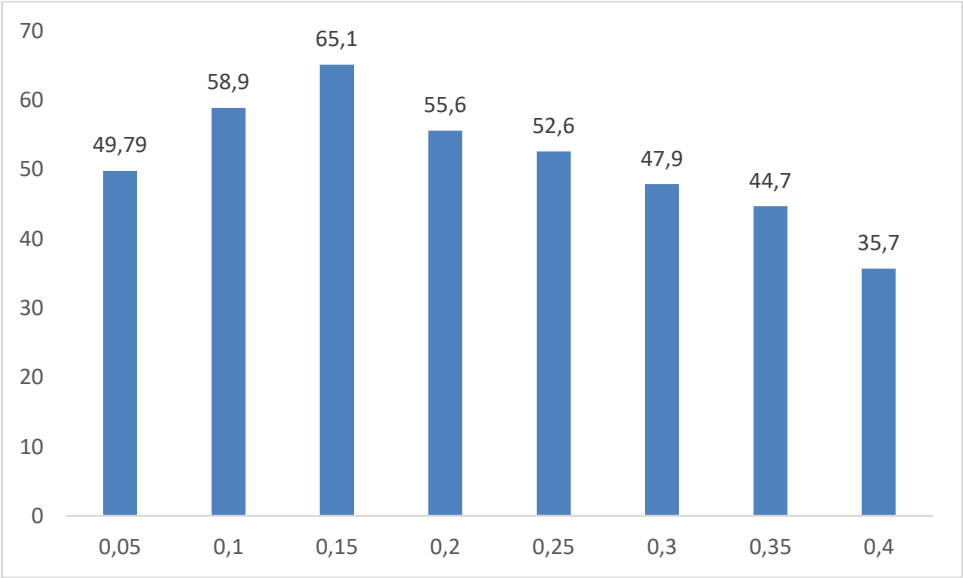

Figura 8.6: Acurácia por confusão.

É possível observar que quanto mais aumenta o valor de confusão pior fica a acurácia. O parâmetro de confusão sendo maior, aumenta as possibilidades de emoções para cada ponto de confusão o que fez o método não aplicar padrões corretamente.

8.7. Conclusão

Este capítulo apresentou os resultados dos experimentos feitos com o método. Os experimentos utilizaram uma base de contos infantis com o subconjunto utilizado por Ghazi e colegas (GHAZI; INKPEN; SZPAKOWICZ, 2010) para efeitos de comparação dos resultados. Foram apresentados os resultados gerados por cada uma das etapas das fases que compõem o método e ao final comparado os resultados com a classificação do método da Mariza (DOSCIATTI, 2015) e Ghazi e colegas (GHAZI; INKPEN; SZPAKOWICZ, 2010). As medidas de desempenho que foram utilizadas para medir foram: precisão, cobertura, F1 e acurácia. E para analisar estatisticamente os resultados destas medidas foi aplicado o teste *t-student*, onde foi possível observar a variação do *p-value* e provar que o método proposto obteve resultados interessantes. Ao final foi colocado em evidencia a quantidade de correções e inserções de erros do método com cada algoritmo de extração de padrões e descritas conclusões sobre este fato.

Considerações Finais

Classificar emoções vem sendo, ao longo dos anos, um desafio crescente. A identificação de emoções em texto, proporciona diversas aplicações no mercado, sobretudo para aquelas empresas que desejam conhecer a emoção de seus clientes sobre seus produtos. Também, proporciona conhecimento e pesquisa para a academia, instigando os pesquisadores a melhorarem medidas desempenho constantemente.

Unir a área de Análise de Sentimentos, identificando emoções em textos, com o auxílio da área de MPS proporciona um novo leque de pesquisas. Algoritmos comumente utilizados em outras áreas que trabalham com sequencias, como biologia Molecular, aplicados na AS, propicia um caminho de investigação entre as duas áreas de conhecimento.

Neste trabalho foi proposto um método de identificação de emoções baseado em Mineração de Padrões Sequenciais. Na fase de treinamento foi desenvolvida com o objetivo de validar a primeira hipótese que afirma que é possível extrair padrões sequenciais a partir de textos rotulados com emoções. Foi identificada uma base de contos infantis, onde as rotulações da base seguiam a ordem cronológica do texto o que possibilitou gerar padrões com *FP-Growth, PrefixSpan* e *FP-Align*.

