SISTEMA AUTONÔMICO PARA DETECÇÃO DE MUDANÇAS EM EVENTOS A PARTIR DE NOTÍCIAS

Douglas Fonseca Alves Paranhos

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia de Sistemas e Computação.

Orientador: Geraldo Bonorino Xexéo

Rio de Janeiro
Setembro de 2018

SISTEMA AUTONÔMICO PARA DETECÇÃO DE MUDANÇAS EM EVENTOS A PARTIR DE NOTÍCIAS

Douglas Fonseca Alves Paranhos

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Examinada por:

Prof. Geraldo Bonorino Xexéo, D.Sc.

Prof. Jano Moreira de Souza, Ph.D.

Prof. Leandro Guimaraes Marques Alvim, D.Sc.

RIO DE JANEIRO, RJ – BRASIL
SETEMBRO DE 2018

Paranhos, Douglas Fonseca Alves

Sistema autonômico para detecção de mudanças em eventos a partir de notícias/Douglas Fonseca Alves Paranhos. – Rio de Janeiro: UFRJ/COPPE, 2018.

XII, 57 p.: il.; 29, 7cm.

Orientador: Geraldo Bonorino Xexéo

Dissertação (mestrado) – UFRJ/COPPE/Programa de Engenharia de Sistemas e Computação, 2018.

Referências Bibliográficas: p. 42 – 51.

1. Detecção e Rastreio de Tópicos. 2. Sistemas Autonômicos. I. Xexéo, Geraldo Bonorino. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Título.

Dedicado à
minha família.

Agradecimentos

Gostaria de agradecer primeiramente à minha família , principalmente minha mãe Maria Auxiliadora e meu pai Lincoln, que sempre me apoiaram, incentivaram e deram suporte para que este trabalho fosse realizado.

Gostaria de agradecer também ao meu orientador Geraldo Bonorino Xexéo, pela enorme paciência e conselhos de extremo valor, não somente acadêmicos.

Ao amigo e professor Fabrício Raphael Pereira, que teve parte de seu trabalho utilizado pelo meu e quem mais contribuiu saneando dúvidas e ajudando na construção do mesmo, tendo uma paciência quase de monge.

Aos amigos e companheiros de CAPGOV, Juan Baptista, Bernardo Blasquez, Rômulo Freires e Augusto Acioli. Aos também companheiros de CAPGOV Caroline Lima, Jéssica Costa e Edberg Franco pelo incontáveis almoços e em especial à Débora Lima, Xiao Yuan Kong pelos valorosos conselhos e ajudas.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

SISTEMA AUTONÔMICO PARA DETECÇÃO DE MUDANÇAS EM EVENTOS A PARTIR DE NOTÍCIAS

Douglas Fonseca Alves Paranhos

Setembro/2018

Orientador: Geraldo Bonorino Xexéo

Programa: Engenharia de Sistemas e Computação

Detecção e Rastreio de Tópicos (TDT) tem sido um tema de bastante pesquisas desde que foi definido no final dos anos 90 e começo dos anos 2000 e tem por objetivo identificar eventos do mundo real a partir de informação não-estruturada. Computação Autonômica, do mesmo modo, também tem crescido bastante à partir dos anos 2000 e é designado para sistemas que tem capacidade de medir seu próprio desempenho automaticamente, sendo aplicado nas mais modernas tecnologias. Muitos trabalhos foram desenvolvidos em ambos os temas, porém poucos que unissem estes dois importantes conceitos, reduzindo assim a necessidade de intervenção humana na importante tarefa de analisar informações não-estruturadas. O presente trabalho tem por objetivo criar um sistema autonômico para detecção de modificações em eventos a partir de notícias.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

AUTONOMIC SYSTEM FOR CHANGE DETECTION IN EVENTS FROM NEWS

Douglas Fonseca Alves Paranhos

September/2018

Advisor: Geraldo Bonorino Xexéo

Department: Systems Engineering and Computer Science

Topic Detection and Tracking (TDT) has been a topic of many researches since it was defined in the late 90's and early 2000's and the main goal is to identify real-world events from non-structured information. Autonomic Computing, in the same way, has been growing since the early 2000's and is designated for systems which are capable of measuring its own performance automatically, used in latest and modern technologies. Many works were developed in both topics, nevertheless only a few unite these two important concepts, minimizing human intervention to analyze non-structured information. The present work aims to create an autonomic system for change detection in events from news articles.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

1.1 Motivação

Com o advento da internet, a informação tem se mostrado cada vez mais abundante e o acesso à mesma foi facilitado de inúmeras maneiras, seja por blogs, microblogs, sites de notícias online, enriquecendo e aumentando ainda mais a variedade de fontes já existentes. Com essa explosão de informação, em todo seu volume, variedade e velocidade, acompanhar acontecimentos do mundo real noticiados, tanto em meio virtual como físico, fica dificultado, à medida que um número extremamente grande de fatos acontecem todos os dias. Tendo em vista as dificuldades citadas, uma área de conhecimento foi criada nos anos 90 nos EUA denominada *Topic Detection and Tracking* (TDT).

Acrescido à dificuldade inerente de acompanhar eventos do mundo real pelos motivos já citados, torna-se ainda mais complicado detectar quando um evento se modificou e montar uma linha do tempo dos eventos apontando quando o evento foi modificado, fazendo assim com que o mesmo seja compreendido com facilidade.

Alguns trabalhos no sentido de montar uma *timeline* do evento, ou agrupando notícias de eventos já foram desenvolvidos, tais como Acrópolis (SCHNEIDER, 2015), o *LeadLine* (DOU *et al.*, 2012) e também o *Event Registry* (LEBAN *et al.*, 2014a). Acrópolis é uma rede social onde o usuário manualmente monta uma linha do tempo de notícias, e por ser uma rede social, usuários podem interagir de alguma forma com linhas do tempo feitas por outros usuários. Já o *Event Registry*, que será melhor explicado mais adiante, agrupa diferentes notícias de diferentes lugares do mundo e em diferentes línguas quando elas tratam de um mesmo evento, disponibilizando-as em uma *API*. E, por fim, o *LeadLine*, que é um sistema de análise visual interativa de eventos que foram identificados pelo mesmo.

1.2 Objetivo

O presente trabalho tem por objetivo monitorar eventos acontecidos no mundo real que foram noticiados e automaticamente identificar quando eventos sofrerem modificação.

Para atingir este objetivo utiliza a técnica de identificar respostas relativas ao evento tais como: Quem estava envolvido? (*Who*?), Quando o evento aconteceu? (*When*?), Onde o evento aconteceu? (*Where*?), O que aconteceu? (*What*?), Como aconteceu? (*How*?) (5W1H) (SMITH, 2012; WOLFE, 2010) e, monitorando estes atributos, identificar quando algum deles se modifica, reconhecendo então que algum dos pilares do evento foi alterado, decretando assim a mudança do evento.

1.3 Estrutura da Dissertação

O restante desta dissertação está estruturado da seguinte forma. Capítulo 2 faz uma breve revisão sobre Detecção e Rastreio de Tópicos, Sumarização, *Concept Drift* e Redes Neurais. Capítulo 3 revisa o conceito de Sistemas Autonômicos. Capítulo 4 apresenta a proposta do trabalho, além da implementação da mesma. O capítulo 5 discute seus resultados. Finalmente o capítulo 6 apresenta a conclusão e trabalhos futuros.

Capítulo 2

Revisão da Literatura

Este Capítulo tem por objetivo discutir assuntos utilizados no trabalho, dar uma visão geral do tema além de se aprofundar nos pontos efetivamente utilizados na dissertação.

2.1 Topic Detection and Tracking

Em 1997 um estudo patrocinado pela Agência de Projetos e Pesquisa de Defesa Avançados (do inglês, DARPA) dos EUA teve início e tinha por objetivo criar um sistema capaz de monitorar e alertar especialistas de acontecimentos específicos que eram de interesse dos mesmos (ALLAN, 2002).

Deste estudo gerou-se o conceito de nome Detecção e Rastreio de Tópicos (do inglês, *Topic Detection and Tracking*), ou simplesmente TDT, e inicialmente era composto por três tarefas básicas: segmentar um fluxo de dados, tais como texto e principalmente áudio e vídeo, em diferentes histórias, identificar dentre essas histórias quais abordam um novo evento ainda não noticiado e identificar quando uma notícia nova no *stream* pertence a um tópico já previamente identificado (ALLAN *et al.*, 1998).

TDT, por ser baseado em eventos, possui muitas técnicas similares ou que foram derivadas de Filtragem de Informação (do inglês, *Information Filtering*) (ALLAN *et al.*, 2000). O estudo piloto foi realizado de duas formas: NED (do inglês, *New Event Detection*), que posteriormente também foi chamado de detecção online e RED (do inglês, *Retrospective Event Detection*) também conhecido como detecção offline.

A abordagem NED trata-se das notícias chegando em forma de *streaming*, classificando assim cada notícia do *stream* conforme fossem chegando em alguma história já encontrada ou em uma história completamente nova. Já a abordagem RED, como o escopo é fechado (todo o conjunto de notícias é conhecido à priori, diferentemente do NED) apenas define em tópicos os documentos existentes (ALLAN *et al.*, 1998).

Ao final do estudo, foram identificados cinco tópicos pertinentes à TDT, que serão aprofundados mais adiante: *Segmentação de Histórias, Clusterização, Detecção de Primeiras Histórias, Monitoramento* e *Detecção de Ligação entre Histórias* (ALLAN, 2002).

2.1.1 Segmentação de Histórias

Esta tarefa tem por objetivo detectar corretamente o fim de uma notícia e o início de outra em um contínuo *streaming* de notícias, de áudio e vídeo, principalmente. Ou seja, deve-se delimitar corretamente cada história, em um fluxo das mesmas. (ALLAN *et al.*, 1998) HAUPTMANN & WITBROCK (1998), CARBONELL *et al.* (1999), HSU *et al.* (2003) e ROSENBERG & HIRSCHBERG (2006) apresentam abordagens para o cumprimento da presente tarefa com relação à vídeo. Utilizam informações no vídeo, no áudio e nas legendas dos áudios.

HAUPTMANN & WITBROCK (1998) propõe um modelo para detectar intervalos comerciais onde utiliza quatro fontes de informação: processamento de imagem, *speech recognition*, legendas (*closed-caption*) e processamento de áudio. Relativo ao texto, é baseado basicamente em BEEFERMAN *et al.* (1997) onde é apresentado um modelo probabilístico que separa um texto em trechos coerentes e em HEARST & PLAUNT (1993) que separa regiões dentro de um parágrafo e sejam coerentes dentro de um tópico.

CARBONELL *et al.* (1999) se baseia em BEEFERMAN *et al.* (1999) e treina um modelo probabilístico onde leva em consideração as palavras em si e probabilidades de palavras aparecerem juntas em um mesmo contexto ou contextos diferentes.

HSU *et al.* (2003) também utiliza um modelo probabilístico porém trabalha com a ideia de "pontos candidatos" que basicamente são possíveis pontos onde é possível que haja uma troca de histórias, já ROSENBERG & HIRSCHBERG (2006) apesar de também trabalhar com a ideia do "ponto candidato", acrescenta a ideia de "fronteira da frase", buscando assim quando os pontos candidatos e fronteiras das frases coincidem.

2.1.2 Clusterização

O objetivo desta tarefa é detectar a qual tópico previamente detectado a história atual pertence (ALLAN, 2002). É basicamente feita de algoritmos de clusterização, onde a história atual é identificada pertencendo a algum *cluster* já encontrado baseado em determinadas características do evento, portanto pertencente a um tópico previamente estabelecido.

É possível a utilização de praticamente qualquer algoritmo de clusterização para essa tarefa, clusterização *fuzzy* (NEO *et al.*, 2006), *k-means* (ALLAN, 2002) e cus-

tomizados (SHI *et al.*, 2017), além de poder ser utilizado em qualquer meio que se deseja a detecção de eventos, tais como video (NEO *et al.*, 2006), notícias textuais (NALLAPATI *et al.*, 2004; SETTY *et al.*, 2017) e conteúdos de mídias sociais (SHI *et al.*, 2017).

2.1.3 Detecção de Primeiras Histórias

O objetivo desta tarefa é receber uma história e marcá-la como sendo primeira ou não-primeira em um determinado tópico (ALLAN *et al.*, 2000). ALLAN *et al.* (1998) utiliza a abordagem tradicional de representar cada notícia em um espaço vetorial e, calculando a similaridade utilizando o tradicional TF-IDF, verifica a similaridade da história nova com as já encontradas. Caso esteja fora de um certo limite é considerada primeira história.

KUMARAN & ALLAN (2004) estuda o impacto da utilização de algoritmos de classificação de texto bem como quando reconhecimento de entidades nomeadas podem ser úteis ou não para esta tarefa. PETROVIĆ *et al.* (2010) desenvolve um algoritmo baseado em *Locality-Sensitive Hashing* (LSH) para esta tarefa, e em PETROVIĆ *et al.* (2012) este algoritmo é aprimorado com a introdução de paráfrases. STOKES & CARTHY (2001) aborda o problema construindo dois classificadores, um semântico e um sintático, utilizando uma combinação dos dois.

2.1.4 Monitoramento

Dado um tópico com N histórias, monitorar as histórias subsequentes do tópico. Esta tarefa é semelhante à Filtragem de Informação. (ALLAN *et al.*, 2000)

YANG *et al.* (2000) aborda o problema testando diversas variações do algoritmo de classificação de k vizinhos mais próximos (kNN), bem como a classificação de Rocchio. FUKUMOTO & SUZUKI (2000) propõe um algoritmo que utiliza um critério chamado por ele de "dependência do domínio", onde cada termo tem sua relevância para o tópico levada em consideração para esta tarefa, enquanto YAMRON *et al.* (1998) utiliza *Hiden Markov Model* (HMM) como solução.

2.1.5 Detecção de Ligação entre Histórias

Determinar se duas histórias têm um mesmo evento como assunto. Detecção de Ligação entre Histórias e Detecção de Primeiras Histórias, apesar de parecerem, não são a mesma tarefa, CHEN *et al.* (2003); FARAHAT *et al.* (2003) comprovam esta afirmação aplicando vários métodos como *Part of Speech Tagging (PoS tagging)*, métricas de similaridades, lista de *stopwords* extendida, do mesmo modo que métricas de avaliação como Precisão e Revocação, tendo resultados diferentes para ambas

tarefas. CHEN *et al.* (2004) testa uma combinação de métodos e métricas de similaridades, como *Support Vector Machine* (SVM), árvores de decisão, etc, e métricas de similaridade como distância dos Cossenos, distância de *Hellinger* (BERAN *et al.*, 1977), distância de Tanimoto (ROGERS & TANIMOTO, 1960), entre outros.

SHAH *et al.* (2006) utiliza reconhecimento de entidades nomeadas para Detecção de Ligação entre Histórias, enquanto ŠTAJNER & GROBELNIK (2009) emprega resolução de entidades nomeadas para este fim. BROWN (2002) utiliza o método básico, apresentado por ALLAN *et al.* (1998) porém aplica um diferente método de remoção de *stopwords*, sendo esta lista de *stopwords* não uma lista estática, mas sim dinâmica. SHAH & EGUCHI (2009) testa um novo tipo de pesos para termos baseado na importância do mesmo não só para a coleção, mas sim para o tópico em que ela se inclui.

2.1.6 Evolução de Eventos

Introduzido por MAKKONEN (2003) o conceito de Evolução de Eventos tem por base entender e identificar como eventos evoluem ao longo do tempo. No método proposto por MAKKONEN (2003) o evento é decomposto em uma tupla de 4 vetores, e é representado da seguinte forma <TERMOS, LOCALIZAÇÕES, NOMES, TEMPORAIS>, onde cada vetor desta tupla é determinado da seguinte maneira:

- A lista de termos possui elementos que podem indicar o que aconteceu no evento, representando o atributo *WHAT* dos 5W1H.
- A lista de localizações contém todas as informações de locais encontrados no texto, corresponderia, teoricamente ao atributo *WHERE* dos 5W1H.
- A lista de nomes abrange os nomes próprios do documento, correspondendo portanto, ao atributo *WHO* dos 5W1H.
- A última lista, de temporais, envolve todos termos referentes a datas e tempo de alguma forma, equivalendo ao atributo *WHEN* dos 5W1H.

Com a representação do evento em mãos, calcula a similaridade de cada atributo levando em consideração a especificidade de cada um. E finalmente compara 2 representações de evento calculando se a média dos atributos de algum evento é maior que algum valor fixado.

