CENTRO UNIVERSITÁRIO FEI

CLAUDIO APARECIDO BORGES JUNIOR

PEL217 - ENGENHARIA DE SOFTWARE EM EXPERIMENTOS CIENTÍFICOS:

Projeto de Pesquisa - Um estudo sobre os métodos de treinamento de word embedding em diferentes domínios

São Bernardo do Campo

2020

SUMÁRIO

1 INTRODUÇÃO

O Processamento de Linguagem Natural (PLN) é um campo da Inteligência Artificial e Linguística que busca fazer com que computadores entendam sentenças ou palavras escritas em linguagem humana. Dessa forma, usuários não precisam aprender novas linguagens específicas de máquinas. Usuários podem, portanto, utilizar as linguagens de seu domínio para se comunicarem com os computadores (KHURANA et al., 2017).

Segundo Khurana et al. (2017), o PLN possui diversas aplicações. A tradução automática é uma aplicação do PLN e geralmente traduz frases de uma linguagem para outra com a ajuda de motores estatísticos. Outra aplicação do PLN é a categorização de textos, em que grandes quantidades de texto são inseridos nos sistemas de categorização e esses sistemas atribuem os textos a categorias pré-definidas. Os sistemas de categorização de textos são frequentemente utilizados na identificação de publicidade em massa recebida via correio eletrônico. A extração da informação também é uma área do PLN que busca identificar frases de interesse em textos escritos em linguagem humana. A extração de informação pode ser aplicada na identificação de palavras chave, preparação de índices e propaganda direcionada. Devido à grande quantidade de dados disponíveis, a sumarização tem tido um papel importante como aplicação do PLN. A sumarização não busca apenas reconhecer e entender as informações importantes em um texto, mas também compreender os significados emocionais. Os sistemas de diálogo, sejam eles por texto ou voz, também são uma aplicação do PLN previsto por grandes fornecedores de aplicativos de usuário final. Exemplos de sistemas de diálogo são o *Google assistant*, *Windows Cortana*, *Apple Siri* e *Amazon Alexa*.

Um dos desafios do PLN é a representação da palavras. Uma notória coleção de modelos de representação é conhecida como *word embedding*. Segundo Wang et al. (2019), *word embedding* é uma representação vetorial de palavras em números reais. Essa representação captura o significado semântico e sintático de grandes *corpora* não etiquetados. Portanto, essa representação busca quantificar e categorizar similaridades entre itens linguísticos baseados em suas propriedades distribucionais em grandes amostras de dados de uma linguagem.

Diferentes modelos de *word embedding* produzem representações vetoriais distintas. De acordo com Yaghoobzadeh e Schütze (2016), existem alguns critérios que os modelos devem buscar para prover uma melhor representação. Um é o de separação das evidências de contextos, já que cada contexto pode inferir atributos específicos e outro é a comunalidade de palavras raras em PLN. Modelos de *word embedding* devem aprender representações úteis baseadas com um

número pequeno de contextos, precisando também ser robustos contra ambiguidade. Além disso, eles devem ser capazes de representar corretamente os diversos domínios das palavras.

Vários métodos de avaliação foram propostos para testar as qualidades dos modelos de *word embedding*. Esses métodos são classificados em avaliações intrínsecas e extrínsecas. Os métodos de avaliação intrínseca são experimentos em que modelos de vetoriais são comparados com o julgamento humano. Um conjunto de palavras é normalmente utilizado e o julgamento sintático e semântico de palavras é comparado entre humanos e esses modelos. Os métodos de avaliação extrínseca utilizam essas representações como entrada de tarefas de PLN mais específicas. Os resultados de tais tarefas são comparados entre os diversos modelos de *word embedding* utilizando medidas específicas de suas aplicações (BAKAROV, 2018).

O aprendizado de representações vetoriais de alta qualidade é de extrema importância na área de PLN. Porém, a questão "o que é um bom modelo de *word embedding*?"continua um problema em aberto. Segundo Bakarov (2018) ainda não há um consenso na comunidade científica sobre quais métodos de avaliação devem ser utilizados.

A pergunta "o que é um bom modelo de *word embedding*?"pode ser analisada de diferentes formas. O presente trabalho considera que diferentes modelos de representação vetorial podem gerar resultados distintos em nas diversas aplicações de processamento de linguagem natural. Essa hipótese existe porque os modelos são concebidos considerando certos atributos da linguagem. Esses atributos podem estar presentes em aplicações específicas e portanto essas aplicações podem ser beneficiadas com a utilização de desses modelos.

Desse modo, o seguinte problema de pesquisa é definido: quais modelos de *word embedding* são indicados para os diferentes tipos de aplicações de processamento de linguagem natural e por que alguns modelos possuem melhores qualidades do que outros modelos em aplicações específicas mesmo sendo treinados com o mesmo *corpus*?

O objetivo principal deste trabalho é comparar os principais modelos de *word embedding* utilizando-os como entrada a aplicações de processamento de linguagem natural afim de avaliar quais modelos possuem melhores qualidades nessas aplicações.

O objetivo secundário do trabalho é identificar quais componentes da linguagem possuem maior influência nos principais modelos de *word embedding* e quais componentes da linguagem são mais relevantes nas diferentes tarefas de processamento de linguagem natural.

2 EXPERIMENTAÇÃO EM ENGENHARIA DE SOFTWARE

A revisão bibliográfica sobre *Experimentação em Engenharia de Software* foi realizada através de um resumo de cada capítulo de Wohlin et al. (2012). Os capítulos foram divididos em diferentes seções no presente trabalho, sendo que, cada seção possuí indicação do capítulo resumido.

2.1 CAPÍTULO 1 - INTRODUÇÃO

Software se tornou parte de vários produtos, desde torradeiras até carros autônomos, e tem sido desenvolvido extensamente. O processo de desenvolvimento não é fácil e é uma tarefa altamente criativa. Os produtos de software estão sujeito a falta de funcionalidades, sobre-custo, perda de prazo e baixa qualidade. O termo "engenharia de software"foi dado para descrever uma disciplina focada em sistemas de desenvolvimento de softwares intensivos.

O objetivo do livro é discutir como estudos empíricos e experimentação se encaixam no contexto de engenharia de software. Três aspectos são importante; a segmentação de processos de software dependendo das fases do ciclo de vida; a necessidade de métodos sistemáticos e disciplinados; e a quantificação.

A complexidade do desenvolvimento de produtos de software implica em uma complexidade nos processos de software, dificultando a otimização e a procura por um processo de software ótimos. Empresas buscam a melhora contínua, e a fazem também melhorando os processos de software, para que o desenvolvimento de software seja cada vez mais sistemático e disciplinado. Os processos de melhoria incluem duas atividades, avaliação do processo de software e avaliação da proposta de melhoria do processo de software. A primeira identifica possíveis áreas de melhoria, enquanto a segunda efetivamente propõe uma alteração no processo de software. Nesse caso, avaliar uma proposta de melhoria é importante antes que essa alteração ocorra, para que o processo não seja alterado sem que haja informações suficientes.

Avaliar uma proposta de melhoria não é uma tarefa fácil sem que haja um envolvimento humano direto. Para produtos, podemos construir um protótipo e verificar se esse protótipo pode ser evoluído e eventualmente se transformar em um resultado final desejado. Essa abordagem não é possível em processos; porém podemos fazer simulações e comparar os resultados de modelos propostos. O único modo de testar efetivamente os resultados seria ter pessoas usando o modelo. Desse modo, estudos empíricos são de grande importância para avaliar processos de

software, sendo que a experimentação prove um método sistêmico, disciplinado, quantificado e controlado de avaliar atividades humanas. Devemos então utilizar métodos e avaliar estratégias quando realizamos pesquisas em engenharia de software.

Engenharia de software é uma disciplina que se estende por várias áreas. O desenvolvimento de software é uma tarefa que exige recurso humano, não sendo possível produzir software como em uma linha de montagem. É uma disciplina baseada na criatividade e engenhosidade das pessoa. Todavia devemos tratá-la como uma disciplina científica, usando métodos científicos para pesquisa e para o desenvolvimento do software. Podemos reduzir a 4 o número de métodos de pesquisa no campo da engenharia de software, sendo eles: o método científico, onde o mundo é observado e um modelo é construído baseado na observação; o método de engenharia, onde a solução atual é estudada e mudanças são propostas, e então avaliadas; o método empírico, onde um modelo é proposto e avaliado através de estudos empíricos; e o analítico, em que há uma proposta de teoria formal que é comparada com observações empíricas. Tanto o método de engenharia quanto o método empírico são variações do método científico.

Tradicionalmente, o método analítico é utilizado em áreas mais formais como algoritmos ou eletromagnetismo. Já o método científico é usado em áreas com maior mais aplicadas, como simulação de redes de telecomunicações. A engenharia de software é intimamente relacionada com comportamentos humanos, isto ocorre porque fundamentalmente existem pessoas desenvolvendo software. Desse modo, regras formais não são comuns em engenharia de software, com exceção de aspectos estritamente técnicos. Portanto, a utilização de estudos empíricos é muita utilizada, como também é utilizada em ciências humanas. Isso não quer dizer que os outros métodos não se apliquem, mas cada qual possui sua especificidade e sua utilização.

Estrategias empíricas em engenharia de software incluem a criação de experimentos formais; estudos de projetos reais em empresas, ou seja, estudo de caso; e realização de pesquisas como, por exemplo, entrevistas. Estudos podem utilizar uma combinação dessas estratégias, e portanto existem similaridades e diferenças entre as estratégias.

A principal razão em utilizar experimentação em engenharia de software é para entender e identificar relações entre diferentes fatores. O aperfeiçoamento do entendimento é a base para mudança e melhoria do modo como trabalhamos, portanto estudos empíricos em geral e experimentação são importantes. O foco é prover um guia e suporte para realizar experimentos em engenharia de software. Além disso, deve ser ressaltado que experimentos verdadeiros, onde há uma completa randomização, são difíceis de serem realizados no campo de engenharia

de software. Experimentos nesse campo são quasi-experimentos, ou seja, onde participantes não são selecionados totalmente randômicos.

A intenção do livro é prover uma introdução ao estudo empírico e experimentação, ressaltando as oportunidades e benefícios em realizar experimentos em engenharia de software. A pesquisa empírica deve e pode ser utilizada com mais frequência em engenharia de software.