A segunda fase do método proposto foi desenvolvida para validar a segunda hipótese afirma que que é possível melhorar a identificação de emoções em textos utilizando padrões sequenciais obtidos através da rotulação. Os experimentos foram realizados com a mesma base, utilizando diversos métodos para comparar com os resultados do método proposto. Nos experimentos apresentados, o método proposto utilizando o algoritmo *FP-Align* obteve o melhor resultado comparado com outros algoritmos (*FP-Growth* e *PrefixSpan*). O método foi comparado com outros métodos (GHAZI; INKPEN; SZPAKOWICZ, 2010) e *Stanford* que já trabalharam com a base e obteve também melhores resultados.

Os testes estatísticos que foram executados comprovaram que o método obteve na medida Precisão um resultado significativo de melhora comparado com Ghazi e colegas (GHAZI; INKPEN; SZPAKOWICZ, 2010).

A pesquisa deste trabalho tem sequência melhorando o número de inserções de erros

nas substituições de padrões na fase de classificação. Utilizar melhores descritores de padrões como: Mineração de Episódios, para ajudar na melhora dos resultados. É importante salientar que são necessárias novas bases para que o método possa ser testado. Em direção à generalização do método realizar experimentos com novas bases que não seguem uma ordem cronológica na rotulação para identificar se o método se comporta bem nestas condições.

Referências Bibliográficas

ALM, C. O.; ROTH, D.; SPROAT, R. EMOTIONS FROM TEXT : MACHINE LEARNING FOR TEXT-BASED EMOTION PREDICTION. HUMAN LANGUAGE TECHNOLOGY CONFERENCE/CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, P. 579–586, 2005.

A. Gomariz, M. Campos, R. Marin, and B. Goethals, "ClaSP: An efficient algorithm for mining frequent closed sequences," The Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 50–61, 2013.

AMAN, S.; SZPAKOWICZ, S. IDENTIFYING EXPRESSIONS OF EMOTION IN TEXT. TEXT, SPEECH AND DIALOGUE, V. 4629, P. 196–205, 2007.

ARNOLD, M. B. EMOTION AND PERSONALITY. VOL. I. PSYCHOLOGICAL ASPECTS. EMOTION AND PERSONALITY PSYCHOLOGICAL ASPECTS, V. 1, 1960.

BALAHUR, A. ET AL. SENTIMENT ANALYSIS IN THE NEWS. PROCEEDINGS OF THE SEVENTH INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC'10), P. 2216–2220, 2010.
BALAHUR, A.; HERMIDA, J. M.; MONTOYO, A. BUILDING AND EXPLOITING EMOTINET, A KNOWLEDGE BASE FOR EMOTION DETECTION BASED ON THE APPRAISAL THEORY MODEL. IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, V. 3, N. 1, P. 88–101, 2012.

BALAHUR, A.; MIHALCEA, R.; MONTOYO, A. COMPUTATIONAL APPROACHES TO SUBJECTIVITY AND SENTIMENT ANALYSIS: PRESENT AND ENVISAGED METHODS AND APPLICATIONS. COMPUTER SPEECH AND LANGUAGE, V. 28, N. 1, P. 1–6, 2014.
BAYARDO, R. EFFICIENTLY MINING LONG PATTERNS FROM DATABASES. PROCEEDINGS OF THE ACM SIGMOD, SEATTLE, WA, 1998

BOGUSKI, M.S. (1998). BIOINFORMATICS - A NEW ERA. TRENDS GUIDE BIOINFORMATICS, 1-3.

BR, A.; JONASSEN, I.; EIDHAMMER, I.; GILBERT, D.: APPROACHES TO THE AUTOMATIC DISCOVERY OF PATTERNS IN BIOSEQUENCES. JOURNAL OF COMPUTATIONAL BIOLOGY 5(2): 277-304 (1998)

BREJOVA, B.; DIMARCO, C.; VINAR, T.; HIDALGO, S.R.; HOLGUIN, G.; PATTEN, C. : FINDING PATTERNS IN BIOLOGICAL SEQUENCES. PROJECT REPORT, DEPARTMENT OF BIOLOGY, UNIVERSITY OF WATERLOO, 2000.

C. C. Aggarwal, Data mining: the textbook, Heidelberg:Springer, 2015.

Carletta, J. (1996). Assessing Agreement on Classification Tasks: The Kappa Statistic. Computational Linguistic, Vol. 22, N. 2, pp. 249-254.