DEL CORSO *et al.* (2005) propõe um método em que é elaborado um ranking de notícias em um *stream*, detectando assim as mais representativas notícias de acordo com o tempo, podendo por conseguinte, manualmente ser detectadas mudanças em um evento, porém não é detectada a evolução pelo *framework*; um ser humano teria que acessar o ranking gerado e tirar suas conclusões.

WEI & CHANG (2007) trabalha com o conceito de “intra-eventos” e “extra-eventos”. Utiliza um método que basicamente utiliza a clusterização para buscar generalizações entre eventos usando como base as características de cada documento.

ZHAO *et al.* (2008) desenvolve um método de *queries*, que nada mais são que um vetor com os termos que definem o evento, variáveis, onde ele constantemente atualiza a *query* de acordo com os documentos e, de acordo com 2 limiares arbitrários, calcula a similaridade entre o evento e o documento. Caso a similaridade seja menor que ambos limiares a notícia não pertence ao evento, caso seja maior que ambos a notícia pertence ao evento, e, se porventura, estiver entre ambos, a *query* é recalculada.

WEI *et al.* (2009) apresenta uma métrica que ele chama de “TF-IDFtempo” para extração e seleção de *features* para caracterização do evento onde essa métrica é a clássica TF-IDF que leva em consideração o tempo, sendo diretamente proporcional à “idade” dos termos, quanto mais antigo, menor é seu valor.

DENG *et al.* (2011a) adota uma estratégia de dividir eventos em “eventos atômicos”, dividindo um documento em um ou mais eventos atômicos, identifica a correlação e similaridade entre eles e constrói um grafo não direcionado.

GRUENHEID *et al.* (2015) apresenta o *framework StoryPivot* onde procura identificar evolução de eventos sob diferentes domínios e perspectivas. O processo é dividido em duas fases: Identificação de Histórias, onde o evento é divido em <CARACTERÍSTICAS, FONTE, *TIMESTAMP*>, sendo o *timestamp* a data em que a notícia foi coletada, fonte é de onde foi retirada e características pode ser uma grande variedade de características, como entidades nomeadas, título, etc, e Alinhamento de História, segunda fase do processo, onde é utilizada uma janela deslizante de tempo para comparação das novas histórias identificadas com as anteriores.

Alguns artigos utilizam o mesmo termo "Evolução de Eventos", porém com o significado ligeiramente diferente do original daquele apresentado por MAKKONEN (2003).

YANG & SHI (2006) elabora um grafo de evolução de eventos que nada mais é um grafo direcionado onde cada vértice é um evento e uma aresta é uma relação entre dois eventos. É calculada a similaridade dos cossenos entre 2 eventos e a proximidade temporal entre eles para que tal objetivo seja alcançado.

DENG *et al.* (2013) utiliza textos de *microblogging* para avaliar como a opinião pública influencia em cadeias de eventos.

LI *et al.* (2014) aborda o tema diferenciando eventos simples, de eventos complexos (que também são chamados de história por outros autores como em ALLAN (2002)). Ele busca gerar um grafo de relacionamento entre “eventos simples” procurando expressões chave como “é influenciado por”, “causado por”, etc. E com base nessas informações monta o grafo de relacionamento entre eventos.

CAI *et al.* (2015) não utiliza notícias, mas sim *tweets* do *Twitter* e considera 4 tipos de eventos e suas características, criação, absorção, separação e junção para determinar a dinâmica e evolução de eventos relatados pelo *Twitter*. Utiliza uma lista invertida de eventos e *tweets* relativos ao evento, classificando-o utilizando o método de classificação kNN.

HUANG *et al.* (2018) aborda o tema dividindo o problema em duas partes: a primeira verifica a similaridade dos documentos combinando aspectos temporais, aspectos textuais e entidades presentes, e em seguida monta um grafo com nós, que são *clusters* de documentos, e finalmente gera uma linha do tempo com uma notícia representativa de cada *cluster*.

Alguns trabalhos não utilizam o conceito de Evolução de Eventos, entretanto trabalham com eventos que evoluem no tempo.

YANG *et al.* (2014a) procura identificar padrões em evoluções de eventos, ou seja, estágios de evoluções em eventos, tais como estágios em proliferação de doenças, etc, e classificar esses padrões, para tal gera um modelo probabilístico levando em consideração características dos eventos, sendo o temporal o principal.

İLHAN & ÖĞÜDÜCÜ (2015) identifica como comunidades evoluem em redes sociais baseado em eventos, como crescimento, sobrevivência, etc. Identifica para cada evento características do mesmo e monitora comunidades identificando semelhanças entre a comunidade e os eventos.

MELE *et al.* (2017) aplica *Latent Dirichlet Allocation* (LDA) para a detecção de tópicos em fatias do *stream* de notícias, e os tópicos descobertos são ligados utilizando uma *Hidden Markov Model* (HMM), determinando assim probabilidades de transição de um tópico para outro.

DOU *et al.* (2012) cria um sistema chamado *LeadLine*, que é uma ferramenta interativa de visualização para eventos extraídos tanto de notícias como de dados redes sociais, ele aplica LDA com a intenção de detectar tópicos na coleção e de acordo com a quantidade de palavras chave em um curto período de tempo, detecta o evento. Também aplica reconhecimento de entidades nomeadas a fim de completar a tupla <TÓPICO, TEMPO, PESSOAS, LUGARES>.

NALLAPATI *et al.* (2004) utiliza alguns algoritmos de clusterização com diferentes tipos de métricas possíveis para a identificação de relacionamento entre eventos.

Não há *dataset* padrão para avaliar ou mensurar as técnicas de Evolução de Eventos, nem mesmo uma padronização de métricas para avaliação, embora grande parte utilize as tradicionais Precisão, Revocação e *f-measure* como métricas.

Cada autor coletou manualmente seu *dataset* durante uma quantidade de tempo variável, com quantidade de documentos variável, das mais diversas fontes como *BBC News World* (DEL CORSO *et al.*, 2005), *The New York Times* (WEI & CHANG, 2007), CNN (DOU *et al.*, 2012; YANG & SHI, 2006), *Washington Post*

(WEI & CHANG, 2007), *Financial Times* (WEI & CHANG, 2007), portal de notícias da *Wikipedia* (HUANG *et al.*, 2018) e tópicos como finanças (WEI & CHANG, 2007), desastres naturais (DENG *et al.*, 2011a; ZHAO *et al.*, 2008), alguns usaram notícias em outros idiomas que não o inglês (DENG *et al.*, 2011a; HUANG *et al.*, 2018; LI *et al.*, 2014; ZHAO *et al.*, 2008), e alguns sequer chegaram a avaliar (GRUENHEID *et al.*, 2015; MAKKONEN, 2003).

2.2 Sumarização

Sumarização consiste em formas de gerar um sumário, ou seja, aplicar alguma técnica de sumarização automática em um texto tal que, ao final dela, deve-se ter a maior quantidade de informação possível do(s) documento(s) original(is) com menos frases.

Uma técnica de sumarização automática, em relação ao número de documentos, pode trabalhar com vários documentos ao mesmo tempo ou um documento apenas. Quanto ao tipo de texto gerado no sumário, a técnica pode ser extrativa ou abstrativa, onde extrativa é quando a mesma escolhe sentenças do texto do(s) documento(s) original(is) e as retorna como sendo o sumário do documento, enquanto a abstrativa, é gerado um sumário com palavras-chave e conceitos do documento original, porém em um texto artificialmente produzido pela técnica. Ainda sobre o texto produzido, pode ser genérico ou focado em *query*, este é gerado direcionado à uma *query* indicando quais aspectos que devem aparecer no sumário, já aquele, como o próprio nome diz, é genérico. As técnicas ainda podem ser supervisionadas ou não-supervisionadas, além de trabalhar com somente uma língua ou várias línguas (GAMBHIR & GUPTA, 2017).

Em relação às abordagens das técnicas extrativas, podem ser de 5 tipos (ou mistura entre elas), segundo GAMBHIR & GUPTA (2017):

- Baseado em dados estatísticos: Leva em consideração dados do texto, tais como dados de cada frase, por exemplo posição da mesma, semelhança com o título do texto, presença de entidades nomeadas, presença de palavras consideradas positivas ou negativas, tamanho relativo, entre tantos outros. Cada frase, baseado nos dados apresentados, obtém uma pontuação e, alicerçado nessa pontuação além da taxa de compressão (que define o quanto se deve considerar ou descartar do texto), extrai as frases do texto que melhor constituirão o sumário. Este tipo tem como virtude não necessitar de muito processamento computacional e, por levar somente dados estatísticos do texto em si, pode ser utilizado em qualquer língua.

- Baseado em Aprendizado de Máquina: Utiliza técnicas de Aprendizado de Máquina para aprender e melhorar a montagem do sumário, podendo estas serem

supervisionadas, semi-supervisionadas ou ainda não-supervisionadas. Há registros na literatura de usos de *Support Vector Machine* (SVM), Regressões Lineares e Logísticas, Classificador de *Naïve Bayes*, Redes Neurais, Árvores de decisão, entre outros, para os casos dos métodos supervisionados, *Hidden Markov Model* (HMM) e clusterização para os não-supervisionados.

- Baseado em tópicos: Sobre o que o documento trata. Alguns tratamentos deste modo existem, como tratar um tópico como um conjunto de termos representativos deste ou ainda tratá-lo como um conjunto de entidades nomeadas e datas, etc.

- Baseado em grafos: Monta um grafo onde cada frase ou palavra é um vértice de um grafo e arestas relacionam nós que são semanticamente conectados e então o problema torna-se basicamente definir importâncias para cada vértice. Quando um vértice se liga a outro vértice ele está, a grosso modo, recomendando o outro vértice, então quanto mais recomendações um vértice possuir, melhor posicionado e mais relevante ele é para o sumário (MIHALCEA & TARAU, 2004).

- Baseado no discurso: Relações no discurso são consideradas aqui, que nada mais são do que conexões entre partes e frases do texto. São avaliados aspectos tais como coesão e abordagens linguísticas.

Quanto à avaliação da sumarização, podem ser levados em consideração alguns itens que se deseja mensurar, sendo a métrica *ROUGE* (LIN, 2004), que compara o sumário gerado automaticamente com sumários criados por humanos levando em consideração palavras em comum, N-Gramas, entre outros, a mais amplamente utilizada. As tradicionais métricas de Recuperação da Informação, Precisão, Revocação e *f-measure* também são bastante utilizadas.

2.2.1 Métodos baseados em dados estatísticos

Esses métodos possuem em comum usar apenas características dos textos, e utilizando uma fórmula que leva em consideração essas características, atribui pontuações a cada sentença.

KO & SEO (2008) apresenta um método dividido em duas partes: a primeira divide o documento em um bi-grama de pseudo-sentenças (que são sentenças adjacentes no texto) levando assim informações contextuais em consideração e utiliza dados estatísticos em consideração para selecionar os mais proeminentes bi-gramas e são divididos em 2 pseudo-sentenças. A fase seguinte do método é constituída da escolha das melhores, entenda-se por melhores as com maior pontuação, pseudo-sentenças escolhidas para a montagem do sumário final.

FERREIRA *et al.* (2013) identifica, implementa, compara e avalia, tanto qualitativamente quanto quantitativamente 15 algoritmos de geração de pontuações para sumarizadores extrativos. Os diversos algoritmos são testados em alguns *datasets*, entre eles um *dataset* de notícias.

2.2.2 Métodos baseados em grafos

Esses algoritmos têm em comum basicamente os seguintes passos:

- identifica unidades de texto (que podem ser frases, palavras, etc) e os adiciona em vértices de um grafo, sendo cada vértice uma dessas unidades.
- Identifica relações entre essas unidades de texto e cria arestas, que podem ser dirigidas ou não-dirigidas e ponderadas (conter pesos) ou não, entre os respectivos vértices.
- Itera no algoritmo baseado em grafos até um determinado ponto de parada.
- Ordena os vértices de acordo com cada pontuação, sendo os mais representativos do documento os que obtiverem maior pontuação.

MIHALCEA & TARAU (2004) desenvolve um método chamado *TextRank* onde é aplicado o *PageRank* (BRIN & PAGE, 1998) para o contexto de sumarização. Primeiro é feita a tokenização do texto e são anotados papéis semânticos, depois palavras são consideradas candidatas a vértices de um grafo não-dirigido e não-ponderado. Em seguida, é utilizada a correlação entre palavras, seguindo o simples conceito de distância entre elas no documento em uma janela de tamanho mínimo 2 e máximo 10. Em seguida o algoritmo do *PageRank* é utilizado com mesmos parâmetros e finalmente são escolhidos os X melhores vértices, sendo esse X um percentual do tamanho total do texto. BARRIOS *et al.* (2016) propõe variações na forma de cálculo da similaridade a fim de melhorar a performance do método.

ERKAN & RADEV (2004) apresenta o *LexRank* que é uma sumarização multi-documento que aplica uma clusterização nos documentos existentes e calcula o centróide de cada *cluster* encontrado, que contém as palavras com maiores IDF de toda a coleção. Finalmente calcula quais documentos e sentenças possuem mais palavras do seu respectivo centróide, montando um grafo não direcionado e ponderado. Escolhendo, por fim, as sentenças mais bem pontuadas.

BARALIS *et al.* (2013) introduz uma abordagem também multi-documento chamada *GraphSum*, que além de aplicar o básico de pré-processamento (lematização, remoção de *stopwords*, etc) considera cada *cluster* como termos que têm alta correlação, utiliza uma variante do *PageRank* para determinar a pontuação de cada nó e por fim seleciona as sentenças com maior *score*.

2.2.3 Métodos baseados em *Machine Learning*

Os métodos dessa categoria utilizam pelo menos uma técnica de aprendizado supervisionado ou não-supervisionado e geralmente são usados para trabalhar características dos textos.

YEH *et al.* (2005) propõe duas técnicas e utiliza uma combinação de ambas para sumarizar o texto. A primeira trata-se de um classificador (treinável) que leva em consideração algumas características, tais como, posição da sentença, semelhança entre a sentença e o título, etc, que, utilizando-se de uma função que é treinada por meio de um algoritmo genético, gera a pontuação, sendo as mais bem pontuadas integrantes do sumário. A segunda utiliza *Latent Semantic Analysis* (LSA) para produzir uma matriz semântica do documento e utiliza esta matriz para gerar um mapa de relacionamento semântico do documento.

FATTAH & REN (2009) traz um gerador de sumário que considera os seguintes dados estatísticos de cada sentença: posição, palavras positivas, palavras negativas, centralidade, semelhança com o título, se inclui Entidade Nomeada ou não, a presença de dados numéricos, tamanho relativo, entre outros e, de posse desses dados, treina um Algoritmo Genético e uma Regressão para definir os pesos de cada um desses elementos. Também utiliza esses componentes para treinar uma Rede Neural *FeedForward*, uma Rede Neural Probabilística, além de um Modelo de Mistura Gaussiana, gerando um sumarizador para cada modelo.

KIKUCHI *et al.* (2014) cria um método em que utiliza árvores semânticas aninhadas, onde cada documento é representado por duas árvores sendo uma a árvore de documento e outra a árvore de sentenças. Aquela possui sentenças como nós e seus relacionamentos, enquanto esta trata-se da árvore onde os nós são palavras e seus relacionamentos. Com estas estruturas montadas, a tarefa de sumarizar torna-se um problema de otimização computacional de aparar a árvore, preservando o máximo de conteúdo do texto, com o menor tamanho possível.

MENDOZA *et al.* (2014) utiliza um método para sumarização de um documento por vez, um algoritmo genético que utiliza os tradicionais dados estatísticos já mencionados amiúde neste capítulo como características.

YANG *et al.* (2014b) propõe um aprimoramento de sumários baseados em tópicos, onde essa melhora se dá em criar um modelo probabilístico para o *cluster* de cada tópico, fazendo assim com que a tarefa de pegar a sentença (ou palavra-chave) mais representativa de cada *cluster* seja facilitada.

FATTAH (2014) elabora um método que utiliza uma combinação de 3 classificadores: SVM, *Naive Bayes* e Máxima Entropia, onde cada um é trabalhado em cima das características do texto, e uma mescla dos mesmos são utilizados para a realização da sumarização.

2.3 *Concept Drift*

Trata-se da mudança da distribuição dos dados em ambientes de constantes mudanças (SCHLIMMER & GRANGER, 1986; WIDMER & KUBAT, 1996). Existem quatro tipos de *Concept Drift* e cada vertente pode ser abordado de algumas formas.