2.2 CAPÍTULO 2 - ESTRATÉGIAS EMPÍRICAS

A pesquisa exploratória estuda os objetos em sua condição natural e as descobertas emergem da observação. Um desenho de pesquisa flexível é necessário para se adaptar as mudanças. Esse tipo de pesquisa é também conhecida como qualitativa. A pesquisa indutiva tenta explicar fenômenos com base em explicações que pessoas fazem.

A pesquisa explicativa está principalmente preocupada em quantificar o relacionamento ou comparar dois ou mais grupos a fim de encontrar uma relação de causa e efeito. A pesquisa é feita em ambientes controlados onde alguns fatores são fixados antes do experimento produzindo uma pesquisa quantitativa de desenho fixo. Através da pesquisa quantitativa é possível realizar comparações e realizar analises estatísticas.

Tanto a pesquisa qualitativa quanto a quantitativa podem ser utilizadas para um mesmo assunto buscando responder perguntas de forma diferentes. Enquanto a quantitativa busca efeitos de um tratamento, a qualitativa buscar explicar o razão do resultado da pesquisa quantitativa.

Existem 3 grandes grupos de estratégia empíricas: pesquisa, estudo de caso e experimento ou quasi-experimento.

A pesquisa é normalmente uma investigação de algo passado onde se busca coletar dados quantitativos e qualitativos através de entrevistas e questionários. Os resultados são analisados e generalizados para toda a população da amostra da pesquisa. Desse modo, uma pesquisa amostral não ocorra para entender o comportamento de uma amostra da população, mas sim de toda a população.

As pesquisas podem ser descritivas, explicativas e explotarias. A pesquisa descritiva pode ser conduzida para habilitar afirmações sobre uma população, buscando entender qual a distribuição de um atributo e não a razão de sua existência. Pesquisa explicativa buscar explicar teorias de uma população através de relações entre o que se procura explicar e variáveis explicativas. Pesquisa exploratória é utilizada para ganhar mais conhecimento sobre um determinado

assunto e assim evitar situações não previstas. Assim ela não responde uma pergunta, mas sim prove mais possibilidades de analise.

O estudo de caso em engenharia de software é uma investigação empírica com várias fontes de evidências para investigar uma instância no mundo real onde o fenômeno é normalmente difícil de ser destacado em seu ambiente. Dados são coletados e então uma análise estatística é realizada. Onde normalmente se busca o efeito de uma atributo ou o relacionamento entre atributos. Os casos de uso são muito empregados em ambientes industriais.

O estudo de caso pode ser utilizado como comparação entre um resultado obtido e algo comparável. Para isso, é necessário existir uma base sólida de comparação. Essa base pode ser a comparação de um novo método contra uma linha base de uma empresa; ou a comparação de dois projetos, um utilizando um novo método enquanto o outro utilizando o método corrente; ou a aplicação do novo método em partes de um todo de forma aleatória, e comparando o os componentes onde o novo método foi empregado com os quais não o foram.

A principal diferença entre estudo de caso e experimento é que o experimento faz amostras variando normalmente um único fator por vez, enquanto o estudo de caso seleciona um conjunto de variáveis que normalmente representam uma situação. Assim, o estudo de caso é mais fácil de planejar a mais realistas, porém são mais difíceis de serem generalizados.

Experimentos em engenharia de software são investigações empíricas que manipulam um único fator ou variável da configuração do estudo. Desse modo, é possível entender o efeito que uma variável possui sobre um conjunto de fatores. Sendo que esse entendimento somente é possível quando há uma completa aleatoriedade. Porém, muitas vezes é difícil atingir uma perfeita aleatoriedade, produzindo os quasi-experimentos.

Experimentos são próprios para confirmar teorias, confirmar conhecimento comum, explorar relações, verificar a precisão de um modelo, validas medidas. Para isso, os experimentos seguem diferentes etapas, como levantamento de escopo, planejamento, operação, análise e interpretação, apresentação e consolidação.

Dependendo do controle sobre execução do experimento, do controle sobre a medição, do custo de investigação, e também da possibilidade desejada de replicação pode-se definir uma estratégia empírica de investigação. Pesquisas e experimentos são de possuem uma replicabilidade muito superior ao estudo de caso. Pesquisas são possuem um custo menor do que estudo de caso que possui um custo menor do que experimentos. Estudo de caso e experimentos possuem controle sobre a medição porém não há em pesquisa. E finalmente um experimento é a única estratégia que possui controle sobre a execução.

A replicação de um experimento envolve realizar novamente a investigação com uma população diferente. Se a aleatoriedade foi suficiente, replicar um experimento com uma amostra da população gerará o mesmo resultado, caso contrária, haverá diferenças devido a incapacidade do experimentos em descrever todos os aspectos de forma generalizada. A replicação pode se dar utilizando o mesmo processo que o experimento original, ou utilizando um processo diferente deliberadamente. Existem vantagens em cada método de replicação. Ao utilizar o mesmo procedimento, fatores conhecidos são controlados, aumentando a confiabilidade do resultado; porém existe uma grande tendência em um viés experimental. Ao utilizar um processo diferente deliberadamente, mais conhecimento sobre o tema pode ser adquirido.

Teoremas em engenharia de software fornecem o conceito mais básico sobre um dado mecanismo. Teoremas são bastante raros no campo de engenharia de software e não há muitos teoremas aceitos academicamente. Alguns autores dividiram o conhecimento em teoremas, hipóteses, e conjetura. Outros dividiram os teoremas em contextos como analise, explicação, predição, explicação e predição, e desenho e ação. Outros autores propuseram a criação de um framework para teoremas em engenharia de software com partes como construção, proposição, explicação e escopo. Porém nenhuma deses sistemas de teoremas tiveram impacto na engenharia de software até agora.

Agregar informações de múltiplos estudos empíricos é necessário para que haja construção do conhecimento. Algumas vezes, um único estudo empírico não é suficiente para responder certas perguntas, e sua agregação com outros experimentos adicionam solidez a conclusão. A agregação pode ser pode ser por uma revisão bibliográfica sistemática, em que se busca busca responder uma pergunta científica específica. Em casos onde o campo de pesquisa é mais geral e menos explorado, perguntas mais abrangentes podem ser utilizadas, buscando um mapeamento de estudos e consequentemente a situação atual das pesquisas e as tendências.

O capítulo do livro então apresenta um framework para avaliação de alterações de processos de software. Os requisitos básicos propostos são: compreensão do processo e do produto de software; definição de qualidades de processo e produto; avaliação de sucessos e fracassos; feedback da informação para controle do projeto; aprendizado com base na experiência.; embalagem e reutilização de experiência relevante.

É importante realizar uma avaliação empírica de alteração do processo atual, e para isso deve-se escolher uma estratégia de pesquisa com base no risco e no tamanho da alteração. Pode-se também utilizar o paradigma da melhoria da qualidade, onde se caracteriza o ambiente; define-se metas; escolhe um processo apropriado; executa as alterações; analise os dados; e

consolida as experiencias. Para a construção da experiencia, é proposto que haja metas, e que cada meta tenha 1 ou mais questões a serem respondidas; e que acada questão tenha 1 ou mais métrica associadas.

Experimentos científicos possuem grande relevância para a industria, e a industria possui processos com complexidade de interesse da academia. Desse modo, um processo de de avaliação e proposta de alteração de processo é necessário. O capítulo divido essa troca de experiência em alguns passos que podem ser responsabilidade da industria ou da academia. A definição do problema pela industria; a formulação do problema pela industria e pal academia; estudo do estado de arte pela academia; a proposta de soluções candidatas pela industria e pela academia; a validação das soluções candidatas pela academia; uma análise estática pela industria; uma validação dinâmica pela industria;e finalmente a liberação da proposta pela industria.

Devido a natureza da informação que pode ser coletada durante uma experimentação, aspectos éticos devem ser considerados. Desse modo os sujeitos da experimentação devem ser informados e estar de acordo com a sua participação; o estudo deve ser motivado cientificamente; os pesquisadores devem fazer o máximo possível para manter a confidencialidade até mesmo face a um conflito durante a publicação. Adicionalmente, é importante que haja um feedback os sujeitos da pesquisa para que haja uma cooperação de longo alcance.

2.3 CAPÍTULO 3 - MÉTRICAS

A medição do software é importante para ser possível controlar projetos, produtos e processos. Além disso a medição desempenha um papel fundamental nos estudos empíricos. É através de medições que podemos fazer comparações, estimativas e afirmações.

Medição é o processo pelo qual números ou símbolos são atribuídos a atributos de uma entidade do mundo real de modo a defini-la em termos de regras claras. As medidas são esses números ou símbolos que são atribuídos a entidades no processo de medição. As métricas são utilizadas para denotar entidades que são medidas. O principal objetivo em mapear uma medição a um valor é para caracterizar esse valor, como por exemplo dizer que um objeto A é menor que o objeto B, ou que A e B são do mesmo tipo.

Em um estudo empírico é importante preservar a validade das medidas observadas, ou seja, se uma propriedade for observada, a medição deve preservar essa propriedade dentro de certos limites de erros.

A medição de medidas de atributos podem ser realizadas em diferentes escalas, e pode ser de interesse transformar entre diversas escalas. Essa transformação pode ser admissível desde que o relacionamento entre os objetos medidos seja preservado. De modo semelhante, se afirmações são verdadeiras em diferentes escalas utilizadas, pode-se considerar que as afirmações fazem sentido, caso contrário não a fazem.

Existem 4 principais tipos de escalas: a nominal, a ordinal, a de intervalo e a de razão. A escala nominal é a que traz consigo a menor quantidade de informação do atributo medido. Ela mapeia um atributo de uma entidade a um nome ou símbolo como um tipo de classificação. A escala ordinal adiciona mais informação ao atributo do que a nominal. Ela proporciona uma ordenação entre os atributos de objetos medidos, desse modo pode-se observar critérios como "maior do que", "menor que", "mais complexo que"entre outros. A escala de intervalo é utilizada quando há sentido em estabelecer uma diferença entre duas medidas. Assim como a escala ordinal, a de intervalo também possui uma ordenação entre os valores, mas adiciona o conceito de distancia entre esses valores. Por último a escala de razão onde existe a noção de ordenação, há uma distancia relativa entre os valores e também a proporcionalidade dos valores faz sentido. É mas comum encontrar no campo de engenharia de software as escalas nominais ou ordinais, e as escalas de intervalo e de razão são mais comuns em medidas de grandezas físicas.