D. SCHWEIZER, M. ZEHNDER, H. WACHE, H. F. WITSCHEL, D. ZANATTA, AND M. RODRIGUEZ, "USING CONSUMER BEHAVIOR DATA TO REDUCE ENERGY CONSUMPTION IN SMART HOMES: APPLYING MACHINE LEARNING TO SAVE ENERGY WITHOUT LOWERING COMFORT OF INHABITANTS," IEEE INTERNATIONAL CONFERENCE ON MACHINE LEARNING AND APPLICATIONS, PP. 1123–1129, 2015.

DAMASIO, A. R. AO ENCONTRO DE ESPINOSA: AS EMOÇÕES SOCIAIS E A NEUROLOGIA DO SENTIR. 2003.

DARWIN, C. THE EXPRESSION OF THE EMOTIONS IN MAN AND ANIMALS. 1872.

DAVE, K. ET AL. MINING THE PEANUT GALLERY: OPINION EXTRACTION AND SEMANTIC CLASSIFICATION OF PRODUCT REVIEWS. PROCEEDINGS OF THE 12TH INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, P. 519–528, 2003

DAVIDSON, R. J. AFFECTIVE NEUROSCIENCE AND PSYCHOPHYSIOLOGY: TOWARD A SYNTHESIS. PSYCHOPHYSIOLOGY, V. 40, N. 5, P. 655–665, 2003.

DOSCIATTI, M. M.; FERREIRA, L. P. C. ; PARAISO, E. C. . ANOTANDO UM CORPUS DE NOTÍCIAS PARA A AS: UM RELATO DE EXPERIÊNCIA. IN: PROCEEDINGS OF SYMPOSIUM IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY, 2015. PROCEEDINGS OF SYMPOSIUM IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY. P. 121-130.

DOSCIATTI, M. M.; FERREIRA, L. P. C. ; PARAISO, E. C. . IDENTIFICANDO EMOÇÕES EM TEXTOS EM PORTUGUÊS BRASILEIRO USANDO MÁQUINA DE VETORES DE SUPORTE EM SOLUÇÃO MULTICLASSE.. IN: ENCONTRO NACIONAL DE INTELIGÊNCIA ARTIFICIAL E COMPUTACIONAL (ENIAC), 2013, FORTALEZA. ENCONTRO NACIONAL DE INTELIGÊNCIA ARTIFICIAL E COMPUTACIONAL (ENIAC), 2013. P. 1-12.

E. SALVEMINI, F. FUMAROLA, D. MALERBA, AND J. HAN, "FAST SEQUENCE MINING BASED ON SPARSE ID-LISTS," THE INTERNATIONAL SYMPOSIUM ON METHODOLOGIES FOR INTELLIGENT SYSTEMS, PP. 316–325, 2011.

EKMAN, P. AN ARGUMENT FOR BASIC EMOTIONS. COGNITION & EMOTION, V. 6, N. 3, P. 169–200, 1992.

EKMAN, P. FACIAL EXPRESSION AND EMOTION. THE AMERICAN PSYCHOLOGIST, V. 48, N. 4, P. 384– 392, 1993.

ESULI, A.; SEBASTIANI, F. SENTIWORDNET: A PUBLICLY AVAILABLE LEXICAL RESOURCE FOR OPINION MINING. PROCEEDINGS OF THE 5TH CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, P. 417–422, 2006.

FEHR, B.; RUSSELL, J. A. CONCEPT OF EMOTION VIEWED FROM A PROTOTYPE PERSPECTIVE. JOURNAL OF EXPERIMENTAL PSYCHOLOGY: GENERAL, V. 113, N. 3, P. 464–486, 1984.

FELLOUS, J. M. FROM HUMAN EMOTIONS TO ROBOT EMOTIONS. 2004 AAAI SPRING SYMPOSIUM, ARCHITECTURES FOR MODELING EMOTION: CROSS-DISCIPLINARY FOUNDATIONS, V. TECHNICAL, P. 37, 2004.

FLORATOS, A.: PATTERN DISCOVERY IN BIOLOGY: THEORY AND APPLICATIONS. PH.D. THESIS, DEPARTMENT OF COMPUTER SCIENCE, NEW YORK UNIVERSITY, JAN. 1999

FRIJDA, N. H. THE EMOTIONS. THE EMOTIONS, P. 544, 1986.

GHAZI, D.; INKPEN, D.; SZPAKOWICZ, S. PRIOR AND CONTEXTUAL EMOTION OF WORDS IN SENTENTIAL CONTEXT. COMPUTER SPEECH & LANGUAGE, V. 28, N. 1, P. 76–92, JAN. 2014.