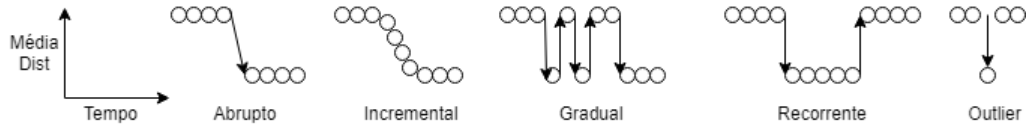

Figura 2.1 – Tipos de *Concept Drift*. Fonte: Adaptado de GAMA *et al.* (2014)

O tipo abrupto acontece quando um conceito, medida, sensor, etc, muda seus valores repentinamente, como um sensor de velocidade que começa a apresentar um defeito e começa a marcar a velocidade máxima do automóvel constantemente. Incremental é o tipo onde ao mudar os valores, ele passa por estágios intermediários, mudando portanto de forma onde cada vez fica mais próximo do valor final, um exemplo seria um velocímetro de um carro que vai freando e o automóvel para por completo. O formato gradual é quando a troca de conceitos não possui valores intermediários mas é feita aos poucos, e por fim, recorrente é quando o conceito de tempos em tempos volta a ser encontrado, como por exemplo, fases da lua ou marés do mar. A figura ainda traz um exemplo de *outlier*, que não é considerado *Concept Drift*.

Métricas de avaliação de *Concept Drift* podem ser algumas, dependendo do que o sistema se propõe a monitorar: Precisão e Revocação, *f-measure* para sistemas com tarefas como a de busca e recuperação da informação, *RSME* em sistemas de recomendação, *MAE* para regressões, probabilidade de falso positivo, demora na detecção, etc.

GAMA *et al.* (2014) explicita e caracteriza métodos para lidar com os vários tipos de *Concept Drift*, que podem possuir as seguintes características:

2.3.1 Quanto à memória

Este módulo especifica como o método trata os dados a serem levados em consideração para o aprendizado. Por sua vez é dividido em duas sub-áreas: Dados e Mecanismo de Esquecimento.

- Dados: Quais dados serão levados em consideração para a elaboração de um modelo preditivo. Tem por base o pensamento que dados mais recentes têm maior relevância. Pode ser de duas formas: Utilizar uma instância por vez, ou vários, que seriam organizados por uma janela de tempo de tamanho fixo ou variável.

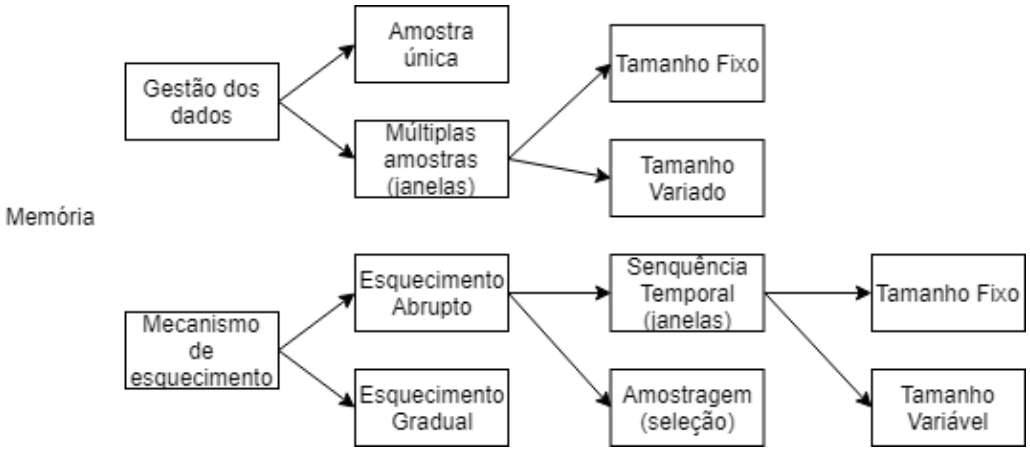

Figura 2.2 – Tipos de tratamento de *Concept Drift* relativos à memória. Fonte: Adaptado de GAMA *et al.* (2014)

- Mecanismo de Esquecimento: Sinaliza como o sistema lidará com dados obsoletos. Este tópico é de suma importância, já que um modelo que tenha curta memória será muito suscetível a ruídos, enquanto um que tenha memória muito longa demoraria a detectar mudanças em demasia. É dividido em dois tipos: Esquecimento abrupto, quando os dados são simplesmente desconsiderados e o gradual, quando nada é esquecido, porém os dados recebem pesos, sendo os mais antigos ganhando pesos menores em relação aos mais recentes, até que, no longo prazo, praticamente se tornam irrelevantes.

2.3.2 Quanto à detecção de mudanças

Este módulo contém as técnicas que detectam a mudança nos dados explicitamente e são baseados em 4 tipos: análise sequencial, controle estatístico, diferenças entre duas janelas temporais e heurísticos.

- Análise Sequencial: Uma sequência de dados monitorados pelo sistema possui uma distribuição D_0, quando a partir de um dado x, a distribuição muda para a distribuição D_1, então a partir de x os valores tendem a ser diferentes anteriores a ele, tendo "diferentes" como maiores em módulo a um certo limiar.

- Controle Estatístico: Gera uma distribuição Binomial para os erros dos dados, que tendo um número suficiente de observações, se aproxima da distribuição Normal. Essa distribuição para os erros dos dados não deve mudar, se e somente se houver mudança na distribuição dos dados, ocorrendo assim o *Concept Drift*.

- Diferenças entre duas janelas temporais: Gera duas janelas, uma com os dados atuais e outra com um sumário das informações passadas. A detecção do

Concept Drift é feita entre a comparação das duas janelas utilizando métodos estatísticos.

- Heurísticos: Utiliza informações contextuais como informação para calcular diferenças, tais como *timestamp* do dado, características específicas de cada domínio em relação aos dados, etc.

2.3.3 Quanto ao aprendizado

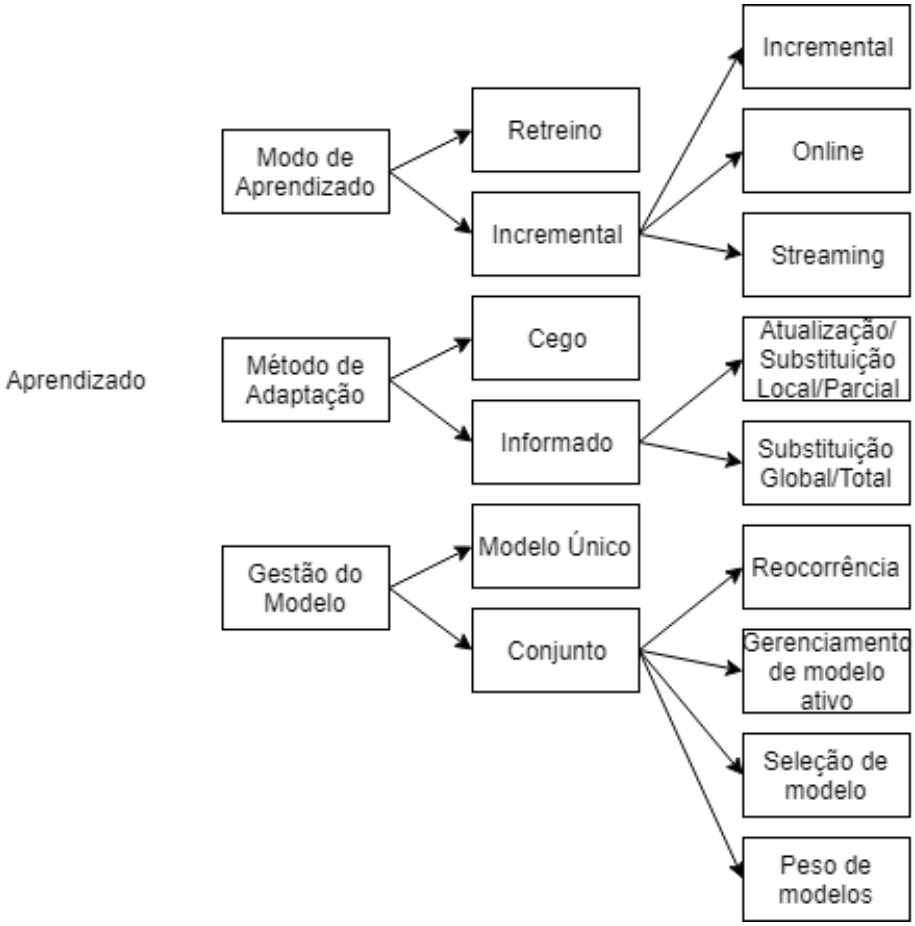

Figura 2.3 – Tipos de tratamento de *Concept Drift* relativos ao aprendizado. Fonte: Adaptado de GAMA *et al.* (2014)

Este módulo se refere à forma como a atualização do modelo preditivo se dá e tem sua representação feita na figura 2.3.

- Modo de Aprendizado: Quando novos dados surgem o modelo preditivo precisa se atualizar. O modelo pode ser retreinado quando o modelo antigo é descartado, ou pode ser atualizado de forma incremental, quando o modelo antigo é aproveitado. Duas formas de modelo incremental são os modelos Online, quando somente os mais recentes são utilizados para treinamento e

Streaming quando o fluxo de dados é de alta velocidade e contínuo, então o modelo não pode ser muito complexo, devido à natureza dos dados.

- Modo de Adaptação do Modelo: Pode ser cego, ou seja, adapta o modelo sem nenhuma detecção de mudança nos dados (geralmente são janelas de tamanho fixo deslizantes). Também pode ser do tipo informado: neste tipo o modelo é adaptado caso algum gatilho seja ativado pelos dados e este modelo então pode ajustar o modelo como um todo ou somente parte dele.

- Modo de Gestão de Modelos: É utilizada uma combinação entre modelos existentes, nesse caso os dados são gerados sendo uma mistura de distribuições, onde há pesos para cada um gerando uma combinação entre as mesmas.

2.4 Redes Neurais Artificiais

O nome "Redes Neurais" é uma alusão ao sistema nervoso central de animais (o cérebro, mais especificamente). Assim como o cérebro, a rede neural artificial possui neurônios interligados em camadas, e é amplamente utilizada em aprendizado de máquina (ANDERSON, 1995). São utilizadas para detectar padrões complexos praticamente indetectáveis por seres humanos, onde, dada uma entrada, é produzida uma saída de acordo com alguma lógica interna.

Semelhanças entre as Redes Neurais do cérebro e as Redes Neurais Artificiais (FAUSETT *et al.*, 1994):

- O processamento da informação ocorre no neurônio.

- Sinais entre neurônios são transmitidos através de impulsos pelas sinapses, que são conexões entre os neurônios.

- Cada sinapse possui um peso associado, amplificando assim o sinal transmitido.

- Cada neurônio aplica uma função não linear à sua entrada para determinar a saída.

Elas possuem vantagens consideráveis em relação à outros métodos:

- São relativamente fáceis de entender comparados à outros métodos que necessitam de conhecimentos específicos, tais como estatísticos, etc.

- Trata-se de um modelo não linear, portanto são bem eficazes aproximando funções não lineares.

Podem ser supervisionadas ou não-supervisionadas. Podem ser, também, com (*recurrent*) ou sem (*feedforward*) realimentação. Em Redes Neurais *feedforward* nenhum neurônio de uma camada, alimenta outro neurônio de uma camada anterior, já nas recorrentes um ou mais neurônios alimentam uma camada anterior e, por conseguinte, a ele mesmo. Umas das mais utilizadas para Processamento de Linguagem Natural (do inglês, *Natural Language Processing*, NLP)) são as *Gated Recurrent Units* (GRU) (CHO *et al.*, 2014a) e as *Long-Short Term Memory* (LSTM) (HOCHREITER & SCHMIDHUBER, 1997a).

O neurônio tem o formato da figura 2.4, e a função aplicada à sua entrada para determinar a saída denomina-se "função de ativação", que pode ser linear, degrau, sigmóide, entre outros.

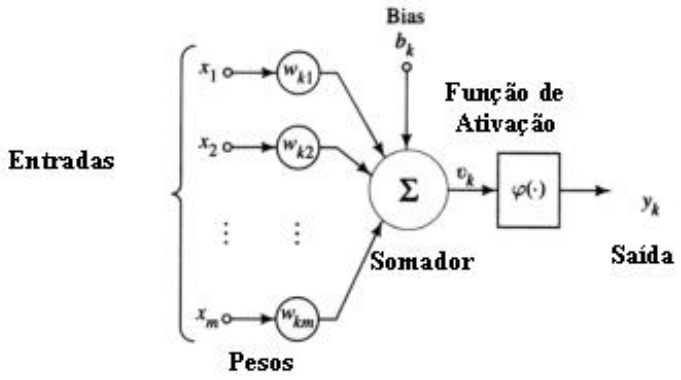

Figura 2.4 – Formato de um neurônio. Fonte: adaptado de HAYKIN (1994)

2.4.1 Redes Neurais *Feedforward*

Redes Neurais que só seguem um sentido sempre. Podem possuir uma ou mais camadas, sendo a mais simples de todas, o *Perceptron* de uma camada, que consiste em uma rede neural onde a saída está conectada diretamente à entrada, e saída é tipicamente 1 ou 0 dependendo da entrada, aplicados os pesos de cada neurônio. O grau de complexidade pode ser acrescido aumentando-se o número de camadas, número de neurônios, ajustados pesos, etc.

2.4.2 Redes Neurais *Recurrent* (*RNN*)

É bastante utilizada em NLP levando em consideração que cada palavra ou frase em um texto não podem ser retiradas de contexto, ou simplesmente desconsiderar todo seu texto ou parágrafos passados, este problema é conhecido como *Long-Term Depedency*. Então as redes neurais recorrentes de certo modo simulam a compreensão humana, já que tem um mecanismo para resgatar conhecimentos passados.

2.4.2.1 Redes Neurais *Long Short-Term Memory*

Introduzida por HOCHREITER & SCHMIDHUBER (1997b) é um tipo especial de *RNN*, que foi especificamente desenhada para resolver o problema de *Long-Term Depedency*. Ela possui um certo esquema de filtragem de reaproveitamento de informações. Ela, em geral, possui o formato da figura 2.6:

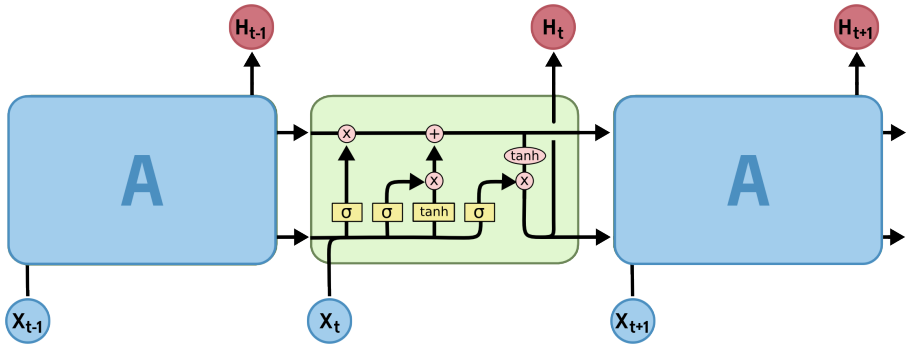

Figura 2.5 – Formato de uma *LSTM*. Fonte: adaptado de http://colah.github.io

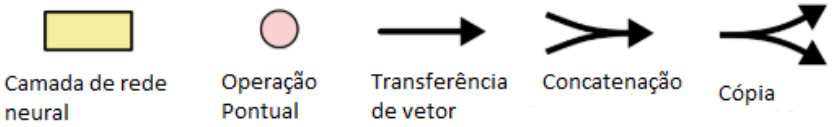

Figura 2.6 – Legenda. Fonte: adaptado de http://colah.github.io

LSTMs possuem 4 camadas de redes neurais, que interagem de uma maneira especial. As redes possuem realimentação, então ela utiliza uma combinação das entradas novas representados pelos $(X_t - 1, X_t, X_t + 1$, e entrada como saída dela mesma, simbolizadas pelas saídas dos A's e geram uma saída que são representadas por $H_t - 1, H_t, H_t + 1$.

2.4.2.2 Redes Neurais *Gated Recurrent Units*

Foi introduzida por CHO *et al.* (2014b) e é uma variante da LSTM, mantém sua principal função que é contornar o problema do *Long-Term Dependency*, porém seu modelo é mais simples em relação à LSTM tradicional, e portanto, mais rápido de treinar, como mostrado na figura 2.7.