As medidas também podem ser objetivas ou subjetivas. As medidas objetivas dependem apenas do objeto medido. Se as medidas forem levantadas varias vezes, o mesmo resultado será obtido dentro de uma faixa de erro. As medidas subjetivas dependem do julgamento da pessoa que contribui com a medição. Desse modo, esse tipo de medida depende tanto do objeto quanto do ponto de vista que o objeto é medido. As medidas subjetivas são normalmente associadas as escalas nominais e ordinais.

Medidas também podem ser diretas ou indiretas. As medidas diretas são diretamente mensuráveis e não necessitam de nenhuma medida de outro atributo. Exemplo como quantidade de linhas de código, ou número de defeitos são característicos de medidas diretas. Medidas indiretas por outro lado envolvem a medição de outros atributos e assim são derivadas desses atributos. Exemplo como densidade de defeitos (numero de defeitos por linha de código) ou produtividade de programadores (linhas de código por esforço de programação) são exemplos de medidas indiretas.

As medidas em engenharia de software são divididas em 3 classes: as de processo, que descreve atividades necessárias para produzir software; as de produto, dos artefatos produzidos

pelo resultado das atividades dos processos; e as de recursos, referentes aos objetos necessários para a realização da atividade do processo como pessoal, hardware ou software. Existe uma distinção entre atributos internos e externos. Os internos podem ser medidos puramente com os termos do objeto. Os externos necessitam de um relacionamento com outros objetos.

Na engenharia de software normalmente queremos fazer afirmações de atributos externos, porém eles são normalmente indiretos que necessitam serem derivados de atributos internal, os quais são normalmente diretos.

Na prática, as métricas são definidas pelo pesquisados e são coletadas durante a fase de execução. Para prover um alto grau de qualidade, é necessários um entendimento das métricas, do tipo de escala e qual distribuição a métrica pertence.

2.4 CAPÍTULO 4 - REVISÃO SISTÊMICA DA LITERATURA

A revisão sistemática da literatura propõe "identificar, analisar e interpretar todas as evidências disponíveis relacionadas a um problema de pesquisa específico". Esse processo deve seguir um método científico. Esse método foi proposto em 3 etapas: planejamento, condução e apresentação da revisão.

O planejamento da revisão é dividido em algumas ações. Primeiramente deve-se identificar a necessidade de uma revisão. Essa necessidade pode se originar de pesquisadores buscando o estado da arte de uma área ou profissionais que desejam utilizar evidências empíricas em suas decisões estratégicas ou atividades de melhoria. Preferencialmente, deve-se utilizar uma revisão sistemática da literatura já existente se for suficiente. Dessa forma, uma revisão sistemática da literatura pode ser compreendida como um método de investigação de produção de revisão da literatura.

Também é necessário que o problema de pesquisa seja especificado. Assim, é possível identificar os estudos primários, a necessidade de extração dos dados desses estudos e sua analise. Alguns aspectos importante para elaboração de um problema de pesquisa são a identificação da população de interesse da revisão; quais ferramentas, tecnologias ou procedimentos serão estudadas no estudo empírico; a comparação a ser realizada antes e após a aplicação das ferramentas, tecnologias ou procedimentos; os aspectos práticos do resultado da experimentação; o contexto em que o estudo será aplicado; os projetos experimentais utilizados no problema de pesquisa.

Um protocolo de revisão também é necessário. Esse protocolo deve definir procedimentos para a revisão sistemática da literatura. Ele também desempenha um papel de diário durante a condução da revisão. Assim, é um documento vivo importante tanto na condução da revisão quanto na sua validade. Esse protocolo deve, preferencialmente, ser revidado por pares para garantir sua consistência e validade. O protocolo de revisão proposto é: antecedentes e justificativa, problema de pesquisa, estratégia de busca para estudos primários, critérios de seleção de estudos, procedimentos de seleção de estudos, listas de verificação e procedimentos de avaliação da qualidade, estratégia de extração de dados, síntese dos dados extraídos, estratégia de divulgação e cronograma do projeto.

A segunda etapa do método de revisão sistemática da literatura é a condução. A condução propõe aplicar o protocolo de revisão na prática. Como atividades tem-se a identificação de pesquisa. A identificação de pesquisa envolve a especificação de textos de procura e sua aplicação em base de dados. Também é necessário a pesquisa manual em revistas ou periódicos, atas de conferências, web sites e até mesmo perguntas diretamente a pesquisadores. Também pode ser utilizado a busca por pesquisa em outros estudos (conhecido como "snowballing"). A estratégia de pesquisa é um compromisso entre encontrar todos os estudos primários relevantes e não obter um número elevado de falsos positivos, que posteriormente precisará ser removido. É compreensível, porém, que os artigos encontrados sejam uma amostra de todos os arquivos de um tópico específico. Os estudos publicados tendem a ter um viés onde os resultados positivos possuem uma maior probabilidade de serem publicados do que resultados negativos. Assim, teses, publicações rejeitados e trabalhos ainda não concluídos devem ser pesquisados. Finalmente, as ações tomadas devem ser armazenas, preferencialmente em um sistema de gestão de referências.

A seleção de estudos primários da-se através dos critérios de inclusão ou exclusão. Critérios esses que devem ser definidos previamente, mas podem sofrer alterações ao decorrer da condução da revisão sistemática da literatura. A identificação desses artigos pode ser feita lendo o título e o abstract, em alguns caso o entendimento da metodologia e da conclusão podem ser necessários. É recomendável que 2 ou mais pesquisadores avaliem um artigo já que a escolha é apenas uma questão de julgamento, e que estatística dessa concordância sejam apresentadas no estudo.

A avaliação da qualidade dos estudos primários também é importante, principalmente quando estudos apresentarem resultados contraditórios. Não existe uma definição do que é um

estudo de qualidade. Estudos mostram que ao menos 3 pesquisadores são necessários para fazer uma avaliação válida. A utilização de lista de verificação é um meio prático de avaliar estudos.

Uma vez a decidido os estudos primários, o processo de extração de dados se inicia. A extração é baseada no problema de pesquisa. É aconselhável que o processo processo de extração seja executado por 2 pesquisadores para que seja possível verificar a qualidade da extração. A extração dos dados devem ir para um formulário onde o nome do revisor deve ser preenchido.

O método mais avançado de síntese de dados é a meta-análise. A meta-análise assume que os estudos sintetizados são homogêneos ou que a causa é bem conhecida. Ela compara o tamanho dos efeitos e os valores p para avaliar o resultado da síntese. Os estudos a serem incluídos na meta-análise devem ser do mesmo tipo; terem a mesma hipótese de teste; terem as mesmas medidas de tratamento e construções de efeito; apresentarem os mesmos fatores explicativos. O processo de meta-análise é dividido em 3 passos: decidir quais estudos incluir; extração do tamanho do efeito dos estudos ou estimá-los; combinar o tamanho do efeito para estimar e testes os efeitos combinados. Métodos menor formais de síntese de dados incluem síntese narrativa ou descritiva, onde os dados são tabulados proporcionando uma melhor entendimento ao pesquisador.

A apresentação da revisão sistemática da literatura pode ser reportada para diferentes audiências e em diferentes meios. Para audiências acadêmicas, é importando detalhar os procedimentos utilizados.

Quando o problema de pesquisa é amplo, ou a linha de pesquisa não foi tão explorada, um mapeamento de estudo ao invés de uma revisão sistemática da literatura pode ser utilizada. Um mapeamento de estudos visa fornecer uma introdução geral sobre o estado da arte de um tópico. O processo de um estudo de mapeamento é semelhante ao da revisão sistemática da literatura, porém os critérios de inclusão ou exclusão e de qualidade são diferentes. A coleta de dados e a sístase de dados tendem a ser mais qualitativas do que em uma revisão sistemática da literatura.

Foi apresentado em Conforto, Amaral e Silva (2011) um roteiro para a condução de revisão bibliográfica sistemática. O processo foi divido em 3 fases: entrada, processamento e saída. A fase da entrada foi sub-dividida em: definição do problema; construção de objetivos claros e factíveis; definição de estudos primários; definição de palavras chave para busca em banco de dados; definição dos critérios de inclusão dos estudos que devem ser alinhados com o objetivo

da pesquisa; definição de critérios de qualificação com finalidade de avaliar a importância de um estudo; definição de métodos e ferramentas; e definição de um cronograma.

Segundo Conforto, Amaral e Silva (2011), a segunda fase é a de processamento, onde são propostas uma busca, análise e documentação e um processo iterativo de 7 etapas sendo elas: busca por periódicos; filtragem através da leitura do título, resumo e palavras-chave; uma nova filtragem pela introdução e conclusão; uma filtragem através da leitura completa; realizar uma busca cruzada por meio de citações de autores; busca por base de dados; e finalmente catalogar os artigos selecionados em um repositório de arquivos. Também são apresentados formulários de suporte a serem utilizados nas etapas de filtragem. A terceira fase, de saída, consiste em 4 etapas: identificação de periódicos relevantes; cadastro do arquivo; síntese do resultado; construção de modelos teóricos.

Felizardo et al. (2019) apresentou uma abordagem que faz uso de técnicas de mineração Visual de Texto e uma ferramenta para suportar a seleção de estudos primários. A ferramenta apresentada foi a *Project Explorer* (PEx) que gera um modelo gráfico de uma coleção de documentos. Sua proposta foi de agrupar estudos altamente relacionados automaticamente e refinar com apoio de recursos baseados em conteúdo e recursos baseados em rede de citações com a finalidade de convalidar ou alertar sobre decisões de incluir ou excluir um estudo primário em uma revisão sistemática. Os recursos baseados em conteúdo podem ser obtidos através do histórico de exclusões, onde alguns agrupamentos podem ser excluídos e outros devem ser revisados; e classificação da qualidade dos estudos onde a qualidade de alguns estudos são avaliadas e assim agrupamentos podem ser excluídos ou necessitarem de avaliação. Os recursos baseados em rede de citação podem ser representados como grafos, sendo possível visualizar estudos que não estão conectados a outros e estudos altamente relacionados.

Marshall e Brereton (2013) mapeou o estado atual das ferramentas que suportam a revisão sistemática da literatura. Para tal, dividiu a metodologia em 4 etapas: problema de pesquisa; processo de pesquisa; critério de inclusão e exclusão; e extração dos dados. O problema de pesquisa foi formulado em 3 perguntas. No processo de pesquisa, foi detalhado as ferramentas utilizadas bem como as palavra chaves pesquisadas, e também foi utilizado a estratégia "snowballing". Os critérios de inclusão e exclusão foram detalhados sendo 3 os critérios de inclusão e 2 os de exclusão. No processo de extração de dados, 2 autores realizaram o processo, sendo que o processo de resolução de discordâncias foi detalhado.

