Flavia Pezoti, Aline Murakami, Eric Campanatti, Gabriel Alves

Protótipo de um sistema para análise de sentimentos em redes sociais usando computação cognitiva

São Paulo

2017

Sumário

1 Introdução

Neste capítulo serão apresentados os pontos que foram decisivos na escolha do tema do projeto, e também quais serão as atividades a serem efetuadas para alcançar os objetivos estabelecidos

1.1 Motivação

Atualmente, o uso de redes sociais, como Facebook, Twitter, Instagram e outros, ocupa uma parcela significante da rotina das pessoas. Bilhões de internautas as utilizam diariamente não apenas para entrar em contato com seus amigos, conhecer novas amizades e divulgar informações que julgam interessantes, mas também expressar e compartilhar, por meio de postagens, vídeos e fotos, seus pontos de vista a respeito de uma extensa gama de assuntos.

Uma postagem criada por um usuário pode gerar muitas informações a respeito de seu estilo de vida: em qual cidade ele mora, onde costuma fazer compras, quais suas preferências musicais, etc. Estas informações extraídas da maneira correta se transformam em dados sociais. A crescente disponibilidade destes dados sociais é extremamente benéfica para tarefas como branding, análise de produtos, gerenciamento de reputação corporativa e marketing de mídia social (PORIA et al., 2015).

No entanto, o volume de dados criado pelas redes sociais torna ineficiente que pessoas obtenham, pesquisem e classifiquem dados sem ajuda computacional (AKAICHI, 2014). Além disso, a descoberta de informações úteis a partir desta enorme quantidade de dados não estruturados continua a ser um desafio aberto (BRYNIELSSON; JOHANSSON; WESTLING, 2013).

A natureza complexa dos conteúdos compartilhados em redes sociais requer técnicas avançadas de aprendizado de máquina e processamento de linguagem natural, para que se possa extrair opiniões dos usuários sobre um determinado tema (MUTHUTANTRIGE; WEERASINGHE, 2016). Embora esses dados sejam facilmente compreensíveis aos seres humanos, eles não são adequados para o processamento automático: as máquinas ainda não conseguem interpretar de forma eficaz e dinâmica o significado associado ao texto em linguagem natural em ambientes muito grandes, heterogêneos, barulhentos e ambíguos como a Web (PORIA et al., 2015).

Emular e compreender o cérebro humano é um dos principais desafios da inteligência computacional, que envolve muitos problemas-chave da inteligência artificial, incluindo a compreensão da linguagem humana, raciocínio e emoções. Dessa forma, com este trabalho

procura-se compreender e aplicar técnicas de inteligência computacional e processamento linguístico para desenvolver um algoritmo capaz de entender e identificar sentimentos associados a textos de mídias sociais dentro de uma determinada temática.

1.2 Objetivo

O objetivo deste trabalho é desenvolver o protótipo de um sistema capaz de identificar os sentimentos humanos expressados em postagens de redes sociais sobre um determinado tema.

Serão pré-determinados "grupos de sentimentos", nos quais as postagens serão enquadradas a partir de indicadores linguísticos a serem estudados ao longo da pesquisa. A expectativa é que o sistema seja capaz de enquadrar os textos em sentimentos obtendo um resultado similar ao que uma pessoa teria realizando o processo manualmente.

Ao final do projeto, será realizada uma comparação dos resultados obtidos pelo algoritmo desenvolvido durante o trabalho com um algoritmo aleatório e com os resultados obtidos por ferramentas de computação cognitiva populares no cenário atual.

1.3 Método de Trabalho

A pesquisa se dividirá em duas partes fundamentais: o estudo dos conceitos base que fundamentam o trabalho e o desenvolvimento de um algoritmo e de um protótipo, que colocam em prática os conhecimentos adquiridos durante este estudo. As macroatividades que compõem estão descritas abaixo.

Pesquisa Bibliográfica

- Levantamento de bibliografias que abordem os temas: Computação cognitiva, Machine learning, Natural Language Processing (NLP), Algoritmos de análise de emoções (Extreme Machine Learning - EML e Suport Vector Machine - SVM) e análises de postagens em redes sociais.
- Levantamento de trabalhos relacionados à temática desta pesqusa, a fim de localizar pontos que possam ser explorados ou estendidos durante a elaboração do projeto.

Estudo e Exploração Tecnológica dos Princípios de Inteligência Artifical

- Avaliação dos bancos de dados léxicos de análise de sentimentos (SentiWordNet, ConceptNet, SentiSense)
- Possibilidade de usar Watson (IBM) e/ou LEX (AWS) para análise dos dados

- Confecção do capítulo 2 do TCC, a partir dos tópicos pesquisados.

Coleta dos Dados Iniciais

Realizar a análise de sentimentos de tweets requer uma base de dados. Para isso será feito um data mining de posts do Twitter que serão o conjunto inicial para a análise.

- Datamining das postagens das redes sociais e armazenamento em um banco de dados para posterior análise. (WEKA (HALL et al., 2009) ou tweetstream (BLEIGH; AGALLOCO; MICHAELS-OBER,))

Desenvolvimento do Algoritmo para Análise de Sentimentos

- Análise dos algoritmos na literatura, adaptação e desenvolvimento de um algoritmo próprio para o escopo do problema. Consistirá na análise de eficiência dos diferentes algoritmos para textos curtos e não estruturados. E, posteriormente, adaptação para alimentar uma base de dados do Watson.
- Treinamento do Algoritmo: Nesta fase o sistema é treinado com o uso de um conjunto de treinamento constituído por textos previamente classificados para obter a probabilidade de que uma palavra seja positiva, negativa ou neutra dada a classe atribuída ao texto.
- Teste do Algoritmo: O algoritmo é testado para calcular a precisão do método.
- Uso do Algoritmo: Nesta fase, a entrada do algoritmo é um conjunto de textos não classificados e é determinado se são positivos, negativos ou neutros.

Desenvolvimento do Protótipo

Nesta atividade, o protótipo de um sistema será modelado e implementado tendo como núcleo o algoritmo de análise de sentimentos desenvolvido anteriormente. Para dar suporte a este projeto, será utilizado o método de desenvolvimento de software orientado a objetos ICONIX, que consiste nas seguintes etapas:

- Análise de Requisitos
- Análise e Design Preliminares
- Design Detalhado
- Implementação - Escrita de códigos e testes

Avaliação dos Resultados

Será feita uma análise estatística da acuracidade da análise de sentimentos do conjunto de posts de Twitter coletados na etapa anterior. Comparando o algoritmo desenvolvido com os já descritos em literatura.