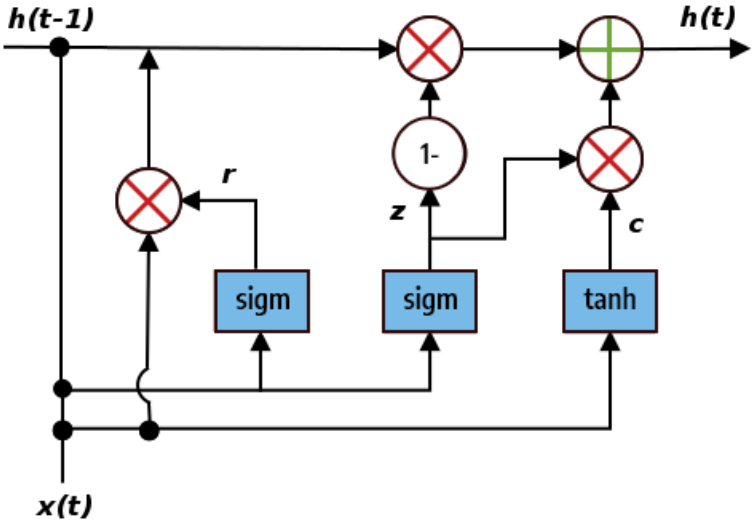

Figura 2.7 – Formato de uma *GRU*. Fonte: adaptado de http://www.safaribooksonline.com

2.4.3 AutoEncoder

AutoEncoders, em geral, tem o formato da figura 2.8:

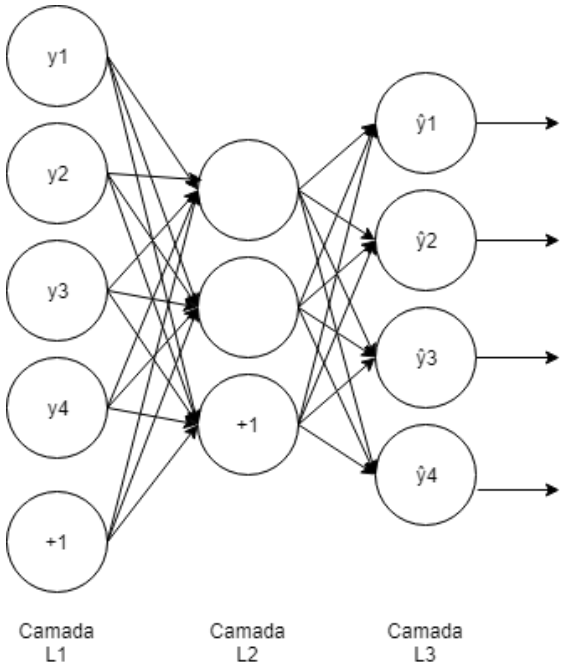

Figura 2.8 – Formato de um AutoEncoder. Fonte: http://ufldl.stanford.edu

Autoencoders são redes neurais que são divididas em duas partes: Codificador e Decodificador. O treinamento dela basicamente é realizado para reproduzir a entrada, então é obrigada a aprender boas representações sobre a entrada, método esse chamado de *backpropagation*. Dada a entrada $y_1, y_2, y_3, \ldots$, onde $y_i \in \Re_n$, o *Autoencoder* tenta aprender a função tal que $hW, b(y) \approx y$.

2.4.3.1 *Autoencoder Neural Event Model* (AutoNEM)

DASIGI & HOVY (2014) desenvolveu o método *Neural Event Model* (NEM) com o objetivo de identificar notícias anômalas baseado em eventos extraídos de notícias. Para tal objetivo, ele utiliza uma técnica de *deep learning* semi-supervisionado (LECUN *et al.*, 2015) considerando manchetes de notícias. Primeiramente é utilizado o SENNA (COLLOBERT *et al.*, 2011) para a anotação de Papéis Semânticos, identificando agentes próximos aos *5W1H*. Em seguida, após o pré-processamento (remoção de *stopwords*, *stemming*) são identificados os termos relevantes, e então é usado o algoritmo não supervisionado *GloVe* (PENNINGTON *et al.*, 2014) para a geração de *word-embeddings* para cada termo. *Word-Embeddings* são representações vetoriais que são úteis para se ter uma representação que possa ser comparada, medida distância, etc, entre objetos que antes não poderiam ser computadas esses valores.

Com os *word-embeddings* gerados, o método então tem duas etapas: uma rede neural recorrente e recursiva é usada para treinar os termos e gerar a melhor composição de argumentos, e em seguida com posse da composição de argumentos é gerada uma próxima camada gerando a representação do evento. Fabricio (PEREIRA, 2018) desenvolveu o método *AutoNEM*, um método baseado no NEM, que tem o mesmo objetivo, e é baseado em 4 passos:

- Detecção e Estruturação do evento: É utilizado o pré-processamento e anotação de Papéis Semânticos.
- Treinamento do AutoNEM: Um modelo é treinado utilizando um *Autoencoder*, a fim de descobrir a melhor representação possível do evento e de seus atributos.
- Geração de *Encoding* dos eventos: Gera um Codificador do evento junto com seus argumentos que foram gerados no passo anterior.
- Relacionamento entre eventos: A representação de cada evento é comparada com o propósito de descobrir similaridades.

Capítulo 3

Metodologia

Este capítulo tem por objetivo introduzir o conceito de Sistemas Autonômicos, o qual a dissertação busca implementar.

3.1 Sistemas Autonômicos

O termo "autonômico" vem da biologia. No corpo humano, determinadas funções são realizadas pelo sistema nervoso central de forma autônoma, como dilatação de pupilas, arrepio de pelos, ajuste de batimentos cardíacos, funções digestivas do estômago e intestinos, etc.

Com o grau de complexidade que os sistemas de computadores alcançaram nos últimos anos e a necessidade dos mesmos estarem, não só sempre funcionando, como estarem sempre funcionando da melhor maneira possível, o termo "Computação Autonômica" foi cunhado em 2001 pela IBM para estes tipos de sistemas.(HORN, 2001; HUEBSCHER & MCCANN, 2008; KEPHART & CHESS, 2003)

O primeiro projeto em que foi utilizado um sistema dito autonômico foi o projeto *Situational Awareness System* (SAS), desenvolvido pela *DARPA* em 1997, onde soldados em local de guerra, precisavam se comunicar com outros soldados além de mandar sua localização para os mesmos, então o método encontrado foi o dispositivo mandar mensagem para o dispositivo mais próximo, funcionando de forma par-a-par, porém o sistema deveria ter uma particularidade: caso o soldado estivesse mais próximo, quer fosse na missão ou superior local dele, a frequência de onda deveria ser menor devido a proximidade física, caso precisasse se comunicar com algum soldado mais distante, como por exemplo a base dele, a frequência deveria ser maior. O dispositivo deveria ser capaz de ajustar essa configuração de forma automática.

Um estudo então foi feito e demarcou que um sistema para ser autonômico tem que possuir determinadas propriedades, que foram definidas por HORN (2001): Auto-Configuração, Auto-Otimização, Auto-Cura e Auto-Proteção. É definido ainda o modo de interação dos sistemas autonômicos com o ambiente externo.

3.1.1 Autoconfiguração

Um sistema autonômico se configura de acordo com objetivos de alto nível para qual foi designado. Em geral, esta propriedade são mais diretrizes, não instruções exatas do que fazer.

3.1.2 Auto-Otimização

Um sistema autonômico deve estar permanentemente buscando otimizar seus recursos e métodos de ação a fim de atingir seu objetivo final. Recursos esses que podem ser dos mais variados tipos, sendo eles relevantes para o sistema e métodos de ação que podem ser melhorados de acordo com alguma métrica adequada para a função desempenhada. Esta tarefa deve ser realizada de forma espontânea e ininterruptamente pelo sistema.

3.1.3 Autocura

Um sistema autonômico deve ser capaz de detectar problemas em si mesmo. Esses problemas podem ser dos mais variados graus e aspectos, como por exemplo alguma parte de seu sistema parou de funcionar por causa de uma peça em seu hardware avariada ou uma função em seu software utilizada de forma errada. O sistema tem que ser capaz de diagnosticá-los e remediá-los de alguma maneira, e não só isso, consertando a si mesmo da melhor maneira possível.

3.1.4 Autoproteção

Um sistema autonômico deve ser capaz de se proteger de ataques ou interferências externas. Os ataques podem ser de diversos tipos e estratégias, *malwares*, ataques físicos (quando couber), etc.

3.1.5 *Loop* Autonômico

Além das 4 propriedades já citadas, também foi definido o modo de interação dos sistemas autonômicos, denominado como *MAPE-K*.

O sistema *Monitora* propriedades do ambiente através de seus *Sensores* que são relevantes para a funcionalidade do sistema ou alguma de suas propriedades, *Analisando* as propriedades, quando são encontrados dados que afetem os objetivos do sistema é *Planejada* uma ação a fim de remediar a situação da melhor maneira possível, *Executando* ações com tais objetivos com *Atuadores* disponíveis. Todas as ações tomadas pelo sistema são baseados em *Conhecimentos* que foram concebidos com

ele ou adquiridos através do tempo em que esteve em execução. Conforme pode ser visto na figura 3.1

COMPUTING *et al.* (2006) introduz o *loop MAPE-K*, que é constituído de cinco passos: *Monitor*, *Analyze*, *Plan*, *Execute* e *Knowledge*.

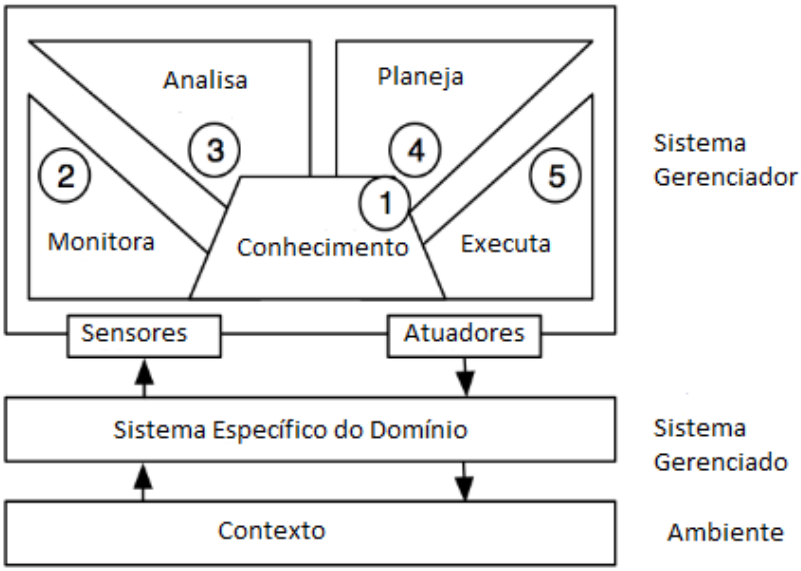

Figura 3.1 – Esquema do *Loop MAPE-K*. Fonte: adaptado de COMPUTING *et al.* (2006)

Diversos estudos já foram feitos com base no *Loop MAPE-K* (ARCAINI *et al.*, 2015; IGLESIA & WEYNS, 2015; RUTTEN *et al.*, 2017), assim como a Computação Autonômica cresceu consideravelmente e é o alicerce de diversos tipos de sistemas, inclusive com as mais modernas tecnologias e conceitos em computação, como *Smart Cities*, *Big Data* e *Cloud Computing.*(AL-SHARIF *et al.*, 2017; HE *et al.*, 2017; LALANDA *et al.*, 2013; POP *et al.*, 2017)

O sistema Autonômico pode ainda possuir vários níveis segundo HORN (2001):

- Básico: Funcionários altamente treinados utilizam ferramentas de monitoramento e fazem alterações manualmente em elementos do sistema.
- Gerenciado: O sistema coleta dados de seu funcionamento de alguma forma inteligente diminuindo o trabalho dos funcionários responsáveis pelo sistema.
- Preditivo: Informações do sistema são coletadas de um jeito mais inteligente que o nível Gerenciado e, além disso, o sistema reconhece padrões no comportamento sugerindo ações a serem tomadas pelos funcionários a fim de combater problemas conhecidos ou melhorar o comportamento do sistema.
- Adaptativo: O sistema utiliza as ferramentas, que são gerenciadas por humanos no nível Preditivo, e atua a fim de minimizar a interação humana ajustando o desempenho próprio automaticamente.

- Totalmente autonômico: O sistema e seus componentes atuam de acordo com regras e políticas a que foi projetado e são controlados por elas, não precisando de nenhuma interação com humanos.

HUEBSCHER & MCCANN (2008) argumenta que estes tipos são muito limitados, restringindo para apenas poucos tipos de sistema, além da descrição e diferenciação entre níveis serem bem vagas, então propõe diferentes elementos da autonomicidade:

- Suporte: Sistemas que focam especificamente em um aspecto ou componente da arquitetura ajudando a melhorar a performance de algum modo.
- Núcleo: Sistemas que focam no objetivo em si, porém não trabalham com objetivos em alto nível.
- Autônomo: Sistemas que possuem um certo grau de inteligência, capazes de se adaptar ao ambiente a fim de enfrentar dificuldades que porventura forem encontradas, entretanto não é capaz de medir a própria performance e adaptar-se a ele para que melhore um objetivo de alto nível.
- Autonômico: Sistemas que descrevem uma arquitetura completa, sendo os objetivos desta arquitetura definidos de forma auto-nível para humanos, e o sistema trabalha independentemente para alcançar tais objetivos.

3.1.6 Sistemas Autonômicos e Notícias

Alguns trabalhos já foram produzidos no sentido de trabalhar com notícias em sistemas autonômicos:

AMOUI *et al.* (2008) discute um método baseado em Aprendizado por Reforço para definir estratégias relativas a qual melhor ação a ser tomada pelo sistema autonômico dentre uma quantidade finita de possibilidades, e testa em um sistema de notícias *web*, adaptando a quantidade e qualidade do conteúdo à propriedades específicas do usuário.

PINHEIRO *et al.* (2009a) propõe um *framework* chamado *Autonomic Data Killing* a fim de filtrar notícias irrelevantes e notícias com conteúdos proibidos em um *feed* de notícias de forma autonômica, posteriormente aperfeiçoa o método criando regras do sistema autonômico baseado em redes Petri (PINHEIRO *et al.*, 2010). Propõe também o *Autonomic collaborative RSS* sistema bastante semelhante e com o mesmo propósito baseado no usuário.(PINHEIRO *et al.*, 2009b)

MIRANDA *et al.* (2017) desenvolve um sistema que automaticamente acha correspondências entre *tweets* e artigos de notícias.

MORAES (2016) não utilizou notícias, mas *tweets*, e desenvolveu um sistema onde consultas sobre eventos eram modificadas à medida que o evento ia se modificando, atuando assim, autonomicamente para modificar as *queries* de busca.

Capítulo 4

Proposta e Implementação

Este capítulo tem por objetivo definir a proposta para a resolução do problema apontado bem como a implementação efetivamente usada.

4.1 Definição do problema

Para chegar à definição do problema, primeiro é preciso definir alguns conceitos básicos. Na literatura de TDT existem algumas definições de "evento" e "tópico". A mais corriqueira é a de que um evento trata-se de "algo que acontece em algum lugar em uma determinada data" (ALLAN *et al.*, 1998), porém esta definição não é muito exata já que levanta alguns questionamentos tais como: essa data é uma data específica ou pode ser um intervalo de tempo? Uma batida de trem acontece em determinado dia, hora, minuto e segundo, já uma guerra por exemplo, pode durar anos, mas ainda sim se enquadraria nesta definição, assim como desastres naturais que tendem a demorar, por exemplo erupção vulcânica, furacões, enchentes, tufões, entre outros.

Para tentar corrigir a definição anterior e cobrir possíveis falhas CIERI *et al.* (2002) definiram evento como "Evento é algo que acontece em uma determinada data e lugar junto com todas pré-condições necessárias e suas inevitáveis consequências", porém, novamente, surgem questionamentos, tais como o que seriam "pré-condições necessárias" ou "consequências inevitáveis"? Quando ocorre a queda de um avião, espera-se que existam mortos, mas isso é consequência inevitável? E se houverem sobreviventes, seria outro evento então? Quais seriam as "pré-condições necessárias" e "consequências inevitáveis" para um evento de queda da bolsa de valores?

Um tópico é definido como "um evento ou uma atividade, junto com todos os eventos e atividades relacionados". (CIERI *et al.*, 2002) Esta definição tenta fazer uma conexão entre evento e tópico. Já a definição de tópico por MAKKONEN (2003) diz "Um tópico é uma série de eventos, uma narrativa que evolui e pode se dividir em vários tópicos". Nota-se que tenta capturar a natureza evolutiva inerente à

eventos. Também pode-se notar que evento e tópico são muito similares e altamente relacionados.

A figura 4.1 tenta exemplificar um conceito mais completo de evento, e o utilizado pelo atual trabalho.