2.5 CAPÍTULO 5 - ESTUDO DE CASO

O estudo de caso é aplicável em muitos tipos de pesquisa em engenharia de software quando os objetos de estudo são fenômenos contemporâneos, que são difíceis de estudar isolados. O estudo de caso é uma estratégia de pesquisa e muitas vezes utilizam de diversas fonte de evidências. São utilizados quando o limite entre o fenômeno e o contexto não são claros.

Um pouco diferente de estudo de caso são as pesquisas-ações. Essas influenciam ou alteram de alguma forma o objeto de pesquisa. Portanto, o estudo de caso é puramente observacional, enquanto as pesquisas-ações alteram de alguma forma o processo. Os estudos etnográficos podem ser considerados especializações de estudos de caso onde o foco é sobre práticas culturais ou estudos de longa duração que envolvem grandes quantidades de dados sobre participantes ou observadores.

Um estudo de caso contem elementos de outras métodos de pesquisa, como por exemplo: entrevistas, pesquisa bibliográfica, análise de arquivos entre outros. Um estudo de caso é caracterizado por: (1) lidar com situações distintas onde existam mais variáveis do que pontos de dados e apenas 1 resultado; (2) se baseiam em múltiplas fontes de dados; (3) se beneficiam de desenvolvimentos passados de proposições teóricas para guiar a coleta de dados. Desse modo um estudo de caso não irá prover conclusões com significância estatística. As características do estudo de caso são: (1) são flexíveis, podendo lidar com características dinâmicas e complexas; (2) suas conclusões são baseadas em uma cadeia de evidências; e (3) se baseiam em teorias previas ou constroem suas próprias teorias.

A pesquisa de um estudo de caso é do tipo flexível, mas isso não significa que o planejamento não é importante. O planejamento é necessário para decidir o método de coleta de dados, quais departamentos ou organizações visitar, quais documentos serão lidos, quais pessoas investigar, com qual frequência conduzir entrevistas, entre outros. O planejamento deve conter: o objetivo, o caso, a teoria, o problema de pesquisa, o método e a estratégia de seleção. O objetivo pode ser entendido como um área de foco com capacidade de expansão ao longo da pesquisa. O problema de pesquisa também pode ser especializado do longo dos ciclos de iterações.

O protocolo de estudo de caso é um documento que pode ser alterado ao longo da pesquisa. Ele tem o objetivo de guiar o estudo e é realizado durante o planejamento. Também possui os objetivos de guia durante o processo de condução da coleta de dados; torna a pes-

quisa concreta durante a fase de planejamento e permite que outros pesquisadores revisem o planejamento.

A coleta de dados podem vir de 3 diferentes níveis. O primeiro nível é o modo direto, onde os pesquisadores estão em contato direito com os sujeitos e coletam os dados em tempo real como por exemplo, entrevistas. No segundo nível estão os métodos indiretos, onde o pesquisador coleta diretamente dados butos sem interagir com os sujeitos como por exemplo: histórico de uso e observações de gravações em vídeo. O terceiro nível contém analises de trabalhos independentes já preparados, como exemplo temos a análise de especificações de requisitos já preparadas ou relatórios de falhar. Tanto no primeiro quanto no segundo nível o pesquisador pode controlar e direcionar os dados coletados.

Existem três tipos de entrevistas, as completamente estruturadas, as semi-estruturadas e as não estruturadas. Na entrevista estruturada todas as questões são planejadas com antecedência e seguem uma sequência pré-determinada, as perguntas são fechadas, o objetivo é descritivo e explicativo, e o pesquisador busca encontrar relações em construções. As semi-estruturadas possuem as questões previamente planejadas mas não necessariamente seguem uma ordem pré-definida, possuem uma mistura de perguntas abertas e fechadas, possuem objetivo descritivo e explicativo, e o foco é entender a experiência do indivíduo de forma quantitativa e qualitativa. As não estruturadas não seguem um questionário previamente estabelecido e portante dependem do desenvolvimento da entrevista e do interesse do pesquisador, possuem objetivo exploratório e o foco é em entender a experiência do indivíduo de forma qualitativa.

A coleta de dados de forma observativa pode ser dividida em pouco ou muito envolvimento do pesquisados e com pouco ou muito conhecimento do sujeito sobre a pesquisa. Observações proveem um entendimento profundo sobre um dado fenômeno. Além disso, esse tipo de coleta de dados tem relevante aplicação quanto há suspeita de exista um desvio. A análise de arquivos também é um método de coleta de dados, porém é de certa forma difícil para os pesquisadores avaliarem a qualidade do dado , já que o dado não foi obtido com foco para a pesquisa em condução.

Existem dois tipos de análises de dados, as análise quantitativa e a qualitativa. A análise quantitativa utiliza-se de estatística descritiva, modelos de correlação, predição e teste de hipótese. A análise qualitativa busca derivar conclusões dos dados mantendo uma clara cadeia de evidências. Desse modo, evidência são claras e proporcionam ao leitor do estudo um mecanismo de derivação das mesmas conclusões se seguidas as decisões do pesquisados.

O processo de análise qualitativo é dividido em duas etapas, a técnica de geração de hipótese e a técnica de confirmação da hipótese. A geração da hipótese busca encontrar hipóteses a partir dos dados. O pesquisados deve portanto não possuir viés e estar aberto para qualquer que seja a hipótese encontrada nos dados. A técnica de confirmação da hipótese consiste na confirmação de que a hipótese é realmente verdadeira, para isso triangulação e replicações são abordagens possíveis.

A análise qualitativa dos dados é conduzida em um sequência de passos. Primeiramente os dados são codificados, isto é, códigos são atribuídos a partes do texto para representar temas, áreas, ou construções. Um código pode ser atribuído a diferentes partes do texto, e uma parte do texto pode ser atribuído a mais de um código. Códigos pode ter uma hierarquia de códigos ou sub-códigos. Os códigos podem ser combinados a reflexões e pensamentos do pesquisador. Com essa etapa concluída, o pesquisador pode passar pelo texto e extrair as primeiras hipóteses. Isso pode ser frases similares em diferentes partes do texto, padrões de dados, entre outros. As hipóteses identificadas poder ser utilizadas em futuras coletas de dados, resultado em um processo iterativo onde a coleta e análise de dados são realizadas em paralelo. Durante as iterações, generalizações podem ser formuladas gerando o resultado final que são as bases de conhecimento.

É importante considerar a validade do estudo de caso desdo começo. Um exemplo de como melhorar a validade são as triangulações, desenvolvendo e mantendo um detalhamento do protocolo do estudo de caso.

O relatório do estudo de caso também é um artefato importante. Ele deve ser direcionado para o publico alvo do estudo de caso, e caso exista mais de um publico alvo, diferentes relatórios devem ser realizados. O relatório deve contar sobre o que o estudou era; comunicar claramente o que foi estudado; prover um histórico de evolução e inquérito; prover os dados básicos para o leitor poder concluir que os resultados são razoáveis; e articular as conclusões. A estrutura do reporta depende do tipo de público alvo, porém uma estrutura analítica linear de problemática, estudos relacionados, metodologia, análise e conclusão é a estrutura mais aceita.

2.6 CAPÍTULO 6 - PROCESSO EXPERIMENTAL

O princípio básico da experimentação se inicia na existência de uma ideia de relacionamento de cause e efeito, essa ideia pode ser uma teoria ou uma formulação de hipótese. Uma hipótese implica na existência de uma declaração formal, tal declaração pode ser de um

relacionamento, por exemplo. Um experimento pode ser utilizado para verificar uma teoria ou hipótese. Existem diversos tratamentos que um experimento possui controle. Um tratamento é um valor que uma variável do estudo pode assumir. Desse modo, o experimento é realizado e um resultado pode ser observado, promovendo assim um teste do relacionamento entre o tratamento e o resultado.

É necessário que alguns termos sejam explicados. Os experimentos possuem variáveis, sendo elas dependentes ou independentes. As varáveis que os experimentos analisam como resultado da alteração de variáveis independentes são as dependentes. Todas as variáveis manipuladas em um processo de experimentação são chamadas variáveis independentes. Um experimento estuda a alteração de uma ou mais variáveis independentes, sendo essas variáveis chamadas de fatores. Um tratamento é um valor particular de um fator. A escolha do fator e o nível que outros variáveis independentes devem possuir é detalhado durante o desenho do experimento. Os tratamentos são aplicados em uma combinação de objetos e sujeitos. O sujeito é a pessoa que participa diretamente do experimento, e o objeto é o material que o sujeito utiliza como fator da experimentação. As características do objeto e do sujeito poder ser variáveis independentes do experimento. Um experimento consiste em um conjunto de testes em que cada teste é uma combinação de tratamento, sujeito e objeto. O numero de testes afeta o erro experimental e ajuda a estimar o nível de confiança.

Experimentos onde o pessoas possuem papel central possuem diversas limitações em relação ao controle do experimento. Pessoas possuem diferentes habilidades, o que pode ser uma variável independente. Pessoas aprendem com o tempo, o que significa que a ordem de aplicação do tratamento pode importar e também que alguns objetos não podem ser utilizados em duas ocasiões. Pessoas também são impactadas por diversos fatores externos. Os experimentos onde tecnologias possuem papel central são mais fáceis de controlar já que tecnologias tendem a ser mais determinísticas. Objetos podem ser variáveis independentes mais difíceis de controlar, já que algumas técnicas são melhores aplicadas a certos tipos de programas e não a outras.

Faz-se então necessário um processo para a experimentação. Um processo deve ser utilizado tanto para experimentos randômicos quanto para quasi-experimentos. Os quasi-experimentos são geralmente usados em engenharia de software quando amostras randômicas, como por exemplo sujeitos, não são possíveis.

Uma vez concretizado o entendimento da necessidade de um experimento, um planejamento deve ser realizado. O processo de experimentação pode ser dividido em levantamento de

escopo, onde o escopo é definido em termos de problemática, objetivo e metas. Planejamento, onde o desenho da experimentação é definido, considerando a instrumentação e as ameaças são avaliadas. A operação do experimento segue o desenho, em que medidas são coletadas, analisadas e avaliadas na etapa de análise e interpretação. Por fim, os resultados são agrupados e apresentados.