1.4 Cronograma

Tabela 1 – Cronograma de Pesquisa

	2017									
	fev	mar	abr	maio	jun	jul	ago	set	out	nov
Pesquisa Bibliográfica	x	x	x	x						
Estudo e Exploração Tecnológica dos Princípios de Inteligência Artifical			x	x	x	x				
Coleta dos Dados Iniciais					x	x				
Desenvolvimento do Algoritmo para Análise de Sentimentos						x	x			
Desenvolvimento do Protótipo							x	x	x	
Avaliação dos Resultados									x	x

1.5 Organização do Texto

No capítulo 1, Introdução, são apresentados o tema, a motivação para a realização da pesquisa, objetivo do trabalho, contextualização dentro da área de Ciência da Computação e o método utilizado para atingir o objetivo.

No capítulo 2, Revisão Bibliográfica, é apresentada a Fundamentação Teórica demandada para a realização dessa pesquisa bem como identificados e analisados Trabalhos Relacionados de apoio.

No capítulo 3, o trabalho de pesquisa completo será apresentado com seus experimentos e atividades. O processo de desenvolvimento do protótipo será descrito em detalhes, com ênfase no algoritmo de classificação de emoções.

No capítulo 4, Avaliação dos Resultados, haverá uma descrição dos resultados obtidos durante o projeto e uma comparação crítica com os resultados esperados durante a idealização do trabalho e revisão bibliográfica.

No capítulo 5, Conclusões, haverá uma avaliação do método utilizado, assim como sugestão de linhas de pesquisa que possam complementar o trabalho no futuro.

2 Fundamentação Teórica

Em relação à fundamentação teórica, é cabível abordar de maneira mais específica os conceitos de Inteligência Artifical, Computação Cognitiva, Análise de Sentimentos, Redes Sociais e o método de desenvolvimento de software ICONIX.

2.1 Inteligência Artifical

> "Atualmente, a IA abrange uma enorme variedade de subcampos, desde áreas de uso geral como aprendizado e percepção, até tarefas específicas como jogos de xadrez, demonstração de teoremas matemáticos, criação de poesia e diagnóstico de doenças. A IA sistematiza e automatiza tarefas intelectuais e, portanto, é potencialmente relevante para qualquer esfera de atividade intelectual humana. Nesse sentido, ela é verdadeiramente um campo universal. " (RUSSEL; NORVIG, 2013)

De acordo com RUSSEL; NORVIG (2013), "Se pretendemos dizer que um dado programa pensa como um ser humano, temos de ter alguma forma de determinar como os seres humanos pensam."

2.2 Computação Cognitiva

Sendo um dos ramos da Inteligência Artificial, a Computação Cognitiva é a capacidade de máquinas pensarem como humanos. Esse termo ganhou grande repercussão quando o Watson da IBM conseguiu vencer o programa de TV Jeopardy em 2011.

No entanto, um dos empecilhos para o avanço da tecnologia é a linguagem. Qualquer idioma é cheio de insinuações, idiossincrasias, expressões idiomáticas e ambiguidades, porém é transmitido muito significado nestes contextos "não exatos", como 4x4 não necessariamente é 16, podendo ser também uma característica de carro.

"IBM Watson é um sistema de NLP(Natural Language Processing) profundo. Ele alcança a precisão tentando analisar a maior quantia de contextos possíveis. Ele obtém este contexto tanto da passagem da pergunta quanto da base de conhecimento (chamada de corpus) que está disponível para ele localizar respostas." (HIGH, 2012)

2.3 Processamento de Linguagem Natural (NLP)

Uma linguagem natural é uma língua falada por pessoas, como Português, Inglês ou Turco. Isto é, uma oposição as linguagens artificias como Java, FORTRAN, ou código

Morse. O NLP é uma parte importante de Inteligência Artificial pois a capacidade de compreender a linguagem está intimamente ligada ao pensamento.

NLP e Computação

A linguagem é uma das ferramentas centrais na vida social e professional das pessoas. Dentre outras coisas, esta age como um meio de transmissão de ideias, informação, opiniões e sentimentos, assim como para persuadir, perguntar por informações, transmitir ordens, entre outros. (KURDI, 2016).

A Ciência da Computação começou a ganhar interesse por análise linguística assim que o próprio campo emergiu, principalmente no campo da Inteligência Artificial (AI). O teste de Turing, um dos primeiros testes desenvolvidos para julgar se uma máquina é inteligente ou não, estipula que para ser considerado inteligente, uma máquina deve possuir habilidades de conversação comparáveis as de um ser humano (TURING, 1950). Isso implica que uma máquina inteligente deve possuir habilidades de compreensão e produção, no sentido mais amplo desses termos (KURDI, 2016). Ou seja, deve ser capaz de processar a linguagem de comunicação, aprender as informações contidas na mensagem, e ser capaz de transmitir o que foi aprendido. Em um contexto prático, a NLP é análoga ao ensino de uma língua para uma criança. Algumas das tarefas mais comuns, como a compreensão de palavras, frases e formação de sentenças gramaticalmente e estruturalmente corretas, são muito naturais para os seres humanos, mas não triviais para computadores. Na NLP, algumas dessas tarefas se traduzem em tokenização, fragmentação, parte de speech tagging, análise, tradução automática, reconhecimento de fala, análise de sentimento, e a maioria deles ainda são os desafios mais difíceis na área de computação(MATHUR; AL, 2016).

Definições importantes

O Processamento de Linguagem Natural, ou NLP, é uma disciplina que se encontra na interseção de vários ramos da ciência, como Ciência da Computação, Inteligência Artificial e Psicologia Cognitiva.

Existem várias definições para a área ainda não completamente consolidadas. Mas como descrito por KURDI (**??**), por exemplo, tem-se que os termos linguística formal ou linguística computacional relacionam-se com modelos ou formalidades linguísticas desenvolvidos para implementação de TI. E os termos Tecnologia de Linguagem Humana ou Processamento de Linguagem Natural, por outro lado, referem-se a uma ferramenta de software de publicação equipada com recursos relacionados ao processamento de linguagem. Além disso, o processamento de fala designa uma gama de técnicas de processamento de sinais para o reconhecimento ou produção de unidades linguísticas, tais como fonemas, sílabas ou palavras. Exceto para a dimensão lidar com o processamento de sinal, não

há grande diferença entre o processamento de fala e NLP. Muitas técnicas que foram inicialmente aplicadas ao processamento da fala encontraram seu caminho em aplicações em NLP, um exemplo são os Modelos de Markov Ocultos (HMM). Finalmente, vale a pena mencionar o termo *corpus linguistics* que se refere aos métodos de coleta, anotação e uso de corpus, tanto na pesquisa lingüística quanto na NLP. Os corpus têm um papel muito importante no processo de construção de um sistema de NLP, especialmente aqueles que adotam uma abordagem de machine learning.