GRAY, J. A. NEUROPSYCHOLOGY OF ANXIETY. 1982

GROSSMAN, D. A.; FRIEDER, O. INFORMATION RETRIEVAL ALGORITMOS E HEURÍSTICA. 2004.

GUSFIELD, D. ALGORITHMS ON STRINGS, TREES AND SEQUENCES: COMPUTER SCIENCE AND COMPUTATIONAL BIOLOGY. [S.1.]: CUP, 1997.

HAGEN, J.B. (2000). THE ORIGINS OF BIOINFORMATICS. NATURE REVIEWS 1, 231-236

HALL, P. A.; DOWLING, G. R. APPROXIMATE STRING MATCHING. ACM COMPUTING SURVEYS, V. 12, N.4, P. 381-402, 1980

HAN, J.; PEI, J.; MORTAZAVI-ASL, B.; CHEN, Q.; DAYAL, U.; HSU, M. FREESPAN: FREQUENT PATTERN-PROJECTED SEQUENTIAL PATTERN MINING. IN PROCEEDINGS OF THE SIXTH ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING (KDD'2000), BOSTON, MA, 2000

HAN, J.; PEI, J.; YIN, Y. MINING FREQUENT PATTERNS WITHOUT CANDIDATE GENERATION. APRESENTAÇÃO EM DATA MINING AND KNOWLEDGE DISCOVERY: AN INTERNATIONAL JOURNAL, KLUWER ACADEMIC PUBLISHERS, 1999.

HUNT, J. W.;SZYMANSKY, T. AN ALGORITHM FOR DIFFERENTIAL FILE COMPARISON COMMUNICATIONS OF THE ACM, V. 20, P. 350-353, 1977.

IZARD, C. E. THE FACE OF EMOTION. P. 468, 1971.

J. AYRES, J. FLANNICK, J. GEHRKE, AND T. YIU, "SEQUENTIAL PATTERN MINING USING A BITMAP REPRESENTATION," ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, PP. 429–435, 2002.

J. HAN, J. PEI, AND M. KAMBER, DATA MINING: CONCEPTS AND TECHNIQUES, AMSTERDAM:ELSEVIER, 2011.

J. HAN, J. PEI, B. MORTAZAVI-ASL, Q. CHEN, U. DAYAL, AND M. C. HSU, "FREESPAN: FREQUENT PATTERNPROJECTED SEQUENTIAL PATTERN MINING," ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, PP. 355–359, 2000.

J. HAN, J. PEI, Y. YING, AND R. MAO, "MINING FREQUENT PATTERNS WITHOUT CANDIDATE GENERATION: A FREQUENT-PATTERN TREE APPROACH," DATA MINING AND KNOWLEDGE DISCOVERY, VOL. 8(1), PP. 53–87, 2004.

J. LIN, E. KEOGH, L. WEI, AND S. LONARDI, "EXPERIENCING SAX: A NOVEL SYMBOLIC REPRESENTATION OF TIME SERIES," DATA MINING AND KNOWLEDGE DISCOVERY, VOL. 15(2), PP. 107–144, 2007.

J. M. Pokou, P. Fournier-Viger, and C. Moghrabi, "Authorship attribution using small sets of frequent part-of-speech skip-grams," The International Florida Artificial Intelligence Research Society Conference, pp. 86–91, 2016.

J. Pei, J. Han, B. Mortazavi-Asl, J. Wang, H. Pinto, Q. Chen, U. Dayal, and M. C. Hsu, "Mining sequential patterns by pattern-growth: The prefixspan approach," IEEE Transactions on knowledge and data engineering, vol. 16(11), pp. 1424–1440, 2004.

J. Pei, J. Han, H. Lu, S. Nishio, S. Tang, and D. Yang, "H-mine: Hyper-structure mining of frequent patterns in large databases," IEEE International Conference on Data Mining, pp. 441–448, 2001.

J. Wang, J. Han, and C. Li, "Frequent closed sequence mining without candidate maintenance," IEEE Transactions on Knowledge Data Engineering, vol. 19(8), pp. 1042–1056, 2007.

JAMES, W. WHAT IS AN EMOTION? MIND, V. 9, N. 34, P. 188–205, 1884.