Não é esperado que o processo de experimentação seja realizado em um verdadeiro modelo cascata. Não é esperado que uma etapa seja completamente finalizada para que outra seja iniciada. Ao longo do processo pode ser identificado a necessidade de voltar algum passo e refinar a atividade antes de continuar com o experimento. A exceção ocorre quando a operação do experimento já foi iniciada, não podendo mais voltar para o levantamento de escopo ou planejamento. Se assim fosse, os sujeitos do experimentos poderiam ser influenciados, necessitando a substituição dos sujeitos.

2.7 CAPÍTULO 7 - ESCOPO

Para definir o escopo de um experimento é necessário definir os objetivos. A utilização de um modelo de objetivos ajuda a garantir que aspectos importantes sejam definidos antes do planejamento e da execução. O modelo de objetivo deve ter: o objetivo do estudo, o propósito, o foco de qualidade, a perspectiva e o contexto. O objetivo de um estudo pode ser produtos, processos, recursos, modelos, métricas ou teorias. O propósito de um experimento indica sua intenção. O foco da qualidade do experimento é o efeito principal do estudo. A perspectiva determina o ponto de vista que os resultados são interpretados. O contexto é o ambiente que o experimento é executado. O contexto pode ser definido pelo numero de sujeitos e objetos envolvidos no estudo. Existes vários tipos de estudos, os que possuem 1 sujeito e 1 objeto, 1 sujeito e N objetos, N sujeitos e 1 objeto, e N sujeitos e M objetos. Eles podem ser experimentos ou quasi-experimentos.

2.8 CAPÍTULO 8 - PLANEJAMENTO

Após a definição do escopo do experimento, a etapa de planejamento ocorre. O planejamento deve conter as etapas de seleção de contexto, formulação da hipótese, seleção das variáveis, seleção dos sujeitos, escolha do tipo de desenho, instrumentação e avaliação da validade.

A seleção do contexto é importante para atingir resultados mais genéricos, para isso, é necessários realizar experimentos em larga escala, em ambiente profissional, em projetos de software real. Porém, a condução de tais experimentos envolvem riscos. Uma alternativa é executar os experimentos em conjunto com experimentos paralelos menores. Essa abordagem pode gerar custos extras. Outra alternativa é realizar experimentos com estudantes, porém são menos genéricos. A definição de escopo busca balancear entre fazer um experimento válido em um determinado contexto ou generalizar em um domínio de software mais amplo. Assim, 4 dimensões podem ser definidas: on-line ou off-line; com estudantes ou profissionais; com problemas reais ou simulações; específico ou geral.

O teste de hipótese é essencial em uma análise estatística. Um experimento deve ter 2 hipóteses. A primeira é a qual define que o experimento não tem relação com o efeito, e é a qual o experimento tenta rejeitas com alto grau de significância. A segunda corresponde ao contrário da primeira e é a qual o experimento tenta ser favorável. Os riscos associados são a rejeição de uma hipótese que é verdadeira (Tipo I), ou a não rejeição de uma hipótese falsa (Tipo II). O poder do teste é definido como a habilidade do teste estatístico em revelar um padrão verdadeiro.

A seleção das variáveis dependentes e independentes devem também ocorrer na etapa de planejamento. É necessário um conhecimento amplo do domínio para selecionar as variáveis corretas. Usualmente existe apenas 1 variável dependente derivada diretamente da hipótese. Normalmente, a variável não é mensurável diretamente e uma medição indireta deve ser realizada.

A seleção dos sujeitos está diretamente relacionada com a generalização dos resultados. Uma seleção representativa proporciona um generalização. A seleção dos sujeitos pode ser probabilística e não probabilística em que a diferença é sabida previamente a probabilística de selecionar uma certa amostra.

Uma análise estatística é necessária para que seja possível ter conclusões significativas. Um experimento consiste em uma série de testes de tratamentos. Um desenho de experimento descreve como os testes são organizados e executados.

Um bom desenho possibilita a replicação do experimento. Os princípios gerais de desenho são randomização, agrupamento, e balancemanto. A randomização é importante para que não os vieses sejam reduzidos, ou ao menos distribuídos uniformemente em toda a população. O agrupamento visa eliminar fatores que podem induzir um determinado efeito, ou seja, grupos são formados e espera-se resultados semelhantes dentro de um mesmo grupo. O balancemendo

implica em possuir a mesma quantidade de tratamentos para um número igual de sujeitos, simplificando e fortalecendo a análise estatística.

Pode-se classificar os tipos de experimentos padrões em 4 categorias. Os experimentos de 1 fator com 2 tratamentos visa investigar dois tratamentos, sendo o usual comparar a média entre duas variáveis dependentes. Os experimentos com 1 fator e mais do que 2 tratamentos visa comparar tratamentos entre eles. Os experimentos com 2 fatores são mais complicados, já que devem possuir uma hipótese para um fator, uma hipótese para outro fator e uma terceira hipótese para a interação entre os fatores. Os experimentos onde devem ser consideramos mais do que 2 fatores são mais complicados. Mesmo que possível, é recomendável que seja utilizado um experimento simples.

O objetivo geral da instrumentação é prover meios de realizar o experimentos e monitoralo, sem afetar o controle do experimento. As medidas são realizadas através da captura de dados. Em experimentos com humanos, é necessário planejar os formulários ou as entrevistas. As diretrizes são utilizadas para guiar os participantes do experimento. Os objetos sob investigação devem ser conhecidos.

A validade do experimento é muito importante. Para isso, é necessário que o experimento seja válido na população onde onde foi realizado e possuir viabilidade de generalização para populações maiores. Ao realizar a conclusão do experimento, é necessário validar o experimento utilizando estatística; ter confiança de que o resultado não foi influenciado por algum fator interno ao experimento; validar o relacionamento entre a teoria e a observação; e possuir uma validação externa, onde seja possível generalizar.

2.9 CAPÍTULO 9 - OPERAÇÃO

A fase seguinte ao desenho do experimento é a operação. Ela consiste na preparara, em que sujeitos são escolhidos e formulário são preparados; execução, em que os sujeitos realizam suas tarefas de acordo com diferentes tratamentos e dados são coletados; e validação do dado, onde os dados coletados são validados.

Uma preparação bem desenvolvida facilita a execução do experimento. Existem 2 aspectos importantes na preparação, a seleção e notificação dos participantes e a preparação do material somo formulários e ferramentas.

É importante que a seleção dos sujeitos do experimento ocorra de forma que haja uma representatividade do que se busca experimentar. Se existe a necessidade de realizar um experi-

mento relacionado ao código C, por exemplo, é importante que os sujeitos tenham experiência nessa área, evitando assim ameças externas a validade do experimento. É essencial também, que os sujeitos entendam a razão do experimento, como os dados serão utilizados e publicados, e que haja um consentimento deles. Desse modo, evita-se que os dados coletados se tornem inválidos. Também é importante que dados sensíveis, isto é, que podem vir a identificar os participantes, sejam mantidos em segredo. Benefícios atrativos também podem ser oferecidos a sujeitos, mas tais benefícios não podem ser grandes o suficiente para motivar os participantes. Se assim for, os participantes podem participar do experimento apenas por essa razão causando um ameaça a ele. É importante também que os sujeitos conheçam os objetivos do experimento, porém, se não for possível, é importante que princípios de ética sejam considerados.

Todos os instrumentos do experimentos devem estar prontos antes que a execução seja iniciada. Os instrumentos são determinados pelo desenho. Se possível, os formulários que os sujeitos coletarão seus próprios dados devem ser anônimos. Muitas vezes, é necessário preparar os instrumentos para cada sujeito, com isso é possível que haja randomização e que cada sujeito possa realizar os testes desenhados.

Existem diversas formas de se executar um experimentos. Alguns experimentos mais simples podem ser executados em uma única execução, em que o experimentador e o sujeito realizam uma reunião e os dados são coletados ali mesmo. Outros experimentos, como quando longos projetos são estudados, não permite ao experimentador participar de todas as etapas.

A coleta de dados pode ser através do preenchimento de formulários pelos participantes, de forma manual com ajuda de ferramentas, em entrevistas, ou de forma automática por ferramentas. As entrevistas possibilitam o esclarecimento de perguntas ao contra ponto de exigir um maior esforço do experimentador. Os formulários exigem menos esforço, porém inconsistência e incertezas são mais difíceis de serem descobertas.

O ambiente que o experimento é conduzido não deve ser alterado. Sendo assim possível identificador os diferentes tratamentos nele. Algumas vezes é necessário que o ambiente seja modificado para que haja uma maior motivação para participação.

A validação dos dados coletados propõe verificar se os dados são razoáveis. Pode ocorrer de os participantes não entenderem corretamente as questões. Nesse caos pode-se realizar um seminário, apresentando os resultados, e explicando-os, para que os participantes possam refletir sobre os resultados que discordam.

2.10 CAPÍTULO 10 - ANÁLISE E INTERPRETAÇÃO

Os dados coletados na etapa de operação são utilizados durante a análise e interpretação. Uma interpretação quantitativa pode seguir 3 passos: caracterização através de estatística descritiva, redução do conjunto de dados através da exclusão de dados anormais, e análise via teste de hipótese.

A estatística descritiva lida com a apresentação e processamento numérico do conjunto de dados. O objetivo desse processo é indicar como o conjunto de dados está distribuído e assim identificar dados anormais, além de proporcionar mais informações sobre a natureza do dado. Abaixo serão apresentadas algumas medidas de estatística descritiva.

A medida central de tendência, como média, mediada e moda, indica a centroide do conjunto de dados. Ela pode ser interpretada como uma estimativa da expectativa de uma variável estocástica de uma amosta do conjunto de dados. A média possui significado para as escalas métricas intervalar ou de razão. A mediana representa o valor central de um conjunto de dados, possuindo significado nas escalas ordinal e métrica. A moda representa a amostra de ocorrência mais comum e é aplicável para todos os tipos de escalas.

Para medir a dispersão do conjunto de dados é necessário estimar o nível de variância da tendência central. Essa variância é a média do quadrado da distância do ponto central. A variância é significativa para escalas métricas. O desvio padrão é o resultado da aplicação da raiz quadrada na variância. Ele possui a mesma dimensão do que os valores e possui significância nas mesmas escalas da variância.