NPL e IA

Inteligencia Artificial (AI) tem como uma de suas descrições o estudo, design e criação de agentes inteligentes. Um agente inteligente é um sistema natural ou artificial com habilidades perceptuais que lhe permite atuar em um determinado ambiente para satisfazer seus desejos ou alcançar com êxito os objetivos (RUSSEL; NORVIG, 2013). O trabalho em AI é geralmente classificado em várias subdisciplinas ou ramos, tais como representação do conhecimento, planejamento, percepção e aprendizagem. Todos esses ramos estão diretamente relacionados à NLP. Isso dá a relação entre AI e NLP uma dimensão muito importante. Muitos consideram a NLP como um ramo da AI, enquanto alguns preferem considerar a NLP uma disciplina independente.

A representação do conhecimento é importante para um sistema NLP em dois níveis: Por um lado, pode fornecer um quadro para representar os conhecimentos linguísticos necessários ao bom funcionamento de todo o sistema NLP, mesmo que o tamanho e a quantidade das informações declarativas no sistema variem consideravelmente de acordo com a abordagem escolhida. Por outro lado, alguns sistemas NLP exigem informações extralinguísticas para tomar decisões, especialmente em casos ambíguos. Portanto, certos sistemas NLP são emparelhados com ontologias ou com bases de conhecimento sob a forma de uma rede semântica, um quadro ou gráficos conceituais (KURDI, 2016).

Em teoria, a percepção e a linguagem parecem distantes umas das outras, mas na realidade, não é esse o caso. Fazer a conexão entre percepção e reconhecimento semântico é crucial, não só para a compreensão, mas também para melhorar a qualidade e interpretação da mensagem contida no texto.

NLP e Ciência Cognitiva

Assim como na análise linguística, a relação entre a ciência cognitiva e NLP vai em duas direções (KURDI, 2016). Por um lado, os modelos cognitivos podem agir para apoiar uma fonte de inspiração para um sistema de NLP. Por outro lado, a construção de um sistema NLP de acordo com um modelo cognitivo pode ser uma forma de testar este modelo. O benefício prático de uma abordagem que imita o processo cognitivo

permanece uma questão aberta porque em muitos campos, construir um sistema que é inspirado por modelos biológicos não se revela produtivo. Deve-se notar também que certas tarefas realizadas por sistemas NLP não têm paralelo em seres humanos, como a busca de informações através de mecanismos de busca ou a busca por grandes volumes de dados de texto para extrair informações úteis. A NLP pode ser vista como uma extensão das capacidades cognitivas humanas como parte de um sistema de apoio à decisão, por exemplo. Outros sistemas de NLP são muito próximos de tarefas humanas, como compreensão e produção.

NLP e Data Science

Com a disponibilidade de mais e mais dados digitais, surgiu recentemente uma nova disciplina: a Data Science (ciência dos dados). Trata-se de extrair, quantificar e visualizar o conhecimento, principalmente a partir de dados textuais e falados (KURDI, 2016). Grande parte da informação encontra-se não estruturada, pois foi especificamente criada para interpretação humana, sendo, dessa forma, de difícil processamento por máquinas. A NLP tem o papel de extrair e processar as informações para torna-las acessíveis e interpretáveis de forma automatizada. Atualmente, dados os inúmeros usos industriais para esse tipo de conhecimento, especialmente nos campos de marketing e tomada de decisões, esta área de estudo se tornou extremamente importante.

Exemplo de Aplicações de NPL na atualidade

Atualmente são gerados petabytes de Weblogs, tweets, feeds do Facebook, bate-papos, e-mails e comentários. As empresas estão coletando todos esses tipos diferentes de dados para uma melhor segmentação de clientes e insights significativos. Para processar todas essas fontes de dados não estruturadas é necessário entender e utilizar NPL. Alguns exemplos de aplicações que utilizam NPL (MATHUR; AL, 2016):

- Corretores de texto (MS Word/ e qualquer outro editor de texto com a funcionalidade)
- Search engines (Google, Bing, Yahoo, wolframalpha)
- Speech engines (Siri, Google Voice)
- Classificadores de Spam (Serviços de e-mail
- News feeds (Google, Yahoo!, e outros)
- Tradutores em máquina (Google Translate, e outros)
- IBM Watson

Para alcançar algumas das aplicações acima e outros pré-processamentos básicos de NLP, existem muitas ferramentas de código aberto disponíveis. Alguns deles são desenvolvidos por organizações para construir seus próprios aplicativos de NLP, enquanto alguns deles são open-sourced. Como exemplo, segue algumas ferramentas de NLP disponíveis (MATHUR; AL, 2016):

- GATE
- Mallet
- Open NLP
- UIMA
- Stanford toolkit
- Genism
- Natural Language Tool Kit (NLTK)

2.4 Análise de Sentimentos

A análise de sentimentos, também chamada de Opinion Mining, tem sido uma das áreas de pesquisa mais ativas no processamento de linguagem natural desde o início de 2000 (LIU, 2012). O objetivo da análise de sentimento é definir ferramentas automáticas capazes de extrair informações subjetivas de textos em linguagem natural, ou seja, opiniões e sentimentos, de modo a criar conhecimento estruturado e acessível para ser usado por um sistema de apoio à decisão.É um campo emergente de análise preocupado com a aplicação de métodos computacionais para o tratamento da subjetividade no texto, com uma série de aplicações em áreas como sistemas de recomendação, publicidade contextual e business intelligence (KUMAR, 2010).

Definição

A análise de sentimentos ou geração de sentimentos é uma das tarefas da NLP. Ela é definida como o processo de determinar os sentimentos por trás de uma sequência de caracteres (MATHUR; AL, 2016). Trata-se de uma tarefa subjetiva, pois fornece as informações sobre o texto que está sendo expresso. Pode ser definida como um problema de classificação em que a classificação pode ser de dois tipos - categorização binária (positiva ou negativa) ou categorização multi-classe (positiva, negativa ou neutra).

Quando combina-se a análise de sentimentos com a mineração de tópicos, ela é referida como análise de sentimento de tópico. A análise de sentimento pode ser realizada

usando um lexicon. O lexicon pode ser específico do domínio ou de natureza geral e podem conter uma lista de expressões positivas, expressões negativas, expressões neutras e palavras de parada. Quando uma sentença de teste aparece, então uma simples operação de consulta pode ser realizada através deste léxico (MATHUR; AL, 2016). Um exemplo de Lexicon de analises de sentimentos é o SentiSense, que compreende 2.190 sínteses e 5.496 palavras baseadas em 14 categorias emocionais (ALBORNOZ; PLAZA; GERVáS, 2012).

Mais formalmente, como definido em (LIU, 2012), uma opinião é uma quíntupla

$$(e_i, a_{ij}, s_{ijkl}, h_k, t_l)$$

Em que e_i é o nome de uma entidade, a_{ij} é um aspecto de e_i, s_{ikjl} é o sentimento no aspecto a_{ij} da entidade e_i , h_k denota o detentor da opinião, e t_l é o momento em que a opinião é expressa por h_k .