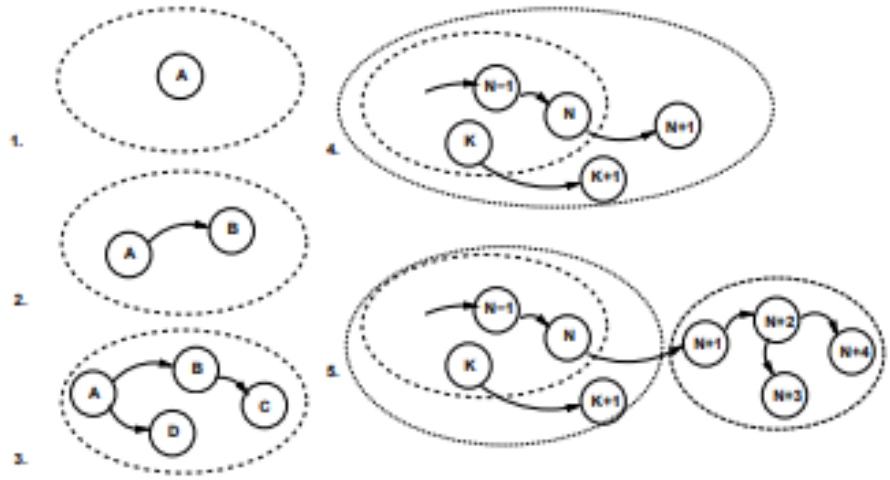

Figura 4.1 – Exemplo de Evolução de eventos. Fonte: MAKKONEN (2003)

A princípio, tem-se somente um documento, que seria uma primeira história A de determinado evento, que está simbolizado pela linha pontilhada no estágio 1. Em seguida, um documento B é detectado também sobre o evento, e já que está fortemente conectado e relacionado ao documento A, existe uma relação entre ambos. Essa relação é feita baseada na similaridade de termos entre ambos os documentos.

Na sequência, os documentos C e D são verificados fortemente relacionados com os documentos B e A, respectivamente, correspondendo ao estágio 3, estando ainda enquadrados ao mesmo evento.

Então, no estágio 4, os documentos K+1 e N+1, possuem vocabulários tão diferentes dos documentos originais, que estariam fora do conceito de evento, entretanto relacionados aos documentos do evento (K e N, respectivamente), sendo então considerados fora do evento.

E finalmente, no estágio 5, os documentos N+2, é encontrado relacionado ao N+1 e os documentos N+3 e N+4 relacionados ao N+2, formando assim um novo evento. O documento K+1, não foi encontrada uma ramificação, então é excluído, não formando nem um evento novo, nem relacionado ao evento já encontrado.

O trabalho tem por objetivo verificar diferenças no evento. E para isso é proposta a arquitetura da figura 4.2 caixa preta para o usuário:

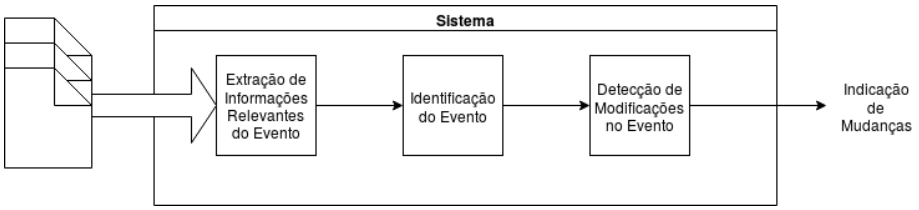

Figura 4.2 – Arquitetura do Sistema

As notícias, em *batches* por intervalo de tempo, passam por um processo onde as informações mais importantes delas são extraídas, em sequência, cada *batch* de notícias é caracterizado e por fim é realizada a verificação das modificações, sendo finalmente detectada a mudança (ou não) do evento. Ao final do processo é mostrado o indicativo de modificação ou não como resultado para o usuário.

4.1.1 Entrada de Dados do Sistema

A entrada do sistema poderia ser um *stream* de notícias, onde cada notícia passa pelo sistema separadamente, entretanto a fim de diminuir a detecção errônea de ruídos, idealmente é a utilização de *batches*. A entrada do sistema são notícias digitais que podem ser coletadas manualmente ou com alguma ferramenta. O conjunto de notícias é contido em uma janela temporal, de tamanho arbitrário, que é chamado de *batch*. Cada *batch* possui um valor atômico em relação às mudanças, ou seja, é considerado que não há modificações dentro de uma mesma janela temporal.

4.1.2 Módulo de Extração de Informações do Evento

Neste módulo são coletadas informações que sejam relevantes do Evento. Pode-se ser utilizada qualquer método, é possível que seja coletando manualmente as informações (como utilizando somente o título da notícia, etc), aplicando-se técnicas de sumarização, entre outros.

4.1.3 Identificação do Evento

Neste módulo os eventos são caracterizados e representados de maneira que possam ser comparados tanto qualitativamente quanto quantitativamente com outros, representação esta que pode ser numérica (cada notícia ser representada por um número, conseguido através de algum cálculo), representação Vetorial, representação Semântica, entre outros.

4.1.4 Discriminação de Modificações do Evento

Neste módulo são computadas as diferenças para a detecção explícita da modificação no evento. As modificações podem ser analisadas manualmente, podem ser calculadas as distâncias de representações, conforme seja a representação de cada evento escolhida no Módulo de Identificação do Evento.

4.2 Implementação

A implementação do sistema se deu dividindo-o em 4 diferentes componentes que cooperam entre si a fim de alcançar o objetivo proposto, podendo ser visto na figura 4.3.

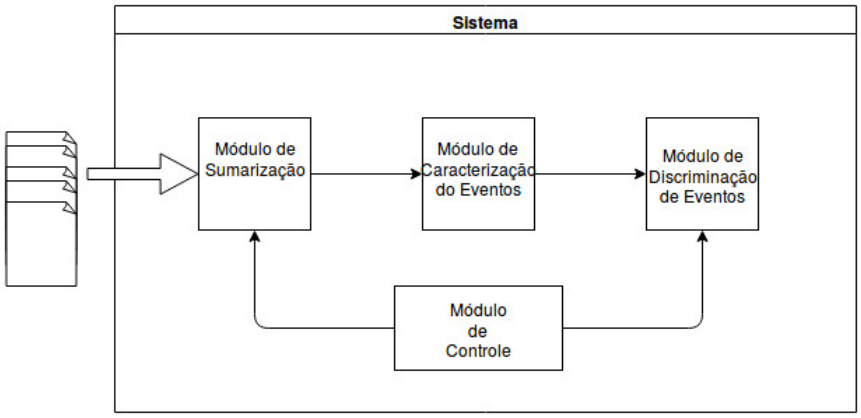

Figura 4.3 – Sistema proposto completo

Uma importante colocação é que o sistema se propõe a ser altamente modularizado, ou seja, caso surjam na literatura métodos com melhores resultados que cada módulo individualmente, basta que a entrada e a saída do módulo sejam respeitadas e o sistema poderá assim ser aperfeiçoado sem maiores esforços.

Outra importante observação é a de que o sistema deve ser inteiramente não-supervisionado, isto é, em nenhuma das etapas do processo deve ter interferência de um ser humano para dar qualquer tipo de *feedback* em relação à corretude de suas avaliações finais.

4.2.1 Módulo de Sumarização

O meio de se extrair as informações importantes no Módulo de Extração da Informação se deu por meio de um método de sumarização.

A abordagem do presente trabalho se propõe a usar uma técnica de sumarização extrativa, que leva em consideração um único documento por vez, genérico, não-supervisionado e apenas no idioma inglês. Quanto ao tipo utilizado, idealmente

seria o estatístico, já que os dados considerados por este tipo são de muito valor para que o propósito deste passo do trabalho seja alcançado. Poderia ser também os baseados em Aprendizado de Máquina, visto que entender e aprender com o fluxo de documentos possivelmente aumentaria a qualidade dos sumários gerados.

Por facilidade e disponibilidade da *API* do *gensim* foi utilizada a variação do *TextRank* disponibilizada em BARRIOS *et al.* (2016) que obteve excelentes resultados com notícias batendo o estado da arte da data da publicação.

O módulo recebe um *batch* de notícias, e aplica um algoritmo de sumarização nele. Como o método da proposta é *single-document*, para cada notícia a técnica é aplicada, obtendo assim uma lista com cada notícia sumarizada.

4.2.2 Módulo de Caracterização de Eventos

O módulo é responsável por gerar a Anotação de Papéis Semânticos, e para cada tupla de Papéis Semânticos encontrados é utilizada a abordagem de geração de representações vetoriais através de um *AutoEncoder*, que são os primeiros passos do *AutoNEM*. (PEREIRA, 2018)

O módulo recebe a saída do módulo de sumarização e, para cada sentença gerada nessa saída é aplicada uma técnica de Anotação de Papéis Semânticos, a fim de extrair e caracterizar os atributos de cada sentença, que são próximos aos desejados 5W1H. No trabalho foi utilizado SENNA (COLLOBERT, 2011; COLLOBERT *et al.*, 2011). Em seguida, o componente de codificação do *Autoencoder* pré-treinado gera um *embedding*, uma representação vetorial, a fim de representar de forma uniforme a semântica do evento para cada tupla. Em uma camada intermediária do Codificador, é gerado uma representação vetorial para cada atributo. Cada atributo com sua representação vetorial é armazenado e utilizado como saída do módulo.

São 5 atributos possíveis encontrados pelo SENNA no AutoNEM: A0, A1, V, AM-LOC, AM-TMP. A0 e A1 são quem praticou a ação e quem a sofreu, podendo ser aproximado a *QUEM* participou do evento. V é a ação em si, pode ser aproximado a *O QUÊ* aconteceu no evento. AM-TMP possui o valor temporal, equiparado a *QUANDO* o evento aconteceu e por fim, AM-LOC são termos que denotam localizações de onde o evento aconteceu, responsável pelo atributo *ONDE* do 5W1H.

4.2.3 Módulo de Discriminação de Eventos

Tendo como entrada a saída do Módulo de Caracterização de Eventos, é responsável por projetar a representação vetorial em um plano e calcular seu centróide para cada atributo do evento. Com o centróide de cada atributo calculado, ele obtém a representação do atributo mais próximo do centróide dada a lista de representações vetoriais da saída do Módulo de caracterização. Do mesmo modo, é calculado o

desvio padrão das distâncias de todos os pontos do *batch* para o centróide, que será utilizado pelo módulo de controle como limiar máximo para a detecção de mudança no evento entre *batches*.

4.2.4 Módulo de Controle

O módulo de controle é responsável por duas tarefas:

4.2.5 Verificação da Sumarização

Verificar que a sumarização está ocorrendo com os melhores efeitos possíveis. O método utilizado pelo Módulo de Sumarização é o *TextRank*, em que um dos parâmetros é a taxa de compressão, que pode ter um valor entre 0 e 1. Quanto maior for a taxa, mais texto o sumário terá, quanto menor, maior a chance do sumário conter somente a sentença em que estão contidos os atributos do evento, descartando portanto sentenças com informações desnecessárias, melhorando assim a etapa seguinte à sumarização (caracterização do evento). Ao passo que não basta apenas botar o menor valor possível, já que as notícias têm tamanho diferente, o percentual para conter o menor número de sentenças possível (idealmente uma só) é variável, logo, o valor da taxa de compressão deve ser adaptado para cada notícia. O código em 4.2.5 representa o algoritmo utilizado.

```
Para cada noticia n em batch:
    taxaSumarizacao = 0.2;
    sumario = aplicarTextRank(n, taxaSumarizacao);

    Enquanto numeroSentencas(sumario) > 1:
        sumario = aplicarTextRank(n, taxaSumarizacao - 0.01);
    Se numeroSentencas(sumario) = 0:
        Enquanto numeroSentencas(sumario) == 0:
            sumario = aplicarTextRank(n, taxaSumarizacao + 0.01);
```

Propriedades autonômicas para o módulo de Sumarização:

- Auto-Configuração: É definido o valor da taxa de sumarização.
- Auto-Otimização: Caso o módulo de controle identifique que o resultado da sumarização está com sentenças demais (idealmente é uma) ele atua a fim de reduzir o número de sentenças obtidos diminuindo o percentual, quando identifica que o número de sentenças é insuficiente (zero), e age para aumentá-las, aumentando a taxa de sumarização.

- Auto-Cura: Como nem sempre é possível que o número de sentenças seja o ideal, o sistema pode ficar alternando entre tentativas onde são obtidas 0 e n > 1 sentenças eternamente. O sistema identifica este problema e interrompe este mal funcionamento, utilizando n como taxa de sumarização.

- Auto-Proteção: Não foi identificada a necessidade da implementação desta funcionalidade.

4.2.6 Detecção de *Concept Drift*

Detectar a mudança, e, caso tenha ocorrido, remediar o *Concept Drift* detectado.

Primeiro objetivo desta tarefa é detectar *outliers*. Dado que eu tenho k medóides $M_0, M_1, \ldots, M_i, \ldots, M_k$, e foi identificado que a distância do medóide M_i para o medóide referência, que a princípio é o M_0 com desvio padrão σ, ultrapassou o limite σ, é guardada essa informação, no *batch* que possui o medóide M_{i+1}, é feita a mesma comparação com M_0, caso exceda o limite novamente, é detectada uma mudança de fato, e M_i não era um *outlier*, então a comparação é feita entre M_i e M_{i+1}, sendo este processo repetido a cada *batch* recebido pelo sistema. Caso o limite entre M_0 e $M_i + 1$ não seja maior que σ, M_i é considerado *outlier* e o medóide de referência permanece inalterado. Esta foi a forma encontrada a se lidar com os possíveis *Concepts Drifts*. O algoritmo pode ser visto em 4.2.6.

```
modificacaoEncontrada = falso
detecacaoMudanca = falso
Para cada medoide Mi em medoides:
    Se medoideReferencia == vazio:
        medoideReferencia = Mi

    Se distanciaCossenos(MedoideReferencia, Mi) > valorLimite:
        Se modificacaoEncontrada == verdadeiro:
            medoidereferencia = Mi
            detecacaoMudanca = verdadeiro
        Senao:
            modificacaoEncontrada = verdadeiro
```

Quanto à classificação proposta por GAMA *et al.* (2014), explicada na seção 2.3 para detecção e tratamento de *Concept Drift*, o sistema se enquadra como:

- Quanto à memória: Dados considerados em uma janela de tempo deslizante fixa e um mecanismo de esquecimento abrupto.

- Quanto à detecção de mudanças: Análise Sequencial.

- Quanto ao aprendizado: o modelo é retreinado de forma cega.

Um exemplo de um mesmo atributo do evento durante toda a duração do mesmo com a dimensionalidade reduzida para duas dimensões pelo *T-SNE*, acrescido de seu atributo temporal, de forma a visualizar a evolução do evento pode ser visto nas figuras 4.4 e 4.5. (MAATEN & HINTON, 2008)

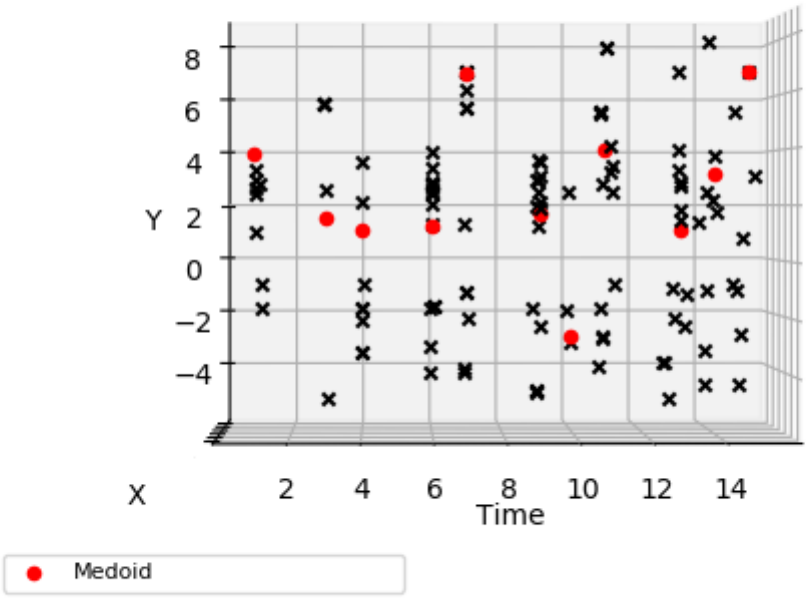

Figura 4.4 – Exemplo de um atributo de notícias e seus medóides nas dimensões tempo e Y

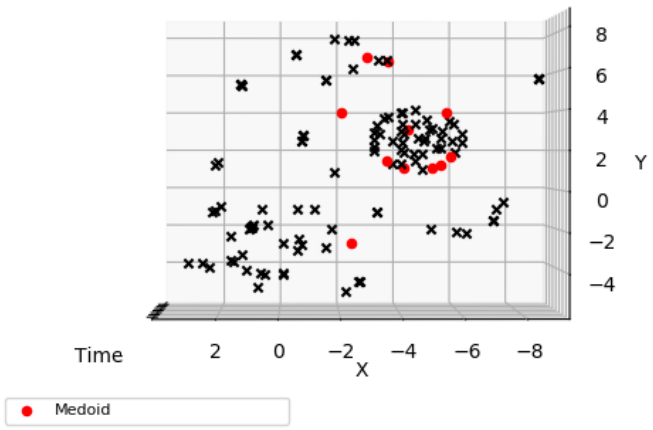

Figura 4.5 – Exemplo de um atributo de notícias e seus medóides nas dimensões X e Y

Propriedades autonômicas para o módulo de controle:

- Auto-Configuração: É definido o ponto de referência. O medóide que deve ser tomado como base para a detecção de mudança é escolhido pelo módulo.
- Auto-Otimização: Não foi identificada a necessidade da implementação desta funcionalidade.
- Auto-Cura: Quando o *Concept Drift* é identificado, o vetor de referência é alterado, já que, teoricamente, dali em diante, caso não houvesse mudança no atributo do evento, ele sempre identificaria mudança erroneamente.
- Auto-Proteção: Não foi identificada a necessidade da implementação dessa funcionalidade.