O intervalo dos dados, calculado através da subtração do maior valor pelo menor valor, podendo ser representado pelo par de valores mínimo e máximo, é um indicativo de dispersão que também é aplicável na escala ordinal. Também é possível representar a dispersão na forma de tabela de frequência, em que cada linha da tabela indica possui a frequência relativa do total de amostras para um determinado valor.

Nos casos em que as amostrar são pares provenientes de duas variáveis estocástica, é interessante analisar a dependência entra essas variáveis. Caso as variáveis estejam relacionadas de forma linear, pode-se aplicar a regressão linear e ajustar a reta minimizando a soma das distâncias quadráticas. Caso a dependência não seja linear, pode ser possível transformar os dados de tal forma que seja possível aplicar a regressão linear.

A covariância pode ser utilizada para quantificar em qual proporção dois conjuntos de dados variam. A covariância possui significância em escalas métricas. Normalizada, possui

intervalo de -1 a +1, sendo 0 a ausência de correlação linear, sendo essa chamada de correlação de Pearson. No caso de escala ordinal ou em uma distribuição não normal, o coeficiente de Spearman pode ser utilizado. Caso existam mais do que 2 variáveis, pode-se utilizar análise de componente principal, análise discriminatória entre outros métodos.

Técnicas de visualizações gráficas poder ser combinadas com os outros métodos e proporcionam boa visualização. Gráficos de dispersão podem ser utilizados onde existam pares de amostras, podendo apresentar dependências entre as variáveis. Histogramas podem apresentar uma visão geral da densidade da distribuição. O gráfico de pizza representa frequência relativa dos dados divididos em grupos de classes.

A qualidade da análise estatística depende dos dados de coletados. Podem existir erros sistemáticos nos dados coletados, mas também podem existir exceções. As exceções são dados muito menores ou muito maiores do que o esperado. Elas podem ser identificados através de análise gráfica ou utilizando métodos estatísticos. As exceções, quando ocorrerem de forma singular, devem ser excluídas.

Em sequência, o teste de hipótese dese ser realizado. Deve-se então verificar se é possível rejeitar a hipótese nula H_0. Caso a hipótese nula seja rejeitada, nada pode ser concluído, porém caso seja rejeitada pode-se dizer que a hipótese é falsa dada uma significância α. α é $P(erro - do - tipo - I)$ e β é $P(erro - do - tipo - II)$, em que o erro do tipo I é $P(rejeitar - H_0|H0 - \text{é} - verdadeiro)$ e o tipo II é $P(\text{não} - rejeitar - H_0|H_0 - \text{é} - falso)$.

Os testes podem ser classificados como paramétricos e não paramétricos. Os testes paramétricos envolvem modelos de uma distribuição específica. Esse tipo de teste necessita que parâmetros sejam medidos em uma escala de intervalo. Os testes não paramétricos fazem suposição mais genéricas e possuem maior generalização do que os testes paramétricos. O poder dos testes paramétricos é maior do que os não paramétricos, traduzindo-se em uma menor quantidade de dados necessários, porém uma análise da aplicabilidade do teste de ser realista.

Existem diversos tipos de testes. O *t-test* é um dos testes parâmetricos mais utilizados e é utilizado para comparar 2 médias, isto é, um fator com dois tratamentos. O *Mann-Whitney* é um teste não paramêtrico, e o *F-test* é um teste paramêtrico utilizado na comparação de duas amostras. Ainda existem os testes *Paired t-test*, *Wilcoxon*, *Sign test*, *ANOVA*, *Kruskal-Wallis*, *Chi-2*, entre outros. Cada tipo de teste pode ser utilizado por uma escolha de desenho diferente. Cada modelo estatístico possui suposições diferentes quanto a normalidade, independência e escala dos dados coletados. Se o conjunto de dado não obedece a suposição, o teste de hipótese é então considerado inválido.

Uma vez concluído o teste de hipótese, é necessário concluir o experimento. Caso a hipótese nula tenha sido rejeitada, pode-se analisar a influência das variáveis independentes sobre as variáveis dependentes. Caso contrário, a única conclusão é que não há significância significativa entre os tratamentos. É importante notar também que o fato da hipótese nula não poder ser rejeitada com uma certa significância não necessariamente significada que a hipótese é verdadeira. Adicionalmente, se for encontrada uma correlação alta entre duas variáveis, não podemos concluir relação causal entre elas, pode haver um terceiro fator que causa efeito na medida.

2.11 CAPÍTULO 11 - APRESENTAÇÃO E EMPACOTAMENTO

Podem existir diversas audiência interessadas no resultado da experimentação, porém a foi enfatizado a apresentação em conferências e revistas acadêmicas. A primeira seção é o resumo abstrato. Os elementos nele presentes são: contexto, objetivos, métodos, resultados e conclusões. Usualmente, o resumo é limitado em número de palavras, como por exemplo 300 palavras.

Em seguida, deve ser apresentada a motivação ou introdução, em que o escopo e o objetivo são apresentados, fornecendo a razão da necessidade do experimento. Os trabalhos relacionados seguem essa seção, provendo uma justificativa de como o experimento se relaciona com trabalhos conduzidos previamente.

O desenho do experimento deve ser apresentado na seção seguinte. Nele deve conter a hipótese que é derivada do problema de pesquisa. Ele deve incluir o tipo do desenho, as variáveis medidas, as variáveis dependentes e independentes e a instrumentação. Também deve ser incluído como os dados serão coletados e analisados, assim como a categorização dos sujeitos e a validade do experimento.

A execução também deve ser apresentada e tem como objetivo facilitar a replicação do experimento. Primeiramente deve ser apresentada a preparação, explicando como o experimento foi conduzido, a preparação dos sujeitos e se houve algum treinamento ministrado. O processo de validação dos dados coletados também devem ser estressados .

A apresentação da análise deve conter os cálculos juntamente com as hipóteses. O tamanho da amostra, nível de significância, tratamento aplicado as amostras devem ser apresentados para situar os leitores.

Em seguida, a interpretação onde a rejeição da hipótese nula ou sua inabilidade. A interpretação deve sumarizar como os resultados do experimento podem ser utilizados.

Por fim, a conclusão e os trabalhos futuros devem ser apresentados. Neles devem estar presentes as discussões sobre o que foi encontrado além de uma sumarização de todo o experimento. Ideias sobre trabalhos futuros também podem ser apresentadas. Caso necessário, também deve-se adicionar apêndices.

3 METODOLOGIA DA REVISÃO SISTEMÁTICA

A fim de entender o estado da arte sobre os tipos de tarefas em processamento de linguagem natural, uma busca sistemática da literatura foi proposta. Os seguintes parâmetros foram utilizados:

a) **Query de busca**: "word E embedding E train", "word E embedding E LSA E word2vec E glove E fasttext E elmo", e "word E embedding E evaluator";
b) **Motor de busca**: *ieeexplore.ieee.org*, *scholar.google.com.br*, e *semanticscholar.org*;
c) **Variável de exclusão**: ano de publicação menor que 2016.

Para cada documento encontrado, a seguinte tabela foi preenchida:

a) Título do artigo;
b) Data da publicação;
c) O objetivo principal do documento;

Quando o objetivo principal do documento encontrado se relacionava com o objetivo principal desse trabalho, os seguintes campos eram preenchidos:

a) A necessidade da pesquisa;
b) A metodologia apresentada;
c) O resultado final e trabalhos futuros;

Ao final dessa etapa, 22 documentos constavam na tabela. Após uma análise das informações da tabela, 4 documentos foram selecionados como relevantes. Um estudo detalhado desses documentos foi realizado e, através do processo de *snowballing*, mais 6 documentos foram incluídos.

4 REVISÃO DA LITERATURA

4.1 TIPOS DE PROBLEMAS

Russell e Norvig (2009) agrupou as tarefas de processamento de linguagem natural em *classificação de texto*, *recuperação da informação* e *extração da informação*. A *classificação de textos* propõe decidir qual conjunto pré-definido de classes um dado texto pertence. A *recuperação da informação* é a tarefa que busca encontrar documentos que são relevantes a necessidade de informação do usuário. A *extração da informação* é a tarefa de identificar ocorrências de uma classe particular de objetos e identificar as relações entre esses objeto. Bakarov (2018) apresentou uma lista não exaustiva de tipos de problemas utilizados para avaliar a representação de palavras. Essa lista foi utilizada como base da pesquisa dos tipos de problemas no campo de processamento de linguagem natural e o relacionamento com os 3 principais grupos de tarefas foram apresentados.

O problema de *separação de textos em blocos* consiste na divisão do texto em blocos de modo que palavras sintaticamente relacionadas tornam-se parte de um mesmo bloco. Os blocos poder ser, por exemplo, frases substantivas (NP) ou verbais (VP). Os blocos não possuem sobreposição, ou seja, uma dada palavra não pode fazer parte de mais de um bloco. Essa tarefa é útil na etapa de pré-processamento de textos para análise e é uma sub-tarefa do grupo de *extração da informação* (TJONG KIM SANG; BUCHHOLZ, 2000).

O *reconhecimento de entidades relacionadas* é uma sub-tarefa do grupo de *extração da informação* com objetivo de classificar elementos de uma sentença em categorias pré-definidas como pessoas, localização, datas e outras classes. Sistemas mais especializados se concentram em uma gama limitada de entidades dado um domínio de interesse. Essas entidades são marcadas e podem ser utilizadas como um dos primeiros passos para a análise semântica de textos, e uma sub-tarefa para sistemas de gerenciamento de documentos, extração da informação, mineração de textos entre outros (COLLOBERT et al., 2011; CARVALHO, 2012).

Análise de sentimento é um caso particular do problema de *classificação de textos*. Ocorre quando um fragmento de texto precisa ser marcado de forma binária, representando uma polaridade positiva ou negativa. Trabalhos nesse área compreendem a identificação de sentimentos de uma única palavras, sentenças, frases e documentos. Estudos mostraram que a classificação de intensidade de sentimento a nível de frases é importante para tarefas práticas de pergunta e resposta (YESSENALINA; CARDIE, 2011; BAKAROV, 2018).

O *anotador de papeis semânticos*, sub-tarefa da *classificação de textos*, busca recuperar a estrutura argumento-predicado de uma sentença com a proposta principal de determinar "quem fez o que para quem", "quanto"e "onde". Dessa forma, busca-se a identificação das relações semânticas existentes entre o predicado e seus participantes e propriedades através de uma lista pré-definida de possíveis papéis semânticos (HE et al., 2017; CARVALHO, 2012).