K. Gouda, M. Hassaan, and M. J. Zaki, "Prism: An effective approach for frequent sequence mining via prime-block encoding," Journal of Computer and System Sciences, vol. 76(1), pp. 88–102, 2010.

K. Y. Huang, C. H. Chang, J. H. Tung, and C. T. Ho, "COBRA: closed sequential pattern mining using bi-phase reduction approach," The International Conference on Data Warehousing and Knowledge Discovery, pp. 280–291, 2006.

KAGAN, J. BEHAVIORAL INHIBITION AS A TEMPERAMENTAL CATEGORY. IN: HANDBOOK OF AFFECTIVE SCIENCES. NEW YORK: OXFORD UNIVERSITY PRESS, 2003. P. 320–331.

LANG, P. J. THE EMOTION PROBE. AMERICAN PSYCHOLOGIST ASSOCIATION, V. 50, N. 5, P. 372–385, 1995.

LAZARUS, R. S. EMOTION E ADAPTATION. OXFORD UNIVERSITY PRESS, 1991.

LIBRALON, G. L. MODELAGEM COMPUTACIONAL PARA RECONHECIMENTO DE EMOÇÕES BASEADA NA ANÁLISE FACIAL. TESE, 2014.

LIU, B. OPINIONS, SENTIMENT, AND EMOTION IN TEXT. CAMBRIDGE UNIVERSITY PRESS, P. 381, 2015

LIU, B. SENTIMENT ANALYSIS AND OPINION MINING. CAMBRIDGE UNIVERSITY PRESS, 2012.

M. De Martino, A. Bertone, R. Albertoni, H. Hauska, U. Demsar, M. Dunkars. Technical Report of Data Mining, INVISIP IST-2000-29640, Information Visualisation for Site Planning, WP No2: Technology Analysis, D2.2, 28.2.2002

M. J. Zaki, "Scalable algorithms for association mining," IEEE Transactions on Knowledge and Data Engineering, vol. 12(3), pp. 372–390, 2000.

M. J. Zaki, "SPADE: An efficient algorithm for mining frequent sequences," Machine

LEARNING, VOL. 42(1-2), PP. 31–60, 2001.

MCDOUGALL, W. AN INTRODUCTION TO SOCIAL PSYCHOLOGY. 1926.

MEHRABIAN, A. FRAMEWORK FOR A COMPREHENSIVE DESCRIPTION AND MEASUREMENT OF EMOTIONAL STATES. GENETIC, SOCIAL, AND GENERAL PSYCHOLOGY MONOGRAPHS, V. 121, N. 3, P. 339– 361, 1995.

MICHAEL S. GAZZANIGA; T; HEATHERTON, ODD F. PSICOLÓGICA: MENTE, CÉREBRO E COMPORTAMENTO. 2005.

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, 2012B.

MORAES, R.; VALIATI, J. F.; GAVIÃO NETO, W. P. DOCUMENT-LEVEL SENTIMENT CLASSIFICATION: AN EMPIRICAL COMPARISON BETWEEN SVM AND ANN. EXPERT SYSTEMS WITH APPLICATIONS, V. 40, N. 2, P. 621–633, 2013.

MOWRER, O. H. LEARNING THEORY AND BEHAVIOR. 1960.

MUNEZERO, M. ET AL. ARE THEY DIFFERENT? AFFECT, FEELING, EMOTION, SENTIMENT, AND OPINION DETECTION IN TEXT. IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, V. 5, N. 2, P. 101–111, 2014.

N. R. MABROUKEH, AND C. I. EZEIFE, "A TAXONOMY OF SEQUENTIAL PATTERN MINING ALGORITHMS," ACM COMPUTING SURVEYS, VOL. 43(1), 2010.

NASUKAWA, T. SENTIMENT ANALYSIS: CAPTURING FAVORABILITY USING NATURAL LANGUAGE PROCESSING DEFINITION OF SENTIMENT EXPRESSIONS. 2ND INTERNATIONAL CONFERENCE ON KNOWLEDGE CAPTURE, P. 70–77, 2003

NEEDLEMAN, S. B.; WUNSCH, C. D. A GENERAL METHOD APPLICABLE TO THE SEARCH FOR SIMILARITIES IN THE AMINO ACID SEQUENCE OF TWO PROTEINS. J. MOL. BIOL., V. 48, P. 433-453, 1970.

NIJHOLT, A. HUMOR AND EMBODIED CONVERSATIONAL AGENTS. CTIT TECHNICAL REPORT SERIES NO. 03-03, 2003..