Capítulo 5

Experimentos e Resultados

Capítulo destinado à explicação de como os experimentos foram montados, bem os resultados e a análise dos mesmos.

5.1 Experimentos

Esta seção tem por objetivo definir como foram realizadas as abordagens utilizadas nos experimentos. Importante mencionar que tanto os *datasets* quanto os métodos foram feitos para a simulação de um sistema *online*, ou seja, o conjunto de documentos utilizados não era fechado, cada *batch* tinha conhecimento apenas dos passados e do presente.

5.1.1 *Datasets*

Para a montagem dos *datasets* foi utilizada a API de consultas do *EventRegistry*[1]. O *EventRegistry* é um site que recupera notícias de todo mundo, inclusive em diferentes línguas, que tratam sobre o mesmo evento e as armazena juntas, sendo o processo quase todo automatizado exceto pela última etapa, onde um humano que diz se realmente é o mesmo evento ou não, então, para o contexto deste trabalho, é tomado como verdade que cada notícia é de fato pertencente ao mesmo evento, e não se trata de um evento novo. (CIERI *et al.*, 2002)

Nos experimentos foram utilizados 5 *datasets*. Os *datasets* foram montados coletando-se manualmente notícias diariamente de eventos ocorridos, por um período de aproximadamente 20 dias corridos, no período de setembro-outubro de 2017. Foram coletadas somente notícias no idioma inglês. O número de notícias de cada *dataset* é variável, e, como nem todos os temas tiveram notícias veiculadas todos os dias, há *datasets* com dias sem uma notícia sequer, ou com pouquíssimas, como 5 ou 10. Há também dias com inúmeras notícias, até mesmo com 1000 notícias.

[1] `http://eventregistry.org/`

Cada *dataset* contém um tema diferente, há um referente ao furacão Maria, furacão este que devastou Porto Rico e diversos outros países da América Central, outro sobre uma série de queimadas e incêndios ocorridos na Península Ibérica, um sobre a crise da separação da Catalunha, onde a província da Catalunha queria se emancipar da Espanha em definitivo e também um sobre a crise diplomática entre EUA e Coréia do Norte, onde o presidente dos EUA *Donald Trump* e o da Coréia do Norte, *Kim Jong Un*, trocaram acusações públicas. Além desses, foi criado um compacto do Furacão Maria, que mais adiante será explicado o seu propósito.

Uma observação importante é que a janela de tempo foi fixada em 1 dia, ou seja, a unidade atômica é o dia, sendo o evento considerado imutável dentro de cada janela de tempo. Sabemos que quanto maior a urgência do evento, mais rápido ele pode se modificar, então essa janela foi arbitrariamente fixada em 1 dia, mas poderia ser 1 hora, 1 semana, 30 minutos, etc, tanto o *Baseline* como a proposta necessitam apenas de intervalos de tempo de tamanhos iguais.

5.1.2 *Baseline*

Para a elaboração do *Baseline* foi utilizada a técnica de *Bag of Words*, onde cada janela de tempo é vetorizada de acordo com as palavras da mesma, e comparada com a seguinte, e assim sucessivamente. Caso a distância (utilizando-se a distância dos cossenos) de uma janela para a outra for maior que um determinado limite calculado previamente, então o indicativo de que houve mudanças é lançado.

Foram utilizados 3 limiares que indicavam a mudança: o valor que maximizava o *f1-score* no *dataset* Furacão Maria compacto, valor este também maximizador da Revocação (0.05), o valor onde Precisão, Revocação e *f1-score* coincidiam (0.11), e por fim, o valor onde a Precisão era máxima (0.13) como mostra a figura 5.1.

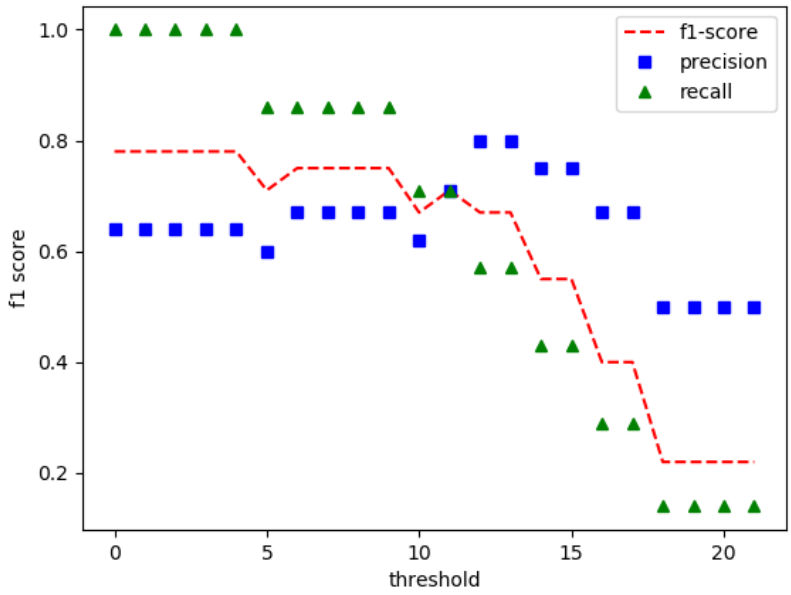

Figura 5.1 – Métricas do método *bag of words* no *dataset* de determinação de parâmetros

5.1.3 Abordagem Proposta

Foram utilizadas 4 abordagens relativas ao módulo de extração da informação para a proposta: Utilizando somente o título, levando em consideração o primeiro parágrafo, usando a notícia completa e notícia completa sumarizada.

A abordagem utilizando somente o título justifica-se no fato de que teoricamente a informação mais relevante ou impactante da notícia está no título, tendo algum grau de informações contextuais.

Os jornalistas quando escrevem as notícias, procuram conseguir a atenção dos leitores no corpo da notícia o mais breve possível, então grande parte da informação relevante sobre a notícia que o artigo traz encontra-se no 1º parágrafo, também chamado de *lead paragraph* (BELL, 1991; MARGOT & PEHA, 2006). Este parágrafo traz mais informações que o título, que teoricamente já identifica as partes mais importantes da notícia.

O corpo da notícia, para averiguar se, levando em consideração todo o conteúdo do documento, é eficiente para o objetivo do trabalho. Por último, é testada a técnica de sumarização descrita em 4.2.1. Todas as técnicas, que não o *Bag of Words* foram testadas considerando-se o identificador de mudança, a variação de apenas 1 atributo ou 2 atributos do evento.

5.2 Resultados e Discussão

	Compact Hurricane Maria			Hurricane Maria			EUA North Korea			Fire Portugal			Catalonia			Média f1-score
	p	r	f	p	r	f	p	r	f	p	r	f	p	r	f	m
Bag of words t=0,05	0,64	**1,00**	0,78	0,50	0,80	0,61	0,57	**1,00**	0,72	**1,00**	**1,00**	**1,00**	0,75	0,42	0,54	0,73
Bag of words t=0,11	0,71	0,71	0,71	0,55	0,50	0,52	0,57	**1,00**	0,72	**1,00**	0,50	0,66	**1,00**	0,28	0,44	0,61
Bag of words t=0,13	**0,80**	0,57	0,66	**0,66**	0,40	0,50	0,42	0,75	0,54	**1,00**	0,50	0,66	**1,00**	0,28	0,44	0,56
Técnica titulo 1	0,60	0,85	0,70	0,47	0,90	0,62	0,57	**1,00**	0,72	**1,00**	**1,00**	**1,00**	0,63	**1,00**	**0,77**	0,76
Técnica titulo 2	0,66	0,85	**0,85**	0,57	0,80	0,66	0,50	0,75	0,60	**1,00**	0,75	0,85	0,57	0,57	0,57	0,70
Técnica primeiro parágrafo 1	0,64	**1,00**	0,78	0,50	**1,00**	0,66	0,50	**1,00**	0,66	**1,00**	**1,00**	**1,00**	0,60	0,42	0,50	0,72
Técnica primeiro parágrafo 2	0,64	**1,00**	0,70	0,60	0,90	**0,72**	0,42	0,75	0,54	**1,00**	0,75	0,85	**1,00**	0,14	0,25	0,61
Técnica corpo 1	0,70	0,85	0,76	0,50	0,50	0,50	0,75	0,75	**0,75**	**1,00**	0,75	0,85	0,50	0,14	0,22	0,61
Técnica corpo 2	0,66	0,28	0,39	0,00	0,00	0,00	**1,00**	0,25	0,40	**1,00**	0,25	0,40	0,00	0,00	0,00	0,24
Técnica sumarização 1	0,63	**1,00**	0,77	0,50	**1,00**	0,66	0,50	**1,00**	0,66	**1,00**	**1,00**	**1,00**	0,83	0,71	0,76	0,77
Técnica sumarização 2	0,70	0,85	0,77	0,53	0,70	0,61	0,42	0,75	0,54	**1,00**	**1,00**	**1,00**	0,50	0,14	0,22	0,62
Média	0,48	0,68	0,55	0,56	0,81	0,62	1,00	0,77	0,84	0,67	0,81	0,71	0,67	0,37	0,42	0,63

Tabela 5.1 – Resultados das técnicas em todos os *datasets*

Primeiro comentário a ser feito é sobre as particularidades de cada *dataset*. O *dataset Fire Portugal* tratou-se de um *dataset* extremamente curto, com apenas 5 *batches* e todos eles possuíam modificações de um *batch* para o seguinte. Isso explica todos os métodos possuírem precisão 1,0, qualquer *batch* que fosse apontada mudança no evento, estaria correto. O *Compact Hurricane Maria*, como já argumentado, foi utilizado para obtenção de valores a serem utilizados como *baseline*.

Analisando os resultados, a proposta utilizando o título obteve bons resultados, porém a expectativa era a de que ele fosse o melhor em todos, devido relevância do título diante ao conteúdo da notícia. Pelo mesmo motivo, havia a expectativa dessa abordagem apresentar melhores resultados aos da sumarização, já que, como descrito na seção 4.2.4, o objetivo da sumarização era deixar com o menor número de sentenças possíveis, obtendo assim, somente a frase que resume a notícia. Isso não aconteceu, provavelmente porque em geral os títulos das notícias tenham tamanhos limitados, apresentando assim informações contextuais insuficientes para uma máquina reconhecer. Observando-se o *f1-score*, não ficou atrás do *baseline* em nenhum *dataset*.

A proposta que utiliza o *lead paragraph*, a fim de contornar os possíveis problemas dos títulos mencionados no parágrafo anterior, acrescenta frases para tornar mais completa a informação, podendo fazer assim com que acrescente informação demais, fazendo com que piore seu desempenho. Nos experimentos ficou a frente do *baseline*, levando-se em consideração o *f1-score* somente no *dataset Hurricane Maria*.

A abordagem utilizando o corpo completo das notícias, que esperava-se ser o pior, confirmou essa suspeita, chegando a ter *datasets* em que ele obteve 0,0 de Precisão e Revocação, além disso, excetuando-se pelo corpo 1 no *dataset EUA North Korea* em todos o melhor corpo (considerando-se 1 e 2) perdeu para o melhor dos *baselines*. Isso se deve ao fato de que as notícias acrescentam muito conteúdo que não o conteúdo central da notícia, impactando dramaticamente o desempenho da detecção dos atributos do evento.

Já a técnica utilizando a sumarização, podemos observar que em todos os *datasets* a proposta utilizando a sumarização (1 ou 2) esteve entre os 3 melhores valores tanto de *f1-score*, quanto de Recall e Precisão, este último excetuando-se o *dataset EUA North Korea*. Quanto aos *baselines*, em todos os *datasets* o melhor da sumarização ficou melhor que o melhor dos *baselines*, exceto pelo *dataset EUA North Korea*, e o de treinamento do *baseline* por 0,01.

Apenas 2 abordagens ficaram acima ou empatado com a média em todos os *f1-scores*: *Bag of Words* 0,05, e Técnica Sumarização 1.

Técnica Sumarização 1 também se mostrou superior na média dos *f1-scores* de todos os *datasets*(0,77), seguido de perto pela Técnica título 1 (0,76) e em 3º lugar *Bag of Words* (0,73).

Há evidências, portanto, de que não só a técnica para a extração da informação da notícia é importante como a quantidade de atributos do evento a ser modificado também é de suma importância. Há evidências também de que eventos mais relacionados a desastres sejam melhores detectados pelo sistema, visto que seu melhor desempenho foi nos eventos do Furacão Maria e do incêndio na Península Ibérica, possivelmente porque toda a informação está explicitamente na notícia, enquanto que eventos como a crise diplomática entre EUA e Coréia do Norte, depende de outros eventos e acontecimentos externos que não são explicitados, dificultando assim a verificação. No apêndice encontram-se alguns exemplos de notícias dos *datasets*.

Capítulo 6

Conclusão

Este trabalho propunha-se a apresentar um sistema autonômico capaz de detectar mudanças em eventos a partir de notícias.

Para isto, utilizando conhecimentos de Detecção e Rastreio de Tópicos que traz o conceito de *Evento*, sistemas autonômicos, que são sistemas capazes de trabalhar sozinhos, o sistema foi construído.

Para facilitar a extensibilidade, o sistema foi dividido em 4 módulos: módulo de extração da informação, de caracterização do evento, discriminação de eventos e por fim, o de controle. Módulos esses que trabalham juntos e cooperam entre si. O sistema também foi capaz de identificar e tratar certos tipos de *Concept Drift*.

Para a avaliação do sistema, foram montados 5 *datasets* com notícias de eventos reais, coletados manualmente do site *Event Registry*, e, simulando um sistema *online*, foi testado obtendo bons resultados, onde a técnica completa (incluindo sumarização) ficou sempre entre os três melhores resultados.

Obteve também a melhor média de *f1-scores* geral, seguido de perto pela média da técnica utilizando o título das notícias, e em terceiro lugar o *Bag of Words* de máximo *f1-score*.

6.1 Constribuições

- Um sistema autonômico com o mínimo de necessidade de intervenção humana possível para a detecção de mudanças em eventos acontecidos no mundo real e noticiados por mídias textuais digitais.
- Um sistema capaz de detectar *outliers* e determinar, além de tratar, um tipo de *Concept Drift*.
- Construção de 5 *datasets* utilizados para o experimento, de diferentes tamanhos e diferentes temas.

6.2 Trabalhos Futuros

Como trabalhos futuros podemos apontar:

- Utilizar diferentes técnicas de sumarização, como sumarização multi-documentos, ao invés da sumarização documento por documento, e alteração do método entre cada notícia, caso o sistema de controle verifique melhores resultados com esse novo método.
- Alterar a forma de auto-correção da taxa de sumarização a fim de reduzir o número de iterações na busca do melhor valor da taxa, entrando assim essa tarefa para a do tipo "otimização" dentre as tarefas do sistema autonômico. Poderia ser feita uma busca binária por exemplo, pelo melhor valor.
- Aplicação de métricas descritas em 2.2 para melhor avaliação da sumarização aplicada.
- Expansão da técnica para outras línguas, visto que o Módulo de Discriminação de Eventos é treinado para a língua inglesa.
- Especificação da abordagem para diferentes áreas específicas, tais como notícias que tratam de política, esportes, etc.
- Integrar a técnica com um sistema de Detecção de Novos Eventos, a fim de fazer um sistema mais completo.
- Identificação de uma forma de alterar o limiar para que melhor identifique mudanças nos eventos.
- Verificar a melhor métrica para cálculo das distâncias durante a clusterização no módulo de Discriminação de Eventos da técnica.
- Melhorar a detecção e tratamentos de *Concept Drift* realizado pelo módulo de Controle.
- Investigar a aplicabilidade e alguma possível melhoria pelo uso de Time Series na solução do problema, principalmente nos pontos referentes aos medóides de cada de cada dia da janela de tempo.
- Investigar de que forma *Topic Drift* se encaixaria e poderia melhorar o desempenho do sistema.