A *identificação de negação de escopo* pode ser considerada uma tarefa do grupo de *classificação de texto* e consiste em identificar se uma dada ação em uma determinada sentença é uma negação ou não. Negações podem aparecer de diversas formas, invertendo não somente o significado de uma palavra mas também de um sentença completa. Negações também podem inverter o sentido de sentenças de forma implícita. (ETTINGER; ELGOHARY; RESNIK, 2016; PRÖLLOCHS; FEUERRIEGEL; NEUMANN, 2016).

A *etiquetagem morfossintática* é uma tarefa que consiste em etiquetar palavras, expressões multi-palavras e sinais de pontuação de uma sentença conforme suas categorias gramaticais, como por exemplo substantivos, verbos, adjetivos, etc. Essa tarefa é requirida por outras aplicações de processamento de linguagem natural, tais como análise gramaticas e tradução automática, e por aplicações de processamento de fala, como por exemplo síntese de fala (DOMINGUES, 2011).

As metáforas e paráfrases são inerentes das linguagens naturais e também são tarefas do processamento de linguagem natural. Segundo Gao2018NeuralMetaphoDetection a *detecção de metáforas* pode se dividir em tarefas de *classificação de textos*, onde busca-se a classificação de um verbo em uma sentença como metafórico ou não, e *extração da informação* quando o desafio for indicar a todas as palavras metafóricas independente de sua classificação morfossintática. Conforme Polastri (2016) a *detecção de paráfrases* e seu reconhecimento pode ser utilizada, por exemplo, na tradução automática, onde paráfrases podem ser utilizadas para aumentar a cobertura estatística; na *sumarização de multidocumentos*, onde a identificação de paráfrases permite o reconhecimento de informações repetidas; e na *geração de língua natural*, onde as paráfrases permitem uma maior fluidez e variados.

5 PLANEJAMENTO DO EXPERIMENTO

O objetivo principal da pesquisa é comparar modelos de representação vetorial no domínio biomédico. Assim, O escopo do experimento é analisar os métodos de representação vetorial de palavras, com o propósito de avaliação, em respeito a efetividade, do posto de vista de pesquisadores, no contexto de mestres e doutores em biomedicina.

Devido a dificuldade em modelar matematicamente a qualidade de uma representação vetorial de palavras, a presente proposta de estudo pretende utilizar-se de um estudo empírico. No caso, devido a inviabilidade operacional na escolha e participação dos sujeitos de pesquisa de forma que seja completamente randômica, considera-se um quasi-experimento.

O quasi-experimento buscará um relacionamo entre causa e efeito graças a possibilidade de controlar os possíveis tratamentos de forma singular.

Segundo Bakarov (2018), os métodos de avaliação da qualidade de *word embeddings* podem ser categorizados em intrínsecos e extrínsecos. Os extrínsecos são baseados na habilidade dos modelos serem utilizados como vetores de características em algoritmos de aprendizado de máquina supervisionados em aplicações e então utilizado como medida da qualidade própria da aplicação.

Os métodos intrínsecos são experimentos em que os vetores representacionais são comparados com o julgamento humano em relação a similaridade de palavras ou a analogia entre as palavras. Esses métodos baseiam-se nas pesquisas de semântica distribucional, a qual afirma que existe uma similaridade semântica entre palavras, e palavras com maior relacionamento aparecem em contextos linguísticos semelhantes (LENCI, 2008).

A proposta de pesquisa utilizará os métodos intrínsecos de similaridade de palavras. A similaridade de cosseno é normalmente utilizada para encontrar a distância entre dois vetores Singhal et al. (2001).

Foram escolhidos 2 algoritmos de vetorização, o *Word2Vec* e *Glove*. Sua escolha se deve ao fato desses dois métodos serem de categorias distintas e possuírem grande influência na literatura.

Os algoritmos sob investigação terão como entrada um conjunto de palavras manualmente elaborados e terão como tarefa encontrar um número fixo de palavras similares. O resultado da aplicação dos algoritmos serão armazenados juntamente com o grau de similaridade encontrado, que é dado na escala estatística de razão.

O experimento será realizado de forma off-line. Os sujeitos de pesquisa terão a tarefa de avaliar os resultados produzidos pelos diferentes algoritmos de vetorização, de 0 a 10, sendo 0 a falta de relacionamento e 10 o completo relacionamento, assim utilizarão da escala estatística intervalar. Tais medidas serão subjetivas, já que estão sujeitas as opiniões dos participantes.

A hipótese é que diferentes métodos de representação vetorial possuem diferentes qualidades. Sendo K um número pertencente ao conjunto dos números inteiros, representando uma avaliação de relacionamento entre 2 palavras, entre 0 a 10, sendo 0 a falta de relacionamento e 10 o completo relacionamento, então: $H_0 : \mu_{K_{Word2Vec}} = \mu_{K_{Glove}}$ e $H_1 : \mu_{K_{Word2Vec}} ! = \mu_{K_{Glove}}$.

As variáveis independentes serão os documentos utilizados durante o processo de aprendizado, as palavras escolhidas como entrada das tarefas dos métodos de vetorização, e os métodos de vetorização. As variável dependente será os resultado da representação vetorial e o relacionamento entre palavras.

Os sujeitos serão mestres e doutores em medicina de universidades da região e também profissionais que atuam em hospitais. Não será possível selecionar os sujeitos aleatoriamente, mas sim de determinadas universidades e hospitais. É possível que a especialidade dos médicos possua alguma influência em sua capacidade de distinguir relacionamentos de palavras. Devido a isso, cada participante deverá preencher um formulário de caracterização, indicando sua especialidade. Assim, caso alguma análise estatística indique uma tendência, os resultados poderão ser agrupados de acordo com essa característica.

A princípio, somente um único grupo de sujeitos será utilizado. O experimento será realizado de forma online, porém anotações refentes ao ambiente, como ruído, luminosidade e horário serão anotados. Dessa forma, caso seja identificado que a variação do ambiente possui uma tendência sobre o resultado, tais características serão agrupadas.

O experimento será dividido em 3 etapas, abertura e explicação (10 minutos), realização da tarefa (30 minutos) e fechamento (5 minutos). Na primeira etapa deve-se apresentar os objetivos do experimento de forma clara e objetiva. O termo de consentimento livre e esclarecido deve ser elaborado e também assinado por cada participante. É importante informar ao sujeito que sua participação é opcional e que o mesmo pode deixar de participar do experimento a qualquer momento sem que haja prejuízo. Seus dados serão armazenados de forma confidencial e nenhuma publicação conterá dados sensíveis.

A realização da tarefa, também de forma online e monitorada, terá duração máxima de 30 minutos. Nessa etapa, cada participante recebera uma lista de relacionamentos produzidos pelos algoritmos de vetorização. Para cada palavra de entrada, a saída será embaralhada e os

sujeitos terão como tarefa avaliar esses relacionamento como descrito anteriormente Estima-se que a avaliação dure 1 minuto por conjunto de palavras, e espera-se um total de 30 conjuntos.

Após a coleta dos dados é necessários identificar as exceções utilizando métodos estatísticos descritivos. Assim, os dados serão representados em um histograma e exceções serão identificadas. Uma vez identificadas, deve-se entender se existe alguma explicação para elas, como por exemplo um sujeito de pesquisa possuir uma especialidade singular, e assim identificar se os dados desse sujeito podem ser comparados com os demais. Caso seja identificada a necessidade de exclusão dos dados, devido a singularidade e não relacionamento, os dados removidos devem ser documentados.

Uma vez que o conjunto de dados estiver adequado inicia-se o processo de análise estatística. Nesse caso, deve-se definir se será utilizado testes paramétricos ou não paramétricos. Assim, é necessário verificar se os dados coletados obedecem uma distribuição normal. Para isso será necessário representar os dados em um histograma, sendo que haverá um gráfico por dupla de palavras, o eixo horizontal o conjunto possível de avaliações e o eixo vertical a quantidade de vezes que uma nota foi atribuída pelos diferentes sujeitos de testes.

Como o desenho do experimento utiliza-se de 1 fator (qualidade do relacionamento semântico) com 3 tratamentos pode-se utilizar os métodos *t-test* e *F-test*, caso os dados sigam uma distribuição normal, ou os métodos não paramétricos *Mann-Whitney* e *Chi-2* caso os dados não estejam normalmente distribuídos.

Os resultados obtidos e analisados serão apresentados na forma de uma dissertação de mestrado.

6 ANEXO I - ESTUDO E REFLEXÕES SOBRE ÉTICA NA PESQUISA

Ser ético pode ser definido como "dar o melhor de si, respeitando a vida e a liberdade", ou "não fazer para os outros o que não gostaria que fosse feito para você", ou "fazer como se sua ação pudesse ser algo universal".

Na pesquisa científica, a ética visa garantir o bem-estar do sujeito da pesquisa. O sujeito pode ser tanto humano quanto animais. Certamente os procedimentos com animais são diferentes do que os procedimentos com humanos. A discussão da ética em experimentos científicos se iniciou após a segunda guerra mundial onde experimentos foram realizados em seres humanos causando grandes prejuízos aos sujeitos. Esses experimentos não somente foram realizados durante a segunda guerra, mas também em outras ocasiões.

A ética visa garantir os direitos e deveres dos sujeitos de pesquisa, com princípios de autonomia, não maleficência, beneficência, justiça e equidade. Isto é, um sujeito de pesquisa deve estar ciente da pesquisa que esta participado, dos benefícios que pode acontecer, e dos possíveis malefícios. Essa anuência é dada através do "Termo de Consentimento Livre e Esclarecido", o TCLE, onde informações da justificativa, métodos, riscos, benefícios, esclarecimentos, são apresentados de forma clara e objetiva. É direito do sujeito se recusar a participação ou deixar o experimento em qualquer que seja a fase.

Existem diversos comitês de ética espalhados pelo Brasil. Os comitês, além de possuírem a função de aprovar e acompanhar procedimentos, também possuem papel educativo, assegurando que pesquisadores possam realizar pesquisas científicas de forma ética. Alguns tipos de experimentos, como por exemplo, experimentos com novos fármacos, ou com povos indígenas, devem ser submetidos à apreciação da CONEP (comissão nacional de pesquisa).

Existe uma diferença entre ética e integridade no campo da pesquisa científica. Enquanto a ética observa a pesquisa com foco no ser humano, a integridade é voltada para a forma de condução de uma pesquisa. Por exemplo, um trabalho integro é aquele onde não há plagio, onde os resultados são reais, precisos e completos.