OATLEY, K.; JOHNSON-LAIRD, P. N. TOWARDS A COGNITIVE THEORY OF EMOTIONS. COGNITION & EMOTION, V. 1, N. 1, P. 29–50, 1987.

ORTONY, A. ON MAKING BELIEVABLE EMOTIONAL AGENTS BELIEVABLE. EMOTIONS IN HUMANS AND ARTIFACTS, P. 189, 2002.

ORTONY, A.; CLORE, G. L.; COLLINS, A. THE COGNITIVE STRUCTURE OF EMOTIONS. NEW YORK: CAMBRIDGE UNIVERSITY PRESS, 1988.

ORTONY, A.; NORMAN, D.; REVELLE, W. AFFECT AND PROTO-AFFECT IN EFFECTIVE FUNCTIONING. IN: WHO NEEDS EMOTIONS? THE BRAIN MEETS THE ROBOT. [S.L.] OXFORD UNIVERSITY PRESS, 2004.

ORTONY, A.; TURNER, T. J. WHAT'S BASIC ABOUT BASIC EMOTIONS? PSYCHOLOGICAL REVIEW, V. 97, N. 3, P. 315–331, 1990

P. Fournier-Viger, A. Gomariz, M. Campos, and R. Thomas, "Fast Vertical Mining of Sequential Patterns Using Co-occurrence Information," The Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 40–52, 2014.

P. Fournier-Viger, A. Gomariz, M. Sebek, M. Hlosta, "VGEN: fast vertical mining of sequential generator patterns," The International Conference on Data Warehousing and Knowledge Discovery, pp. 476–488, 2014.

P. Fournier-Viger, A. Gomariz, T. Gueniche, E. Mwamikazi, and R. Thomas, "TKS: Efficient Mining of Top-K Sequential Patterns," The International Conference on Advanced Data Mining and Applications, pp. 109–120, 2013.

P. Fournier-Viger, C.-W. Wu, A. Gomariz, and V. S. Tseng, "VMSP: Efficient vertical mining of maximal sequential patterns," The Canadian Conference on Artificial Intelligence, pp. 83–94, 2014.

P. Fournier-Viger, C.-W. Wu, and V. S. Tseng, "Mining Maximal Sequential Patterns without Candidate Maintenance," The International Conference on Advanced Data Mining and Applications, pp. 169–180, 2013.

P. Fournier-Viger, R. Nkambou, and E. Mephu Nguifo, "A Knowledge discovery framework for learning task models from user interactions in intelligent tutoring systems," The Mexican International Conference on Artificial Intelligence, pp. 765–778, 2008.

P. Fournier-Viger, T. Gueniche, and V. S. Tseng, "Using partially-ordered sequential rules to generate more accurate sequence prediction, The International Conference on Advanced Data Mining and Applications, pp. 431–442, 2012.

PANG, B.; LEE, L. OPINION MINING AND SENTIMENT ANALYSIS. FOUNDATIONS AND TRENDS IN INFORMATION RETRIEVAL, V. 2, N. 1–2, P. 1–135, 2008.

PANKSEPP, J. AFFECTIVE NEUROSCIENCE: THE FOUNDATIONS OF HUMAN AND ANIMAL EMOTIONS. P. 480, 1998.

PANKSEPP, J. TOWARD A GENERAL PSYCHOBIOLOGICAL THEORY OF EMOTIONS. BEHAVIORAL AND BRAIN SCIENCES, V. 5, N. 03, P. 407, 1982.

PEI, J.; HAN, J.; LU, H.; NISHIO, S.; TANG, D.; YANG, D. H-MINE: HYPER-STRUCTURE MINING OF FREQUENT PATTERNS IN LARGE DATABASES. IN PROCEEDINGS OF THE 2001 IEEE INTERNATIONAL CONFERENCE ON DATA MINING (ICDM'01), SAN JOSE, CALIFORNIA, 2001B.

PEI, J.; HAN, J.; MORTAZAVI-ASL, B.; PINTO, H. PREFIXSPAN: MINING SEQUENTIAL PATTERNS EFFICIENTLY BY PREFIX-PROJECTED PATTERN GROWTH. IN PROCEEDINGS OF THE 2001 INTERNATIONAL CONFERENCE ON DATA ENGINEERING (ICDE'01), HEIDELBERG, GERMANY, 2001

PICARD, R. W. AFFECTIVE COMPUTING. MIT MEDIA LABORATORY PERCEPTUAL COMPUTING SECTION, N. 321, 1995.