O sentimento s_{ikjl} pode ser positivo, negativo ou neutro, ou expresso com diferentes níveis de intensidade / intensidade, como o sistema de 1 a 5 estrelas usado pela maioria dos sites de revisão (por exemplo, em avaliações de items da Amazon).

Análises de Sentimentos em Mídias sociais

A grande difusão das redes sociais e seu papel na sociedade moderna estão entre as novidades mais interessantes dos últimos anos, capturando o interesse de pesquisadores, jornalistas, empresas e governos. A densa interconexão que surge frequentemente entre os usuários gera um espaço de discussão capaz de motivar e envolver os indivíduos, vinculando as pessoas com objetivos comuns e facilitando diversas formas de socialização. Isto dá origem ao chamado "individualismo na rede": em vez de sempre contar com uma única comunidade de referência, graças às redes sociais torna-se possível estimular-se movendo-se entre mais pessoas e recursos, muitas vezes heterogêneos. As redes sociais estão, portanto, criando uma revolução digital. O aspecto mais interessante desta mudança não está unicamente relacionado com a possibilidade de promover a participação política e o ativismo. Esta revolução social influencia a vida de cada indivíduo. É a liberdade de nos expressarmos, de ter um espaço próprio onde possamos ser nós mesmos, ou de ser quem gostaríamos de ser, com poucos limites e barreiras (LIU et al., 2016).

Neste contexto, a análise de sentimentos tenta tornar evidente o que as pessoas pensam fornecendo representações, modelos e algoritmos capazes de passar de "texto simples não estruturado"para "visão complexa". (LIU et al., 2016).

Aplicações de Opinion Mining

Opiniões dizem respeito às crenças e julgamentos expressos pelas pessoas sobre um determinado tópico e podem ser um componente importante usado para tomar decisões

mais precisas em vários cenários(KUMAR, 2010). Empresas, por exemplo, têm um grande interesse em descobrir o que os clientes estão dizendo sobre seus produtos e ofertas de serviços. Os consumidores, por outro lado, beneficiariam do acesso a opiniões e opiniões de outras pessoas sobre os produtos que desejam comprar, uma vez que as recomendações de outros usuários tendem a influenciar essas decisões. O conhecimento das opiniões de outras pessoas também é importante em outros domínios, como o ativismo político, onde, por exemplo, pode ser interessante descobrir o sentimento geral em relação a uma nova legislação ou a partidos políticos e figuras públicas(SHARMA; MITTAL; GARG, 2016); Ou na detecção de viés subjetivo em ambientes onde não deve haver nenhum, como no monitoramento de cobertura de notícias (KUMAR, 2010).

Mídias Sociais e Análises de tendências

Muitas empresas, grandes e pequenas, estão explorando se podem usar os comportamentos e comunicações on-line das pessoas para prever situações do mundo real. Por exemplo, desde 2008, o Google vem explorando se as pesquisas que chegam de usuários de todo o mundo podem permitir o rastreamento ou mesmo a previsão da ocorrência de doenças. A empresa de pesquisa demonstrou o rastreamento de duas doenças: gripe e dengue7. A Figura 1 mostra o site Flu Trends, que foi capaz de prever a incidência de gripe com base em um conjunto de palavras-chave usadas em pesquisas durante um surto local de gripe. (HENDLER; MULVEHILL, 2016).

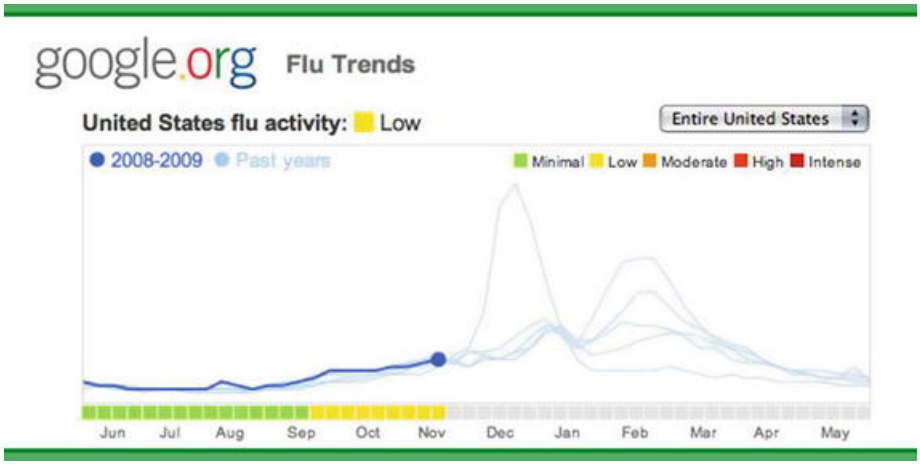

Figura 1 – Google Flu Trends

Atualmente, o Google já não publica estimativas atuais de gripe e dengue com base em padrões de busca, mas continua a oferecer estimativas históricas produzidas pelo Google Flu Trends e Google Dengue Trends. Isto, principalmente devido à falta de acuracidade dos modelos de predição, evidenciados quando o algoritmo subestimou a necessidade de vacinas nos anos de 2011-2013 nos Estados Unidos (BUTLER, 2013).

Outras empresas também estão usando feeds de mídia social, como Facebook, Twitter e seus equivalentes, para entender as tendências em tudo, desde previsões de tendências de consumo quanto em monitoramento de tragédias. A tecnologia AI está sendo usada de muitas maneiras por essas empresas para extrair e entender os dados. Por exemplo, algumas empresas usam técnicas de NLP para interpretar o conteúdo de mídia social, enquanto outras empresas estão explorando como a análise de sentimento das mídias sociais pode ser usada para entender o que as pessoas estão postando. Os resultados atuais indicam que o uso de NLP juntamente com análise de sentimento ajuda a diferenciar entre "eu me sinto bem"e "eu me sinto mal", que se torna um grande diferencial no rastreamento de tendências de saúde (HENDLER; MULVEHILL, 2016).

Desafios

Infelizmente, analisar os feeds de mídia social é muito mais difícil do que parece. Suponha que alguém escreveu um tweet "Meu médico me disse para tomar aspirina. Como eu me sinto melhor #not ". Embora o hashtag #not deixa claro que "me sinto melhor "é sarcástico, se removemos o hashtag, não sabemos se o escritor está sendo sério ou sarcástico.

Os sistemas atuais resolvem este problema através do acoplamento de técnicas de Machine Learning com técnicas de análise de linguagem. No entanto, a fim de serem eficientes, estas técnicas exigem um grande número de exemplos para ser "marcado"por pessoas, o que significa que as pessoas explicitamente precisam indicar recursos como sentimento e sarcasmo. Uma vez que o computador tem essas informações adicionais anotadas, resultados de pesquisas atuais indicam que o computador pode, então, mais precisamente correlacionar grandes conjuntos de dados para descobrir quais palavras prever, quais recursos e usar os resultados para analisar novos conjuntos de dados (HENDLER; MULVEHILL, 2016).