Todas as produções científicas devem ser integras. Porém, nem todas as pesquisas precisam ser submetidas a comissões de ética. Pesquisas que não envolvem seres humanos não a fazem necessário. Uma observação importante são pesquisas onde há coleta de dados com pessoas, essas pesquisas podem precisar serem avaliadas por comissões. A razão é que o risco para uma pessoa pode ser tanto físico, mental quanto espiritual. Ou seja, certos tipos de perguntas de opiniões podem causar prejuízo emocionalmente.

Algumas áreas da pesquisa científica em Engenharia certamente necessitam passar por comissões de ética, e conceitos éticos devem ser levados em consideração. O desenvolvimento de próteses, por exemplo, onde há a experimentação em pessoas ou certos tipos de questionários, são classificados como experimentos em seres humanos e princípios éticos devem ser observados. Ou seja, a utilização de uma prótese não deve causar prejuízo superior a condição atual do sujeito do teste. Tanto a pesquisa quanto o procedimento científico devem ser avaliados por comissões de ética.

Com isso, conclui-se que a ética na pesquisa científica é semelhante tanto na área da Engenharia quanto na área da medicina e saúde. Pois um sujeito de pesquisa deve ser respeitado, deve ser livre para escolher sobre sua participação, conhecer a pesquisa, os benefícios, seus direitos, e a justificativa do experimento.

7 ANEXO II - APLICAÇÃO DO SCRUM EM EXPERIMENTOS CIENTÍFICOS ACADÊMICOS NO ÂMBITO DE MESTRADO E DOUTORADO

Abaixo estão alguns conceitos e sua aplicação no caso de experimentos científicos.

Conceito	**Definição**	**Comparativo com Pesquisa Científica**
Scrum Master	É um facilitador. Ele ajuda a todos os envolvidos a entender os valores do Scrum.	A figura de destaque é o pesquisador. Ele deve possuir o interesse de incentivar a aplicação do Scrum. O orientador também será um facilitador no processo de obtenção do título de mestre, interagindo diretamente com a universidade
Product Owner	É o dono do produto. Ele é o único responsável por definir quais recursos e funcionalidades devem ser construídos e sua ordem. Deve comunicar os objetivos da equipe. Responsável por prover uma visão do produto.	Tanto o orientador quando o estudante podem ser o Product Owner. Caso o projeto de pesquisa tenha sido entregue ao estudante pelo orientador, o orientador pode assumir o papel de product owner. Caso o projeto de pesquisa tenha sido elabora em conjunto, o estudante possui um viés mais forte como Product Owner, já que é de seu interesse a conclusão da pesquisa.
Dev Team	São as pessoas que vão de fato desenvolver o produto. Esse time é responsável por definir como o produto deve ser construído.	O Dev Team é o próprio pesquisador ou conjunto de pesquisadores

Sprint Planning	Ocorre no início de cada sprint e tem como objetivo criar um backlog da sprint. Esse backlog tem como objetivo entregar um incremento do produto, ou seja, uma funcionalidade, um item do product backlog.	Considerando uma relação pesquisar-orientados, a responsabilidade da Sprint Planning é do próprio pesquisador. Ele deve possuir um Product Backlog priorizado, seguindo recomendações do orientador, e realizar o planejamento, criando o Sprint Backlog, com base no que é possível ser realizado dentro do tempo da Sprint. A sprint nesse caso, pode ser entre 2 a 4 semanas. No caso de mestrado, recomenda-se 2 semanas.
Daily Scrum	Reunião diária com os envolvidos no scrum onde busca-se responder 3 perguntas: o que eu fiz ontem? o que eu vou fazer hoje? existe algum impedimento? Desse modo, há uma comunicação clara do estado atual.	A daily scrum deve ser realizada para que seja possível atingir o objetivo de desenvolvimento contínuo. Nesse caso, alterações podem ser realizadas. Sugere-se a utilização de mensagens via celular ou até mesmo e-mail.
Revisão da Sprint	Ocorre ao final da sprint. Possui o objetivo de validar e adaptar o produto que está sendo construído. Verifica-se se o que está sendo feito corresponde ao esperado. Nessa etapa pode ocorrer alterações do product backlog.	A revisão da sprint deve acontecer com todos envolvidos no projeto, seja somente entre o orientador e pesquisados, ou toda a equipe de pesquisadores. Nesse momento, adaptações e alterações podem ser realizadas.
Retrospectiva da Sprint	Possui como objetivo a verificação e adaptações do processo. Nessa etapa é elicitado os pontos positivos e negativos.	A retrospectiva é o momento para adaptar o processo entre os pesquisadores e o orientador. Nesse caso, esse evento pode ser realizado logo após a Revisão da Sprint.

Product Backlog	É a lista de funcionalidades de um determinado produto. O Product Owner é responsável por elaborar essa lista e também priorizá-las.	O product backlog contes as funcionalidades de uma dada pesquisa científica. Por exemplo, seções de capítulos, capítulos inteiros, planejamento de experimento, entre outros. O product backlog é criado pelos pesquisadores sob a supervisão do orientador, se houver.
Sprint Backlog	Tarefas a serem realizadas durante a sprint com o objetivo de fornecer um incremento do produto ao término da Sprint. São criadas pelo time de desenvolvimento.	Deve contar as atividades planejadas de para a conclusão de um incremento da pesquisa. Essas atividades são derivadas do Product Backlog durante a Sprint Planning. Tais atividades são de responsabilidade do time de desenvolvimento, ou seja, os pesquisados.

REFERÊNCIAS

BAKAROV, Amir. A Survey of Word Embeddings Evaluation Methods. **ArXiv**, abs/1801.09536, 2018.

CARVALHO, W. S. **Reconhecimento de entidades mencionadas em português utilizando aprendizado de máquina**. 2012. Dissertação – Instituto de Matemática e Estatística da Universidade de São Paulo.

COLLOBERT, Ronan et al. Natural Language Processing (Almost) from Scratch. **J. Mach. Learn. Res.**, JMLR.org, v. 12, null, p. 2493–2537, nov. 2011.

CONFORTO, E.; AMARAL, D.; SILVA, S. Roteiro para revisão bibliográfica sistemática: aplicação no desenvolvimento de produtos e gerenciamento de projetos. In: CONGRESSO Brasileiro de Gestão de Desenvolvimento de Produto - CBGDP. Porto Alegre - RS, 2011. v. 8.

DOMINGUES, M. L. C. S. **Abordagem para o desenvolvimento de um etiquetador de alta acurácia para o português do Brazil**. 2011. Tese – Universidade Federal do Pará.

ETTINGER, Allyson; ELGOHARY, Ahmed; RESNIK, Philip. Probing for semantic evidence of composition by means of simple classification tasks. In: PROCEEDINGS of the 1st Workshop on Evaluating Vector-Space Representations for NLP. Berlin, Germany: Association for Computational Linguistics, ago. 2016. P. 134–139. Disponível em: <https://www.aclweb.org/anthology/W16-2524>.

FELIZARDO, K. et al. Uma Abordagem Visual para Auxiliar a Revisão da Seleção de Estudos Primários na Revisão Sistemática. In: VI Experimental Software Engineering Latin American Workshop. São Carlos - SP, 2019. v. 6.

HE, Luheng et al. Deep Semantic Role Labeling: What Works and What's Next. In: PROCEEDINGS of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver, Canada: Association for Computational Linguistics, jul. 2017. P. 473–483. Disponível em: <https://www.aclweb.org/anthology/P17-1044>.

KHURANA, Diksha et al. Natural Language Processing: State of The Art, Current Trends and Challenges. **CoRR**, abs/1708.05148, 2017. arXiv: 1708.05148. Disponível em: <http://arxiv.org/abs/1708.05148>.

LENCI, Alessandro. Distributional semantics in linguistic and cognitive research. **Italian journal of linguistics**, v. 20, n. 1, p. 1–31, 2008.

MARSHALL, Christopher; BRERETON, Pearl. Tools to Support Systematic Literature Reviews in Software Engineering: A Mapping Study. **ACM / IEEE International Symposium on Empirical Software Engineering and Measurement**, 2013.

POLASTRI, P. C. **Aprendizado Sem-fim de Paráfrases**. 2016. Dissertação – Universidade Federal de São Carlos.

PRÖLLOCHS, Nicolas; FEUERRIEGEL, Stefan; NEUMANN, Dirk. Negation scope detection in sentiment analysis: Decision support for news-driven trading. **Decision Support Systems**, v. 88, jun. 2016.

RUSSELL, Stuart; NORVIG, Peter. **Artificial Intelligence: A Modern Approach**. 3rd. Upper Saddle River, NJ, USA: Prentice Hall Press, 2009.

SINGHAL, Amit et al. Modern information retrieval: A brief overview. **IEEE Data Eng. Bull.**, v. 24, n. 4, p. 35–43, 2001.

TJONG KIM SANG, Erik F.; BUCHHOLZ, Sabine. Introduction to the CoNLL-2000 Shared Task: Chunking. In: PROCEEDINGS of the 2nd Workshop on Learning Language in Logic and the 4th Conference on Computational Natural Language Learning - Volume 7. Lisbon, Portugal: Association for Computational Linguistics, 2000. (ConLL '00), p. 127–132. Disponível em: <https://doi.org/10.3115/1117601.1117631>.

WANG, Bin et al. Evaluating word embedding models: methods and experimental results. **APSIPA Transactions on Signal and Information Processing**, Cambridge University Press, v. 8, e19, 2019.

WOHLIN, C. et al. **Experimentation in Software Engineering**. Springer-Verlag Berlin Heidelberg, 2012. cap. 1, p. 45–54. Disponível em: <https://www.springer.com/gp/book/9783642290435>.

YAGHOOBZADEH, Yadollah; SCHÜTZE, Hinrich. Intrinsic Subspace Evaluation of Word Embedding Representations. **ArXiv**, abs/1606.07902, 2016.

YESSENALINA, Ainur; CARDIE, Claire. Compositional Matrix-Space Models for Sentiment Analysis. In: PROCEEDINGS of the 2011 Conference on Empirical Methods in Natural Language Processing. Edinburgh, Scotland, UK.: Association for Computational Linguistics, jul. 2011. P. 172–182. Disponível em: <https://www.aclweb.org/anthology/D11-1016>.