PINTO, H.; HAN, J.; PEI, J.; WANG, K. MULTI-DIMENSIONAL SEQUENTIAL PATTERN MINING, PROCEEDINGS OF THE 27TH INTERNATIONAL CONFERENCE ON VERY LARGE DATA BASE (VLDB'01), ROMA, ITALY, 2001

PLUTCHIK, R. A GENERAL PSYCHOEVOLUTIONARY THEORY OF EMOTION. EMOTION: THEORY, RESEARCH, AND EXPERIENCE, V. 1, P. 3–33, 1980.

R. AGRAWAL AND R. SRIKANT, "FAST ALGORITHMS FOR MINING ASSOCIATION RULES," THE INTERNATIONAL CONFERENCE ON VERY LARGE DATABASES, PP. 487–499, 1994.

R. AGRAWAL, AND R. SRIKANT, "MINING SEQUENTIAL PATTERNS," THE INTERNATIONAL CONFERENCE ON DATA ENGINEERING, PP. 3–14, 1995.

R. J. BAYARDO JR. EFFICIENTLY MINING LONG PATTERNS FROM DATABASES, ACM SIGMOD CONFERENCE, 1998.

R. SRIKANT, AND R. AGRAWAL, "MINING SEQUENTIAL PATTERNS: GENERALIZATIONS AND PERFORMANCE IMPROVEMENTS," THE INTERNATIONAL CONFERENCE ON EXTENDING DATABASE TECHNOLOGY, PP. 1–17, 1996.

ROLLS, E. T. THE BRAIN AND EMOTION. 1998.

ROMAN, N. T. EMOÇÃO E A SUMARIZAÇÃO AUTOMÁTICA DE DIÁLOGOS. TESE, 2007.

ROSEMAN, I. J. APPRAISAL DETERMINANTS OF EMOTIONS: CONSTRUCTING A MORE ACCURATE AND COMPREHENSIVE THEORY. COGNITION AND EMOTION, V. 10, N. 3, P. 241–278, 1996.

RUSSELL, J. A. A CIRCUMPLEX MODEL OF AFFECT. JOURNAL OF PERSONALITY AND SOCIAL PSYCHOLOGY, V. 39, N. 6, P. 1161–1178, 1980.

S. ASEERVATHAM, A. OSMANI, AND E. VIENNET, "BITSPADE: A LATTICE-BASED SEQUENTIAL PATTERN MINING ALGORITHM USING BITMAP REPRESENTATION," THE INTERNATIONAL CONFERENCE ON DATA MINING, PP. 792–797,2006.

S. ZIEBARTH, I. A. CHOUNTA, AND H. U. HOPPE, "RESOURCE ACCESS PATTERNS IN EXAM PREPARATION ACTIVITIES," THE EUROPEAN CONFERENCE ON TECHNOLOGY ENHANCED LEARNING, PP. 497–502, 2015.

SANGER, F. (1959). CHEMISTRY OF INSULIN; DETERMINATION OF THE STRUCTURE OF INSULIN OPENS THE WAY TO GREATER UNDERSTANDING OF LIFE PROCESSES. SCIENCE (NEW YORK, N.Y 129, 1340-1344.

SCHERER, K. R.; SCHORR, A.; JOHNSTONE, T. APPRAISAL PROCESSES IN EMOTION: THEORY, METHODS, RESEARCH. EUA: OXFORD UNIVERSITY PRESS, 2001.

SEBE, N.; LEW, M.; HUANG, T. THE STATE-OF-THE-ART IN HUMAN-COMPUTER INTERACTION. COMPUTER VISION IN HUMAN-COMPUTER INTERACTION, V. 2, P. 1–6, 2004.

SELLERS, P. H. THE THEORY AND COMPUTATION OF EVOLUTIONARY DISTANCES: PATTERN RECOGNITION. JOURNAL OF ALGORITHMS, V. 1, P. 359-373, 1980.

SETUBAL, J; MEIDANIS, J. INTRODUCTION TO COMPUTACIONAL MOLECULAR BIOLOGY.[S.1]:BROOKS-COLE, 1997. 320 P.

SMITH, C. A.; LAZARUS, R. S. APPRAISAL COMPONENTS, CORE RELATIONAL THEMES, AND THE EMOTIONS. COGNITION & EMOTION, V. 7, N. 3-4, P. 233–269, 1993.