Embora a pesquisa em PNL tenha feito grandes progressos na produção de comportamentos artificialmente inteligentes, por exemplo, Google, IBM Watson e Apple Siri, nenhuma dessas estruturas NLP por si só realmente entendem o que eles estão fazendo - tornando-os não muito diferentes de um papagaio que aprende a repetir palavras sem nenhuma compreensão clara do que está dizendo. Hoje, mesmo as tecnologias de NLP mais populares visualizam a análise de texto como uma tarefa de correspondência de palavras ou padrões. Tentar verificar o significado de um pedaço de texto, processando-o no nível de palavra, no entanto, não é diferente de tentar entender uma imagem, analisando-a em nível de pixel (CAMBRIA; HUSSAIN, 2015).

Basear-se em palavras-chave arbitrárias, pontuação e frequências de co-ocorrência de palavras funciona bem para textos estruturados, mas uma grande quantidade de conteúdo gerado pelo usuário e o surgimento de fenômenos enganosos, como spam na web

e de opinião, estão causando que os algoritmos de NLP padrão sejam cada vez menos eficientes. Para extrair e manipular adequadamente significados de texto, um sistema NLP deve ter acesso a uma quantidade significativa de conhecimento sobre o mundo e o domínio do discurso. Para este fim, os sistemas de PNL irão gradualmente parar de confiar muito em técnicas baseadas em palavras enquanto começam a explorar a semântica de forma mais consistente e, portanto, dão um salto da Curva de sintaxe para a curva de semântica (Figura 2).

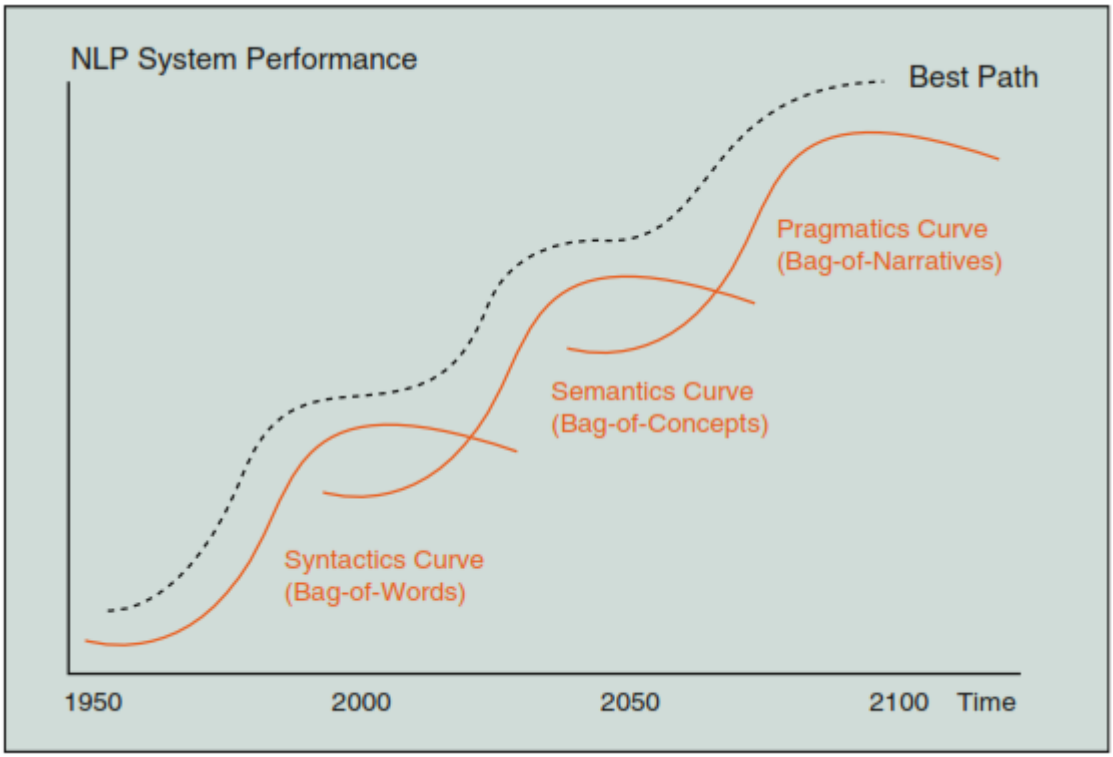

Figura 2 – Evolução prevista da pesquisa da PNL através de três eras ou curvas diferentes (CAMBRIA; HUSSAIN, 2015)

2.5 Sentic Computing

A Sentic Computing (SC), cujo termo deriva do "sentire"latino (raiz das palavras como sentimento e sensibilidade) e "sensus"(como em senso comum), é uma abordagem multidisciplinar para a compreensão da linguagem natural que visa preencher a distância entre o processamento estatístico de linguagem natural (NLP) e muitas outras disciplinas que são necessárias para a compreensão da linguagem humana, como linguística, raciocínio de senso comum, computação afetiva e entre outros (CAMBRIA; HUSSAIN, 2015). Em particular, a SC utiliza de técnicas AI e SemanticWeb, para representação e inferência do conhecimento; Matemática, para realizar tarefas como mineração de gráficos e redução de multidimensionalidade; Linguística, para análise de discurso e pragmática; Psicologia,

para modelagem cognitiva e afetiva; Sociologia, para a compreensão da dinâmica das redes sociais e da influência social; Finalmente ética, para compreender questões relacionadas com a natureza da mente e a criação de máquinas emocionais (BISIO et al., 2017). Na SC a análise da linguagem natural baseia-se em ferramentas de raciocínio de senso comum, que permitem a análise de texto, não apenas em nível de documento, página ou parágrafo, mas também em sentença, cláusula e nível de conceito. A SC é diferente dos métodos comuns para a detecção de polaridade, pois requer uma abordagem multifacetada para o problema da análise do sentimento.

Algumas das técnicas mais populares para Opinion Mining centram-se em frequências de co-ocorrência de palavras e polaridade estatística associadas a palavras. Essas abordagens podem inferir corretamente a polaridade de textos não ambíguos com estrutura de frases simples e em um domínio específico (ou seja, o qual o classificador estatístico foi treinado). Uma das principais características da linguagem natural, no entanto, é a ambiguidade. Uma palavra como "grande"não possui nenhuma polaridade por si só, pois pode ser negativa, por exemplo, no caso de um "problema grande", ou positivo, por exemplo, na "grande jantar", mas a maioria dos métodos estatísticos lhe atribuem uma polaridade positiva, pois o contexto em que aparece muitas vezes é positivo. Trabalhando no nível do conceito, a SC supera isso e muitos outros problemas comuns de framework de opinion-mining que dependem fortemente das propriedades estatísticas das palavras.