Referências Bibliográficas

AL-SHARIF, Z. A., JARARWEH, Y., AL-DAHOUD, A., etal., 2017, "ACCRS: autonomic based cloud computing resource scaling", *Cluster Computing*, v. 20, n. 3, pp. 2479–2488.

ALLAN, J., 2002, "Introduction to topic detection and tracking". In: *Topic detection and tracking*, Springer, pp. 1–16.

ALLAN, J., CARBONELL, J., DODDINGTON, G., etal., 1998, "Topic detection and tracking pilot study: Final report". In: *Proceedings of the DARPA broadcast news transcription and understanding workshop*, v. 1998, pp. 194–218. Citeseer.

ALLAN, J., LAVRENKO, V., JIN, H., 2000, "First story detection in TDT is hard". In: *Proceedings of the ninth international conference on Information and knowledge management*, pp. 374–381. ACM.

AMOUI, M., SALEHIE, M., MIRARAB, S., etal., 2008, "Adaptive action selection in autonomic software using reinforcement learning". In: *Autonomic and Autonomous Systems, 2008. ICAS 2008. Fourth International Conference on*, pp. 175–181. IEEE.

ANDERSON, J. A., 1995, *An introduction to neural networks*. MIT press.

ARCAINI, P., RICCOBENE, E., SCANDURRA, P., 2015, "Modeling and analyzing MAPE-K feedback loops for self-adaptation". In: *Proceedings of the 10th international symposium on software engineering for adaptive and self-managing systems*, pp. 13–23. IEEE Press.

BARALIS, E., CAGLIERO, L., MAHOTO, N., etal., 2013, "GRAPHSUM: Discovering correlations among multiple terms for graph-based summarization", *Information Sciences*, v. 249, pp. 96–109.

BARRIOS, F., LÓPEZ, F., ARGERICH, L., etal., 2016, "Variations of the similarity function of textrank for automated summarization", *arXiv preprint arXiv:1602.03606*.

BEEFERMAN, D., BERGER, A., LAFFERTY, J., 1997, "A model of lexical attraction and repulsion". In: *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pp. 373–380. Association for Computational Linguistics.

BEEFERMAN, D., BERGER, A., LAFFERTY, J., 1999, "Statistical models for text segmentation", *Machine learning*, v. 34, n. 1-3, pp. 177–210.

BELL, A., 1991, *The language of news media*. Blackwell Oxford.

BERAN, R., OTHERS, 1977, "Minimum Hellinger distance estimates for parametric models", *The annals of Statistics*, v. 5, n. 3, pp. 445–463.

BRIN, S., PAGE, L., 1998, "The anatomy of a large-scale hypertextual web search engine", *Computer networks and ISDN systems*, v. 30, n. 1-7, pp. 107–117.

BROWN, R. D., 2002, "Dynamic stopwording for story link detection". In: *Proceedings of the second international conference on Human Language Technology Research*, pp. 190–193. Morgan Kaufmann Publishers Inc.

CAI, H., HUANG, Z., SRIVASTAVA, D., etal., 2015, "Indexing evolving events from tweet streams", *IEEE Transactions on Knowledge and Data Engineering*, v. 27, n. 11, pp. 3001–3015.

CARBONELL, J., YANG, Y., LAFFERTY, J., etal., 1999, "CMU report on TDT-2: Segmentation, detection and tracking". In: *Proceedings of the DARPA broadcast news workshop*, pp. 117–120.

CHEN, F., FARAHAT, A., BRANTS, T., 2003, "Story link detection and new event detection are asymmetric". In: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003–short papers-Volume 2*, pp. 13–15. Association for Computational Linguistics.

CHEN, F., FARAHAT, A., BRANTS, T., 2004, "Multiple similarity measures and source-pair information in story link detection". In: *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*.

CHO, K., VAN MERRIËNBOER, B., GULCEHRE, C., etal., 2014a, "Learning phrase representations using RNN encoder-decoder for statistical machine translation", *arXiv preprint arXiv:1406.1078*.

CHO, K., VAN MERRIËNBOER, B., GULCEHRE, C., etal., 2014b, "Learning phrase representations using RNN encoder-decoder for statistical machine translation", *arXiv preprint arXiv:1406.1078*.

CIERI, C., STRASSEL, S., GRAFF, D., etal., 2002, "Corpora for topic detection and tracking". In: *Topic detection and tracking*, Springer, pp. 33–66.

COLLOBERT, R., 2011, "Deep learning for efficient discriminative parsing". In: *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 224–232.

COLLOBERT, R., WESTON, J., BOTTOU, L., etal., 2011, "Natural language processing (almost) from scratch", *Journal of Machine Learning Research*, v. 12, n. Aug, pp. 2493–2537.

COMPUTING, A., OTHERS, 2006, "An architectural blueprint for autonomic computing", *IBM White Paper*, v. 31, pp. 1–6.

DASIGI, P., HOVY, E., 2014, "Modeling newswire events using neural networks for anomaly detection". In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 1414–1422.

DEL CORSO, G. M., GULLI, A., ROMANI, F., 2005, "Ranking a stream of news". In: *Proceedings of the 14th international conference on World Wide Web*, pp. 97–106. ACM.

DENG, L., DING, Z., XU, B., etal., 2011a, "Exploring event evolution patterns at the atomic level". In: *Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2011 International Conference on*, pp. 40–47. IEEE, a.

DENG, L., XU, B., ZHANG, L., etal., 2013, "Event evolution analysis in microblogging based on a view of public opinion field". In: *Computational Intelligence and Design (ISCID), 2013 Sixth International Symposium on*, v. 2, pp. 193–197. IEEE.

DENG, S., MITSUBUCHI, T., SHIODA, K., etal., 2011b, "Combining technical analysis with sentiment analysis for stock price prediction". In: *2011 IEEE Ninth International Conference on Dependable, Autonomic and Secure Computing*, pp. 800–807. IEEE, b.

DOU, W., WANG, X., SKAU, D., etal., 2012, "Leadline: Interactive visual analysis of text data through event identification and exploration". In: *Visual*

Analytics Science and Technology (VAST), 2012 IEEE Conference on, pp. 93–102. IEEE.

ERKAN, G., RADEV, D. R., 2004, "Lexrank: Graph-based lexical centrality as salience in text summarization", *Journal of artificial intelligence research*, v. 22, pp. 457–479.

FARAHAT, A., CHEN, F., BRANTS, T., 2003, "Optimizing story link detection is not equivalent to optimizing new event detection". In: *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pp. 232–239. Association for Computational Linguistics.

FATTAH, M. A., 2014, "A hybrid machine learning model for multi-document summarization", *Applied intelligence*, v. 40, n. 4, pp. 592–600.

FATTAH, M. A., REN, F., 2009, "GA, MR, FFNN, PNN and GMM based models for automatic text summarization", *Computer Speech & Language*, v. 23, n. 1, pp. 126–144.

FAUSETT, L. V., OTHERS, 1994, *Fundamentals of neural networks: architectures, algorithms, and applications*, v. 3. Prentice-Hall Englewood Cliffs.

FERREIRA, R., DE SOUZA CABRAL, L., LINS, R. D., etal., 2013, "Assessing sentence scoring techniques for extractive text summarization", *Expert systems with applications*, v. 40, n. 14, pp. 5755–5764.

FUKUMOTO, F., SUZUKI, Y., 2000, "Event tracking based on domain dependency". In: *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 57–64. ACM.

GAMA, J., ŽLIOBAITĖ, I., BIFET, A., etal., 2014, "A survey on concept drift adaptation", *ACM computing surveys (CSUR)*, v. 46, n. 4, pp. 44.

GAMBHIR, M., GUPTA, V., 2017, "Recent automatic text summarization techniques: a survey", *Artificial Intelligence Review*, v. 47, n. 1, pp. 1–66.

GRUENHEID, A., KOSSMANN, D., REKATSINAS, T., etal., 2015, "StoryPivot: comparing and contrasting story evolution". In: *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pp. 1415–1420. ACM.

HAUPTMANN, A. G., WITBROCK, M. J., 1998, "Story segmentation and detection of commercials in broadcast news video". In: *Research and Technology Advances in Digital Libraries, 1998. ADL 98. Proceedings. IEEE International Forum on*, pp. 168–179. IEEE.

HAYKIN, S., 1994, *Neural networks: a comprehensive foundation*. Prentice Hall PTR.

HE, H., CUI, L., ZHOU, F., etal., 2017, "Distributed proxy cache technology based on autonomic computing in smart cities", *Future Generation Computer Systems*, v. 76, pp. 370–383.

HEARST, M. A., PLAUNT, C., 1993, "Subtopic structuring for full-length document access". In: *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 59–68. ACM.

HOCHREITER, S., SCHMIDHUBER, J., 1997a, "Long short-term memory", *Neural computation*, v. 9, n. 8, pp. 1735–1780.

HOCHREITER, S., SCHMIDHUBER, J., 1997b, "Long short-term memory", *Neural computation*, v. 9, n. 8, pp. 1735–1780.

HORN, P., 2001, "Autonomic computing: IBM\'s Perspective on the State of Information Technology", .

HSU, W., CHANG, S.-F., HUANG, C.-W., etal., 2003, "Discovery and fusion of salient multimodal features toward news story segmentation". In: *Storage and Retrieval Methods and Applications for Multimedia 2004*, v. 5307, pp. 244–259. International Society for Optics and Photonics.

HUANG, L., LV, S., ZANG, L., etal., 2018, "A Fresh Look at Understanding News Events Evolution". In: *Companion of the The Web Conference 2018 on The Web Conference 2018*, pp. 29–30. International World Wide Web Conferences Steering Committee.

HUEBSCHER, M. C., MCCANN, J. A., 2008, "A survey of autonomic computing—degrees, models, and applications", *ACM Computing Surveys (CSUR)*, v. 40, n. 3, pp. 7.

IGLESIA, D. G. D. L., WEYNS, D., 2015, "MAPE-K formal templates to rigorously design behaviors for self-adaptive systems", *ACM Transactions on Autonomous and Adaptive Systems (TAAS)*, v. 10, n. 3, pp. 15.

İLHAN, N., ÖĞÜDÜCÜ, Ş. G., 2015, "Predicting community evolution based on time series modeling". In: *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, pp. 1509–1516. ACM.

KEPHART, J. O., CHESS, D. M., 2003, "The vision of autonomic computing", *Computer*, , n. 1, pp. 41–50.

KIKUCHI, Y., HIRAO, T., TAKAMURA, H., etal., 2014, "Single document summarization based on nested tree structure". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, v. 2, pp. 315–320.

KO, Y., SEO, J., 2008, "An effective sentence-extraction technique using contextual information and statistical approaches for text summarization", *Pattern Recognition Letters*, v. 29, n. 9, pp. 1366–1371.

KUMARAN, G., ALLAN, J., 2004, "Text classification and named entities for new event detection". In: *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 297–304. ACM.

LALANDA, P., MCCANN, J. A., DIACONESCU, A., 2013, *Autonomic computing*. Springer.

LEBAN, G., FORTUNA, B., BRANK, J., etal., 2014a, "Event registry: learning about world events from news". In: *Proceedings of the 23rd International Conference on World Wide Web*, pp. 107–110. ACM, a.

LEBAN, G., FORTUNA, B., BRANK, J., etal., 2014b, "Event registry: learning about world events from news". In: *Proceedings of the 23rd International Conference on World Wide Web*, pp. 107–110. ACM, b.

LECUN, Y., BENGIO, Y., HINTON, G., 2015, "Deep learning", *nature*, v. 521, n. 7553, pp. 436.

LI, X., ZHENG, Y., DONG, Y., 2014, "Discovering Evolution of Complex Event Based on Correlations Between Events". In: *Web Information System and Application Conference (WISA), 2014 11th*, pp. 47–50. IEEE.

LIN, C.-Y., 2004, "Rouge: A package for automatic evaluation of summaries", *Text Summarization Branches Out*.

MAATEN, L. V. D., HINTON, G., 2008, "Visualizing data using t-SNE", *Journal of machine learning research*, v. 9, n. Nov, pp. 2579–2605.

MAKKONEN, J., 2003, "Investigations on event evolution in TDT". In: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Proceedings of the HLT-NAACL 2003 student research workshop-Volume 3*, pp. 43–48. Association for Computational Linguistics.

MARGOT, C. L., PEHA, S., 2006. "Be a Better Writer". .

MELE, I., BAHRAINIAN, S. A., CRESTANI, F., 2017, "Linking news across multiple streams for timeliness analysis". In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 767–776. ACM.

MENDOZA, M., BONILLA, S., NOGUERA, C., etal., 2014, "Extractive single-document summarization based on genetic operators and guided local search", *Expert Systems with Applications*, v. 41, n. 9, pp. 4158–4169.

MIHALCEA, R., TARAU, P., 2004, "Textrank: Bringing order into text". In: *Proceedings of the 2004 conference on empirical methods in natural language processing*.

MIRANDA, F. F., REIS, Á. P. D. B. B., OTHERS, 2017, "Automatically Finding Matches Between Social Media Posts and News Articles". In: *Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence & Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech), 2017 IEEE 15th Intl*, pp. 1039–1046. IEEE.

MORAES, T. R. D. S., 2016, *Sistema Autonômico de Rastreamento de Tópicos*. Master Thesis, Universidade Federal do Rio de Janeiro, Rio de Janeiro.

NALLAPATI, R., FENG, A., PENG, F., etal., 2004, "Event threading within news topics". In: *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pp. 446–453. ACM.

NEO, S.-Y., ZHENG, Y., CHUA, T.-S., etal., 2006, "News video search with fuzzy event clustering using high-level features". In: *Proceedings of the 14th ACM international conference on Multimedia*, pp. 169–172. ACM.

PENNINGTON, J., SOCHER, R., MANNING, C., 2014, "Glove: Global vectors for word representation". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543.

PEREIRA, F. R. S., 2018, *Relationship Between Detected Events in Online Media*. Ph.D. Thesis, Universidade Federal do Rio de Janeiro.

PETROVIĆ, S., OSBORNE, M., LAVRENKO, V., 2010, "Streaming first story detection with application to twitter". In: *Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics*, pp. 181–189. Association for Computational Linguistics.

PETROVIĆ, S., OSBORNE, M., LAVRENKO, V., 2012, "Using paraphrases for improving first story detection in news and Twitter". In: *Proceedings of the 2012 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pp. 338–346. Association for Computational Linguistics.

PINHEIRO, W., SILVA, M. C. O., RODRIGUES, T., etal., 2010, "Discarding Similar Data with Autonomic Data Killing Framework Based on High-Level Petri Net Rules: An RSS Implementation". In: *Autonomic and Autonomous Systems (ICAS), 2010 Sixth International Conference on*, pp. 110–115. IEEE.

PINHEIRO, W. A., RODRIGUES, T. D. S., DA SILVA, M. A., etal., 2009a, "Autonomic RSS: Discarding Irrelevant News". In: *Autonomic and Autonomous Systems, 2009. ICAS'09. Fifth International Conference on*, pp. 148–153. IEEE, a.

PINHEIRO, W. A., SILVA, M. C., BARROS, R., etal., 2009b, "Autonomic collaborative RSS: An implementation of autonomic data using data killing patterns". In: *Computer Supported Cooperative Work in Design, 2009. CSCWD 2009. 13th International Conference on*, pp. 492–497. IEEE, b.

POP, F., DOBRE, C., COSTAN, A., 2017. "AutoCompBD: Autonomic Computing and Big Data platforms". .

ROGERS, D. J., TANIMOTO, T. T., 1960, "A computer program for classifying plants", *Science*, v. 132, n. 3434, pp. 1115–1118.

ROSENBERG, A., HIRSCHBERG, J., 2006, "Story segmentation of brodcast news in English, Mandarin and Arabic". In: *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pp. 125–128. Association for Computational Linguistics.

RUTTEN, E., MARCHAND, N., SIMON, D., 2017, "Feedback control as MAPE-K loop in autonomic computing". In: *Software Engineering for Self-Adaptive Systems III. Assurances*, Springer, pp. 349–373.