SMITH, T; WATERNAN, M. IDENTIFICATION OF COMMON MOLECULAR SUBSEQUENCES. JOURNAL OF MOLECULAR BIOLOGY, V. 147, P. 195-197, 1981.

SRIKANT R. AGRAWAL, R. MINING SEQUENCIAL PATTERNS. 1995

SRIKANT, R.; AGRAWAL, R. MINING SEQUENTIAL PATTERNS GENERALIZATIONS AND PERFORMANCE IMPROVEMENTS. IN PROCEEDINGS OF THE FIFTH INT'L CONFERENCE ON EXTENDING DATABASE TECHNOLOGY (EDBT). AVIGNON, FRANCE, 1996.

STRAPPARAVA, C. ET AL. SEMEVAL-2007 TASK 14: AFFECTIVE TEXT. PROC. OF SEMEVAL-2007, N. JUNE, P. 70–74, 2007

STRAPPARAVA, C.; MIHALCEA, R. LEARNING TO IDENTIFY EMOTIONS IN TEXT. PROCEEDINGS OF THE 2008 ACM SYMPOSIUM ON APPLIED COMPUTING - SAC '08, P. 1556, 2008.

STRAPPARAVA, C.; VALITUTTI, A. WORDNET-AFFECT: AN AFFECTIVE EXTENSION OF WORDNET. PROCEEDINGS OF THE 4TH INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, P. 1083–1086, 2004

T. Uno, M. Kiyomi, and H. Arimura, "LCM ver. 2: Efficient mining algorithms for frequent/closed/maximal itemsets," IEEE International Conference on Data Mining Workshop on Frequent Itemset Mining Implementations, 2004

TOMKINS, S. S. AFFECT THEORY. APPROACHES TO EMOTION, P. 163–196, 1984.

U. Fayyad, G. P.-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. AI Magazine, 17(3):37-54, Fall 1996.

WATSON, D. ET AL. THE TWO GENERAL ACTIVATION SYSTEMS OF AFFECT: STRUCTURAL FINDINGS, EVOLUTIONARY CONSIDERATIONS, AND PSYCHOBIOLOGICAL EVIDENCE. JOURNAL OF PERSONALITY AND SOCIAL PSYCHOLOGY, V. 76, N. 5, P. 820–838, 1999.

WATSON, J. B. BEHAVIORISM. 1930.

WEINER, B.; GRAHAM, S. AN ATTRIBUTIONAL APPROACH TO EMOTIONAL DEVELOPMENT. IN: EMOTIONS, COGNITION, AND BEHAVIOR. NEW YORK: CAMBRIDGE UNIVERSITY PRESS, 1984. P. 167– 191.

WIEBE, J. M.; BRUCE, R. F.; O'HARA, T. P. DEVELOPMENT AND USE OF A GOLD STANDARD DATA SET FOR SUBJECTIVITY CLASSIFICATIONS. PROCEEDINGS OF THE 37TH ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (ACL'99), P. 246–253, 1999.

WU, S.; MANBER, C. FAST TEXT SEARCHING ALLOWING ERRORS. COMMUNICATION OS THE ACM, V. 35, P. 83-91, 1992.

Y. W. T. Pramono, "Anomaly-based Intrusion Detection and Prevention System on Website Usage using Rule-Growth Sequential Pattern Analysis," The International Conference on Advanced Informatics, Concept Theory and Applications, pp. 203–208, 2014.

Y. Xifeng, H. Jiawei, and R. Afshar, "CloSpan: Mining Closed Sequential Patterns in Large Data Base," SIAM International Conference on Data Mining, pp. 166–177, 2003.

Z. Yang, and M. Kitsuregawa, "LAPIN-SPAM: An improved algorithm for mining sequential pattern," The International Conference on Data Engineering Workshops, pp. 1222–1222, 2005.

ZAKI, M. SPADE: AN EFFICIENT ALGORITHM FOR MINING FREQUENT SEQUENCES, IN MACHINE LEARNING JOURNAL, SPECIAL ISSUE ON UNSUPERVISED LEARNING (DOUG FISHER, ED.), PP 31-60, VOL. 42 NOS. 1/2, 2001.

ZHANG, L.; LIU, B. IDENTIFYING NOUN PRODUCT FEATURES THAT IMPLY OPINIONS. PROCEEDINGS OF THE 49TH ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS HUMAN LANGUAGE TECHNOLOGIES, N. 2008, P. 575–580, 2011.