A detecção de sentimentos é, no entanto, um problema muito desafiador porque as emoções são constructos (ou seja, quantidades conceituais que não podem ser diretamente medidas) com limites distorcidos e variações substanciais da diferença na expressão e experiência individuais (BISIO; ONETOA; CAMBRIA, 2016) . Para superar esse obstáculo, a SC baseia-se em uma modelo de categorização afetiva inspirada biologicamente e motivada psicologicamente que podem potencialmente descrever toda a gama de fatores emocionais e experiências em termos de quatro dimensões independentes, mas concomitantes, cujas diferentes níveis de ativação compõem o estado emocional total da mente (CAMBRIA; HUSSAIN, 2015). Esse modelo é representado como uma ampulheta demonstrado na figura 3.

2.6 Twitter

Twitter é uma rede social onde milhões de curtas mensagens de 140 caracteres são postadas diariamente (microblogging). Esta rede tem um crescente volume de dados graças um grande número de usuários ativos (DESHWAL; SHARMA, 2016).

Por ser uma ferramenta de fácil e rápida utilização, usuários utilizam a rede para comentar situações cotidianas e/ou eventos em tempo real (shows, esportes, premiações), criando o efeito de segunda tela (Tela 1: Evento ou TV, Tela 2: Smartphone conectado

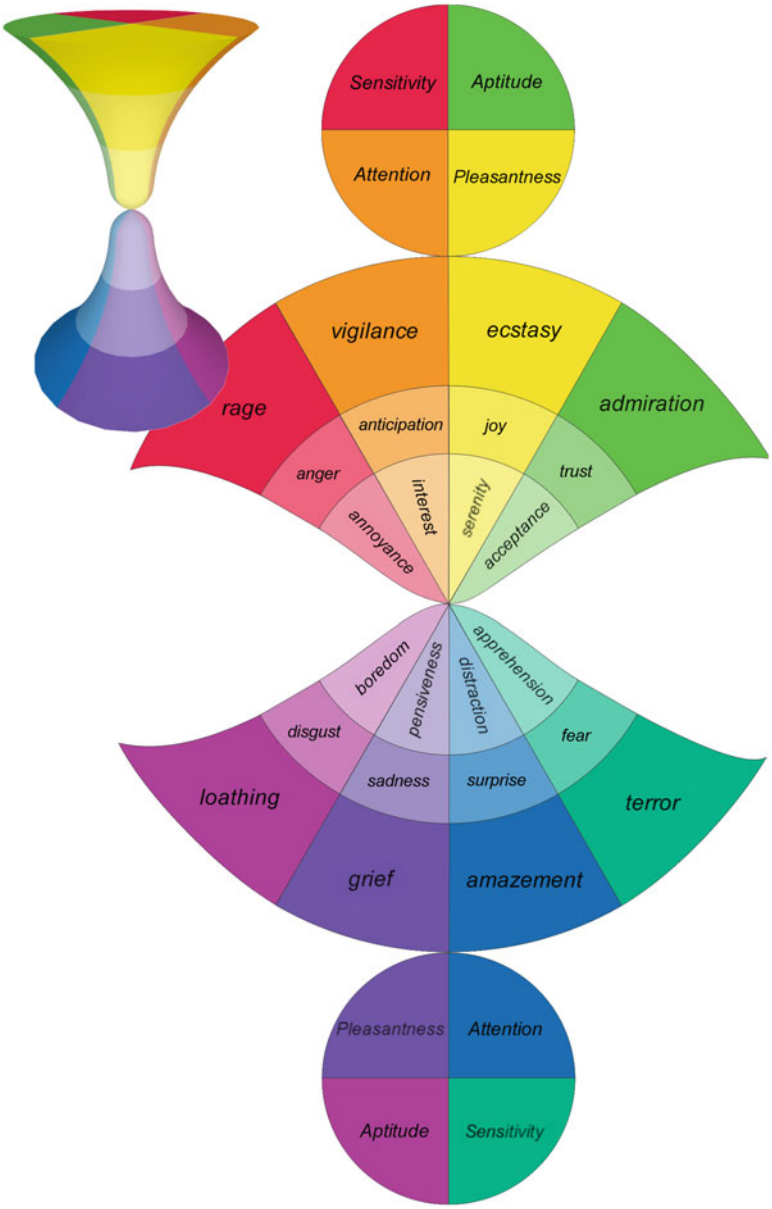

Figura 3 – O modelo 3D e a rede do "Hourglass of Emotions". Uma vez que os estados afetivos vão de fortemente positivo para nulo para fortemente negativo, o modelo assume uma forma de ampulheta. (CAMBRIA; HUSSAIN, 2015)

em redes sociais focados no assunto em si), o que tornou a rede rica em informações abrangendo diversas áreas e diversos públicos (VASUDEVAN et al., 2013).

A extração de dados dos diversos “tweets” se mostrou extremamente útil ultimamente, possibilitando saber qual é a popularidade de tópicos, e qual foi a reação dos usuários neste tópico, ainda com a possibilidade de mapear qual é o público alvo (idade, sexo, localização, etc) pelo cadastro que é efetuado no ato de criação de um perfil na rede.

Segundo o Twitter, em junho de 2016, ele contava com 313 Milhões de usuários (SOBRE...,) postando em média 500 milhões de tweets por dia (TWITTER...,).

2.7 ICONIX

O ICONIX é um método de desenvolvimento de software minimalista, focado em atuar na área entre a elaboração dos casos de uso e o do código (ROSENBERG; STEPHENS, 2007). A ênfase do método está em desenvolver uma boa análise e um bom design a partir de um processo incremental, no qual os diagramas de caso de uso são a base para cada iteração. As fases do desenvolvimento delimitadas pelo ICONIX são: Requisitos, Análise e Design Preliminares, Design Detalhado e Implementação.

Na fase de Requisitos, os requisitos funcionais (que definem quais as capacidades do sistema) pré-elaborados são analisados com o intuito de realizar um modelo de domínio, um dicionário dos termos e objetos reais do seu projeto e de como eles se relacionam superficialmente. A partir do modelo de domínio, estabelecem-se os requisitos comportamentais, que detalham as ações do usuário e como o sistema deve responder a elas. (ROSENBERG; STEPHENS, 2007) Storyboads da GUI são uma ferramenta importante na idenficação destes requisitos, enquanto diagramas de caso de uso são utilizados para documentar cada cenário encontrado.