SCHLIMMER, J. C., GRANGER, R. H., 1986, "Incremental learning from noisy data", *Machine learning*, v. 1, n. 3, pp. 317–354.

SCHNEIDER, D. S., 2015, *Uma Abordagem de Computação Social para a Construção de Histórias e Tramas Noticiosas por Meio da Curadoria Social*. Ph.D. Thesis, Universidade Federal do Rio de Janeiro. `http://core.ac.uk/download/pdf/10885627.pdf` .

SETTY, V., ANAND, A., MISHRA, A., etal., 2017, "Modeling event importance for ranking daily news events". In: *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pp. 231–240. ACM.

SHAH, C., EGUCHI, K., 2009, "Improving document representation for story link detection by modeling term topicality", *IPSJ Online Transactions*, v. 2, pp. 27–35.

SHAH, C., CROFT, W. B., JENSEN, D., 2006, "Representing documents with named entities for story link detection (SLD)". In: *Proceedings of the 15th ACM international conference on Information and knowledge management*, pp. 868–869. ACM.

SHI, L.-L., LIU, L., WU, Y., etal., 2017, "Event detection and user interest discovering in social media data streams", *IEEE Access*, v. 5, pp. 20953–20964.

SMITH, R. D., 2012, *Becoming a public relations writer: A writing workbook for emerging and established media*. Routledge.

ŠTAJNER, T., GROBELNIK, M., 2009, "Story link detection with entity resolution". In: *WWW 2009 Workshop on Semantic Search*, v. 2, pp. 3–2.

STOKES, N., CARTHY, J., 2001, "Combining semantic and syntactic document classifiers to improve first story detection". In: *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 424–425. ACM.

WEI, C.-P., CHANG, Y.-H., 2007, "Discovering event evolution patterns from document sequences", *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, v. 37, n. 2, pp. 273–283.

WEI, C.-P., LEE, Y.-H., CHIANG, Y.-S., etal., 2009, "Discovering event episodes from news corpora: a temporal-based approach". In: *Proceedings of the 11th International Conference on Electronic Commerce*, pp. 72–80. ACM.

WIDMER, G., KUBAT, M., 1996, "Learning in the presence of concept drift and hidden contexts", *Machine learning*, v. 23, n. 1, pp. 69–101.

WOLFE, T., 2010, *The New Journalism*. Reino Unido, Pan MacMillan. ISBN: 1581130155.

YAMRON, J. P., CARP, I., GILLICK, L., etal., 1998, "A hidden Markov model approach to text segmentation and event tracking". In: *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, v. 1, pp. 333–336. IEEE.

YANG, C. C., SHI, X., 2006, "Discovering event evolution graphs from newswires". In: *Proceedings of the 15th international conference on World Wide Web*, pp. 945–946. ACM.

YANG, J., MCAULEY, J., LESKOVEC, J., etal., 2014a, "Finding progression stages in time-evolving event sequences". In: *Proceedings of the 23rd international conference on World wide web*, pp. 783–794. ACM, a.

YANG, L., CAI, X., ZHANG, Y., etal., 2014b, "Enhancing sentence-level clustering with ranking-based clustering framework for theme-based summarization", *Information sciences*, v. 260, pp. 37–50.

YANG, Y., AULT, T., PIERCE, T., etal., 2000, "Improving text categorization methods for event tracking". In: *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 65–72. ACM.

YEH, J.-Y., KE, H.-R., YANG, W.-P., etal., 2005, "Text summarization using a trainable summarizer and latent semantic analysis", *Information processing & management*, v. 41, n. 1, pp. 75–95.

ZHAO, C., PENG, Q., LI, C., etal., 2008, "A New Method of Evolution Event Tracking Based on Variable Query". In: *Fuzzy Systems and Knowledge Discovery, 2008. FSKD'08. Fifth International Conference on*, v. 4, pp. 440–444. IEEE.

Apêndice A

Apêndice

A.1 Exemplos de Notícias de todos os *datasets*

A.1.1 Catalunha

"title": "Spain targets polling stations as Catalan referendum near",

"body": "BARCELONA - Spanish security forces on Tuesday began working to "neutralise"polling stations for a banned independence referendum in Catalonia as US President Donald Trump said Spain should remain united.With five days to go until the October 1 vote, the clash between Catalonia's pro-separatist government and Madrid was increasingly being played out in the arena of logistics and international opinion.During a joint news conference with Spanish Prime Minister Mariano Rajoy in Washington, Trump said it would be "foolish"for Catalonia not to stay in Spain."You're talking about staying with a truly great, beautiful and very historic country,"he added as Rajoy stood at his side.Rajoy urged Catalan officials to return to "common sense"even as Madrid stepped up its effort to stop the vote from going ahead.The chief prosecutor in Catalonia ordered police to seal off buildings that will house polling stations before the day of the referendum and deploy officers on the day of the vote to prevent ballots from being cast.The move comes a day after he ordered regional police to identify those in charge of polling stations on Sunday when the referendum is to be held."The order has been conveyed and it will be executed will all normality,"a spokesman for Catalonia's regional police force, the Mossos d'Esquadra, told AFP.To ensure they will cooperate, Spain's interior minister this weekend put Catalonia's regional police force under its supervision.Losing the battleBy focusing on polling stations, prosecutors appear to have put in place a plan that targets all the logistics needed to stage the referendum, which has been deemed illegal by Madrid.Prosecutors have also threatened Catalan mayors who provide locations for the vote with criminal charges, as well as directors of schools and universities.The election commission set up by Catalan separatists has resigned

after Spain's Constitutional Court threatened to impose daily fines of 12,000 euros (R188715.81).Police have seized nearly 10 million ballots for the vote and have closed down 59 websites that provide information about the referendum.Another 85 sites are in the process of being closed, judicial sources said.Faced with these actions, the separatist leaders of this wealthy northeastern region of Spain, home to around 7.5 million people, have accused Madrid of "repression".The website of the foundation of former Spanish dictator Francisco Franco "remains operational"but not the referendum websites, said Catalan government spokesman Jordi Turull.The pro-separatist camp has been buoyed by editorials published in The Financial Times and several other foreign newspapers backing a legal referendum in Catalonia or calling for dialogue.Spanish diplomatic source said the government was "aware of this claim that we are losing the battle of communication, but it is more difficult, to tell the truth than to tell lies."central government argues that it is simply applying the constitution, which does not allow this type of referendum, just as in neighbouring France and Italy.Spain's democratic constitution of 1978, which was approved by more than 90 percent of Catalan voters, gave wide autonomy to the regions but affirmed "the indissoluble unity of the Spanish nation".No early electionsRajoy pulled out of an informal summit of European Union leaders in Estonia on Friday so as to be able to attend the last meeting of his cabinet before the referendum.He also announced that the crisis would delay the national budget but ruled out fresh elections."I'm not thinking about calling early elections as a result of what we were seeing,"he said in Washington.the date of the referendum nears, tensions are rising.In Huelva in the southwestern Spain, Guardia Civil officers dispatched to Catalonia for the referendum were cheered on by hundreds of locals with cries of "Go for them!"and draped them with Spanish flags.In Barcelona, residents have been giving out red carnations to regional police.Spain has deployed two-thirds of its riot police to Catalonia, some 2,000 officers, for the referendum, according to daily newspaper El Pais."No matter how much war-like zeal they demonstrate... they are wasting time, and the (referendum) in October will take place,"the Catalan government spokesman said."

A.1.2 Crise EUA x Coréia do Norte

"title": "North Korean diplomat says Trump has 'declared war'",

"body": "UNITED NATIONS – North Korea's top diplomat said Monday that U.S. President Donald Trump's weekend tweet was a "declaration of war"and North Korea has the right to retaliate by shooting down U.S. bombers, even in international airspace.It was the latest escalation in a week of undiplomatic exchanges between North Korea and the U.S. during the U.N. General Assembly's annual ministerial meeting.Foreign Minister Ri Yong Ho told reporters that the United Nations and

the international community have said in recent days that they didn't want "the war of words"to turn into "real action."But he said that by tweeting that North Korea's leadership led by Kim Jong Un "won't be around much longer,"Trump "declared the war on our country."Under the U.N. Charter, Ri said, North Korea has the right to self-defense and "every right"to take countermeasures, "including the right to shoot down the United States strategic bombers even when they're not yet inside the airspace border of our country."Hours later, the White House pushed back on Ri's claim, saying, "We have not declared war"on North Korea. The Trump administration, referring to Trump's tweet, also clarified it is not seeking to overthrow North Korea's government.Trump tweeted on Saturday: "Just heard Foreign Minister of North Korea speak at U.N. If he echoes thoughts of Little Rocket Man, they won't be around much longer!"Trump also used the derisive "Rocket Man"reference to Kim in his speech to the U.N. General Assembly on Sept. 19, but this time he added the word "little."This was not the first time North Korea has spoken about a declaration of war between the two countries. In July 2016, Pyongyang said U.S. sanctions imposed on Kim were "a declaration of war"against the Democratic People's Republic of Korea – the country's official name – and it made a similar statement after a new round of U.N. sanctions in December. The North Korean leader used the words again Friday.The foreign minister's brief statement to a throng of reporters outside his hotel before heading off in a motorcade, reportedly to return home, built on the escalating rhetoric between Kim and Trump."The United States has great strength and patience, but if it is forced to defend itself or its allies, we will have no choice but to totally destroy North Korea,"Trump had told world leaders on Sept. 19. "Rocket Man is on a suicide mission for himself and for his regime."Kim responded with the first-ever direct statement from a North Korean leader against a U.S. president, lobbing a string of insults at Trump."I will surely and definitely tame the mentally deranged U. S. dotard with fire,"he said, choosing the rarely used word "dotard,"which means an old person who is weak-minded."Now that Trump has denied the existence of and insulted me and my country in front of the eyes of the world and made the most ferocious declaration of a war in history that he would destroy the DPRK,"Kim said, "we will consider with seriousness exercising of a corresponding, highest level of hard-line countermeasure in history."All U.N. members and the world "should clearly remember that it was the U.S. who first declared war on our country,"Ri said Monday. Ri ended his brief remarks by saying: "The question of who won't be around much longer will be answered then.

A.1.3 Incêndio Portugal e Espanha

"title": "Nine dead as wildfires ravage northern Portugal, Spain",

"body": "In the northwestern Spanish region of Galicia, authorities said that three people had died, two of them trapped in a auto, as a result of blazes that were threatening inhabited areas prompting the evacuation of thousands.The fires, some of which an official said had been started deliberately, were fanned by strong winds as remnants of ex-Hurricane Ophelia brushed the Iberian coast.At least 27 people died in a massive wave of wildfires raging in central and northern Portugal on Sunday and Monday in the worst such calamity since a blaze killed 64 people in June, civil protection told a news briefing.I nitially, officials put the death toll at six and no one was immediately available to confirm the rising death toll.Firefighters found the bodies inside a burned-out vehicle on the road linking two parishes in the city of Nigran, in the province of Pontevedra.Spain's Interior Minister Juan Ignacio Zoido said some of those responsible had already been identified.On Sunday night, over 350 firemen were battling 17 active fires, aided by 20 aircraft and 160 members of the Military Emergency Response Unit."

A.1.4 Furacão Maria

"title": "Feds rush aid to Puerto Rico, while Trump tweets about debt",

"body": "WASHINGTON (AP) - The U.S. ramped up its response Monday to the humanitarian crisis in Puerto Rico, even as President Donald Trump brought up the island's struggles before Hurricane Maria struck - including "billions of dollars"in debt to "Wall Street and the banks which, sadly, must be dealt with."The Trump administration has tried to blunt criticism that its response to Hurricane Maria has fallen short of its efforts in Texas and Florida after the recent hurricanes there.Five days after the Category 4 storm slammed into Puerto Rico, many of the more than 3.4 million U.S. citizens in the territory were still without adequate food, water and fuel. Flights off the island were infrequent, communications were spotty and roads were clogged with debris. Officials said electrical power may not be fully restored for more than a month.Trump himself pointed out some differences between the two states and the island in a series of tweets Monday night."Texas Florida are doing great but Puerto Rico, which was already suffering from broken infrastructure massive debt, is in deep trouble."Trump also noted that the island's electrical grid was already "in terrible shape."Still, he promised, "Food, water and medical are top priorities - and doing well."Washington, officials said no armada of U.S. Navy ships was headed to the island because supplies could be carried in more efficiently by plane. The Trump administration ruled out temporarily setting aside federal restrictions on foreign ships' transportation of cargo, saying it wasn't needed. The government had waived those rules in Florida and Texas until last week.Though the administration said the focus on aid was strong, when two Cabinet secretaries

spoke at a conference on another subject - including Energy Secretary Rick Perry, whose agency is helping restore the island's power - neither made any mention of Puerto Rico or Hurricane Maria.Democratic lawmakers with large Puerto Rican constituencies back on the mainland characterized the response so far as too little and too slow. The confirmed toll from Maria jumped to at least 49 on Monday, including 16 dead in Puerto Rico."Puerto Ricans are Americans,"said Rep. Nydia Velazquez, D-N.Y., who traveled to Puerto Rico over the weekend to assess the damage. "We cannot and will not turn our backs on them."Trump himself was expected at the end of last week to visit Puerto Rico and the U.S. Virgin Islands, after they had been ravaged by Hurricane Irma. But the trip was delayed after Maria set its sights on the islands.The head of the Federal Emergency Management Agency, Brock Long, and White House homeland security adviser Tom Bossert landed in San Juan on Monday, appearing with Puerto Rico Gov. Ricardo Rossello at a brief news briefing. Though Rossello had urgently called for more emergency assistance over the weekend, he expressed his gratitude for the help so far.The governor said the presence of Long and Bossert was "a clear indication that the administration is committed with Puerto Rico's recovery process."Long said, "We've got a lot of work to do. We realize that."Perry and Interior Secretary Ryan Zinke made no mention of Puerto Rico or the hurricane during a joint appearance before the National Petroleum Council, a business-friendly federal advisory committee. News reporters were not allowed to ask questions.Perry had traveled with Trump to Texas and Florida following hurricanes Harvey and Irma.Energy Department crews are working in Puerto Rico and the Virgin Islands, coordinating with the Puerto Rico Electric Power Authority, FEMA and a team from the New York Power Authority, among others. An eight-member team from the Western Area Power Authority, an Energy Department agency, assisted with initial damage assessments in Puerto Rico and has been redeployed to St. Thomas. A spokeswoman said additional responders would go to Puerto Rico as soon as transportation to the hurricane-ravaged island could be arranged.Zinke's department oversees the U.S. Virgin Islands, along with other territories.The federal response to Maria faces obvious logistical challenges beyond those in Texas or Florida. Supplies must be delivered by air or sea, rather than with convoys of trucks.FEMA said it had more than 700 staff on the ground in Puerto Rico and the U.S. Virgin Islands. They were helping coordinate a federal response that now includes more than 10,000 federal personnel spread across the two Caribbean archipelagos.In Puerto Rico, federal workers supplied diesel to fuel generators at hospitals and delivered desperately needed food and water to hard-hit communities across the island. Cargo flights are bringing in additional supplies, and barges loaded with more goods are starting to arrive in the island's ports.San Juan's international airport handled nearly 100 arrivals and departures on Sunday, including

military and relief operations, according to the Federal Aviation Administration. The Pentagon dispatched the Navy amphibious assault ship USS Kearsarge, which provided helicopters and Marines to help with the relief effort onshore.However, the Trump administration said Monday it would not waive federal restrictions on foreign ships' transportation of cargo as it had following Harvey and Irma. The administration said it will continue to enforce the Jones Act, which requires that goods transported between U.S. ports be carried on U.S.-flagged ships.Department of Homeland Security spokesman David Lapan said the agency had concluded there were already enough US-flagged vessels available.On Capitol Hill, congressional leaders were talking about how to pay for it all. Puerto Rico was already struggling from steep financial and economic challenges before Maria made landfall.Last year, House Speaker Paul Ryan and Democratic leader Nancy Pelosi joined with President Barack Obama to help recession-ravaged Puerto Rico deal with its debt crisis. After the devastating storm, Puerto Ricans will now be eligible to benefit from the same pots of federal emergency disaster aid and rebuilding funds available to residents in Texas and Florida.Lawmakers approved a $15 billion hurricane relief packaged after Harvey hit Texas, but billions more will likely now be needed to respond to Maria.Ryan said Monday that Congress will ensure the people of Puerto Rico "have what they need."Associated Press reporters Ben Fox in San Juan, Puerto Rico, and Jill Colvin, Robert Burns, Matthew Daly, Joan Lowy and Darlene Superville in Washington contributed.