Baseadas nos casos de uso encontrados, as demais fases se repetem a cada iteração, trabalhando alguns poucos casos de uso por vez. Durante a Análise e Design Preliminares, os casos de uso são refinados através de diagramas de robustez, ajudando a complementar o modelo de domínio com a identifição de classes antes ignoradas e dos atributos de cada classe. No Design Detalhado, cada caso de uso gera um diagrama de sequência de mensagens que descreve todas as chamadas de método que ocorrem entre os objetos naquele cenário. Com base nestes diagramas, o modelo de domínio é atualizado com os métodos descritos e torna-se um diagrama de classes.

Quando a fase de Design Detalhado acaba, a modelagem realizada até o momento deve ser capaz de descrever com clareza o código a ser escrito. Inicia-se então a fase de Implementação. As classes descritas no diagrama de classes são geradas, seus métodos são implementados e testes unitários são escritos garantir o funcionamento do sistema. Essencialmente, você deve testar todas as funções identificadas durante a análise de robustez. (ROSENBERG; STEPHENS, 2007) Também são realizados testes de integração e aceitação. Por último, revisa-se o código e atualiza-se o modelo para se preparar para a próxima iteração.

Entre cada uma das fases, atingi-se uma milestone que consiste numa revisão crítica do que foi elaborado até o momento para assegurar que os requisitos estão sendo atendidos corretamente. A elaboração dos casos de uso, diagramas de robustez e de sequência de mensagem são a parte chamada de dinâmica no ICONIX e os diagramas de domínio e de classes são a parte estática. A figura 4 apresenta uma visão geral do processo.

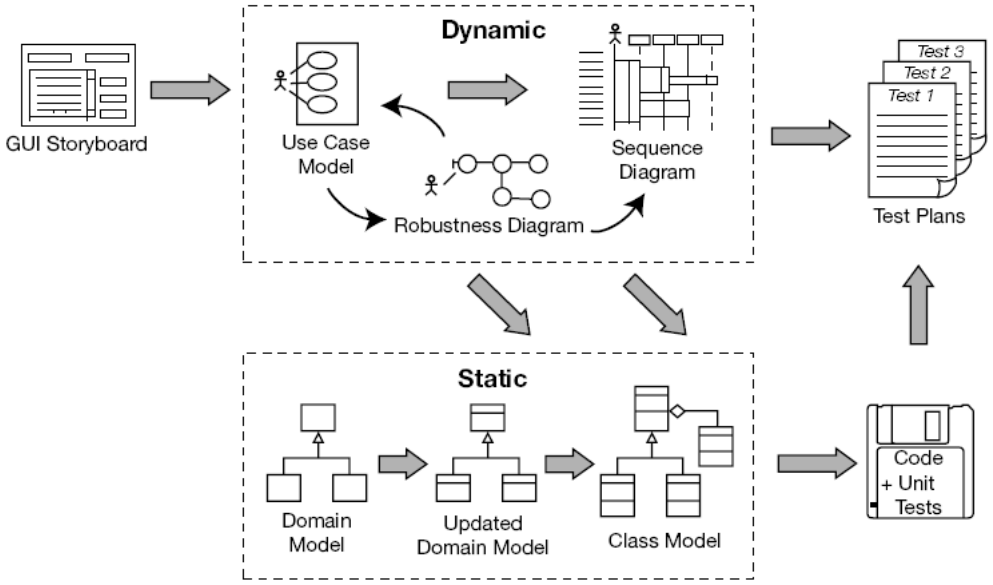

Figura 4 – Visão Geral do Processo ICONIX (ROSENBERG; STEPHENS, 2007)

2.8 Trabalhos relacionados

Emular o cérebro humano é um dos principais desafios da inteligência computacional, que envolve muitos problemas-chave da inteligência artificial, incluindo a compreensão da linguagem humana, raciocínio e emoções. A crescente disponibilidade de dados sociais é extremamente benéfica para tarefas como branding, posicionamento de produtos, gerenciamento de reputação corporativa e marketing de mídia social. Identificar informações úteis a partir desta enorme quantidade de dados não estruturados, no entanto, continua a ser um desafio aberto.

Um dos métodos convencionais de predição, é a utilização da frequência de repetição de palavras chave, para identificar uma tendência. Como descrito no Trabalho de TSUGAWA et al. (2013) foi investigado a eficácia dos registros das atividades no Twitter, correlacionando a frequência do uso de certas palavras para detectar a tendência depressiva dos usuários. Para isso, foi feito um levantamento utilizando 50 homens e mulheres japoneses com aproximadamente 20 anos usando o teste de Zung - Zung's Self-rating Depression Scale (SDS). Este teste é um método popular de análise de tendência depressiva, que se baseia em um questionário de 20 perguntas, respondidas pelo próprio usuário, em que cada questão é pontuada de 1 a 4, e o somatório constitui a pontuação de Zung. Quanto maior a pontuação, maior sua a tendência depressiva. Neste estudo, foram realizadas análises de regressão múltipla para estimar as pontuações de Zung dos participantes a partir de freqüências de palavras usadas em suas mensagens de tweet. Os resultados experimentais mostraram que existe uma correlação positiva média (coeficiente de correlação r aproximadamente 0,45) entre pontuação de Zung calculada pelo questionário e a pontuação

estimada obtida a partir do modelo de regressão. Um dos resultados interessantes deste trabalho, foi identificar que, o uso da análise de frequência de palavras como variáveis independentes não é adequado, pois quando analisadas em conjuntos de até 5 palavras chave, foi encontrado maior coeficiente de correlação.

No entanto, basear-se em palavras-chave arbitrárias, em as frequências de pontuação e ocorrência de palavras pode não ser tão eficiente levando em consideração a explosão de conteúdo da Web e o surto de fenômenos enganosos como o Web trolling e o spam, que diminuem a eficiência desses métodos. Para efetivamente extrair e manipular dinamicamente significados de textos não estruturados, um sistema de NLP verdadeiramente inteligente deve considerar dados ambíguos e ter acesso a uma quantidade significativa de conhecimento sobre os possíveis ruídos dos dados e o domínio do discurso variando no tempo.

No trabalho de PORIA et al. (2015) é descrito um novo paradigma que explora as relações entre os conceitos e padrões linguísticos em um texto para revelar o fluxo de sentimento de conceito a conceito. A principal novidade do artigo consiste em um algoritmo que atribui polaridade contextual a conceitos nos textos e flui essa polaridade através de arcos de dependência para atribuir um rótulo de polaridade final a cada sentença. Este algoritmo extrai o sentimento de cada palavra diretamente de recursos lexicais já existentes como SenticNet (CAMBRIA; OLSHER; RAJAGOPAL, 2014). Em seguida, o algoritmo aplica deslocadores de valor e combina o sentimento de palavras individuais da frase de forma semelhante aos circuitos eletrônicos, onde o sinal das fontes pode sofrer amplificação, inversão e enfraquecimento, combinando-os até obter o resultado final do balanço desses valores. Além disso, o método proposto envolve uma técnica de back-up machine-learning, que funciona nos casos em que não é encontrada informação suficiente nos recursos lexical existentes.

O algoritmo descrito por PORIA et al. (2015) utiliza o conceito de sentic pattern ou padrão de sentimento, foi definido no trabalho do mesmo grupo de pesquisa (PORIA et al., 2014), e pode ser ilustrado na figura abaixo.

O procedimento pode ser compreendido como um algoritmo de coloração de árvores que opera nos nós e arcos da árvore de dependência sintática. O algoritmo determina diretamente a polaridade para as palavras ou relações - conceitos, ou expressões multi-palavras – pertencentes nos recursos lexicais existentes. Em seguida, ele gradualmente estende os rótulos para outros arcos e nós, com as transformações necessárias determinadas por regras de sentic patterns, até obter o rótulo final para o elemento raiz, que é a saída desejada. O grupo denominou a extensão da polaridade como o fluxo do sentimento.

Para complementar a análise quando os dados não estavam mapeados nos Lexicons utilizados, o PORIA et al. (2015) utilizaram uma nova técnica de inteligência computacional chamada máquina de aprendizagem extrema (Extreme Learning Machine - ELM), que é

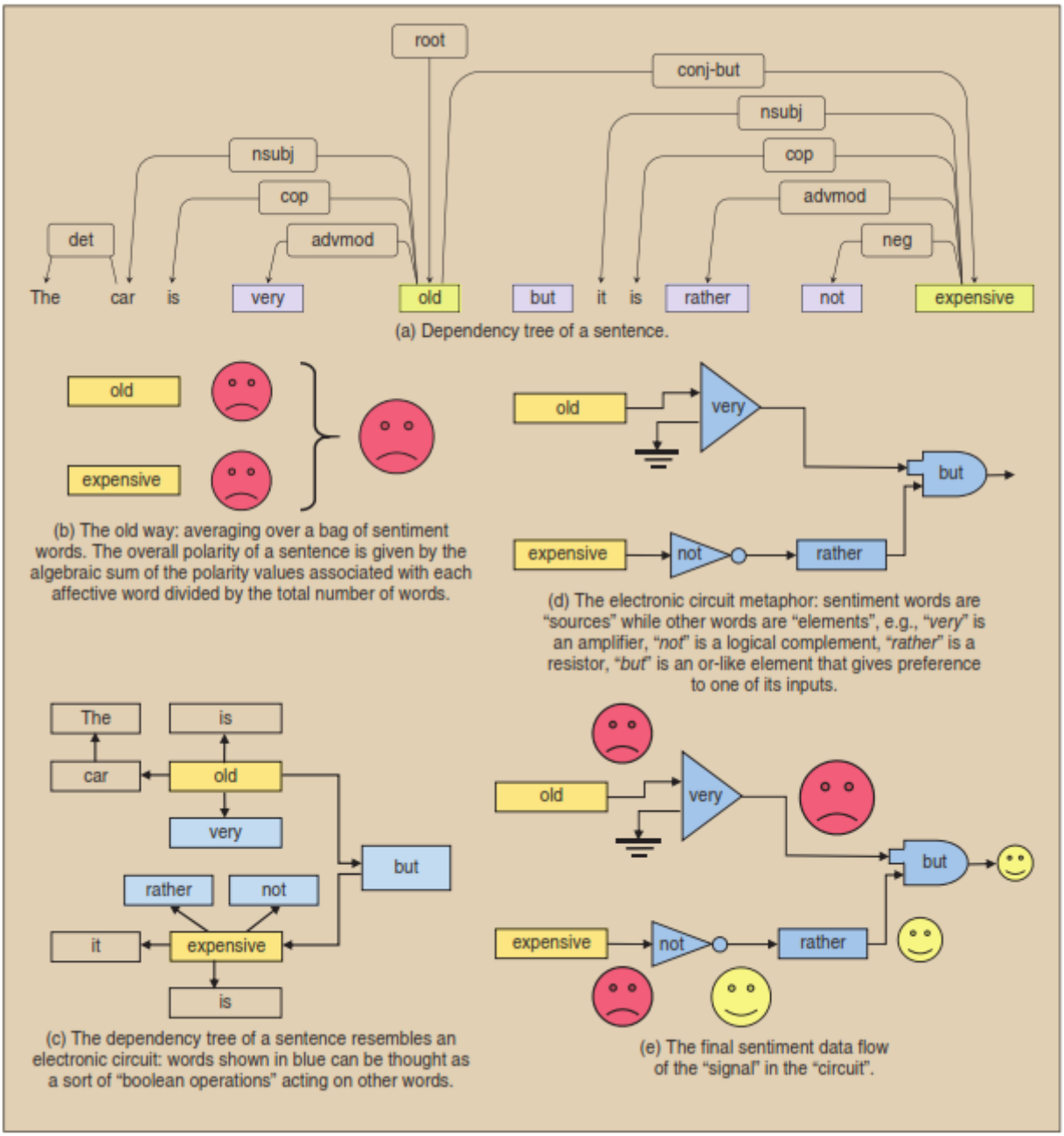

Figura 5 – A ideia principal por trás do sentic patterns: a estrutura de uma sentença é como um circuito eletrônico onde os operadores lógicos canalizam os fluxos de dados de sentimento para obter a polaridade total da sentença. (PORIA et al., 2015)

um tipo redes de feedforward de camada oculta desenvolvido desenvolvida recentemente, com uma camada oculta que não requer ajuste. O ELM superou os métodos de última geração, como o SVM (Support Vector Machines), tanto em termos de precisão como em tempo de treinamento. Nas experiências, Poria et al. (2015) obteu uma precisão global de 71,32% no conjunto de dados final usando ELM e precisão de 68,35% usando SVM. Esses dados foram obtidos a partir de análises em três tipos de conjuntos de dados diferentes.

Referências

AKAICHI, J. Social networks' facebook' statutes updates mining for sentiment classification. In: *Social Computing (SocialCom), 2013 International Conference on*. Alexandria, VA, USA: [s.n.], 2014. Citado na página 2.

ALBORNOZ, J. Carrillo de; PLAZA, L.; GERVáS, P. Sentisense: An easily scalable concept-based affective lexicon for sentiment analysis. In: *The 8th International Conference on Language Resources and Evaluation (LREC 2012)*. [S.l.: s.n.], 2012. Citado na página 11.

BISIO, F. et al. A practical guide to sentiment analysis. In: _____. [S.l.: s.n.], 2017. cap. Concept-Level Sentiment Analysis with SenticNet, p. 173–189. Citado na página 15.

BISIO, F.; ONETOA, L.; CAMBRIA, E. Sentiment analysis in social networks. In: _____. [S.l.]: Elsevier Inc, 2016. cap. Sentic Computing for Social Network Analysis. Citado na página 15.

BLEIGH, M.; AGALLOCO, S.; MICHAELS-OBER, E. *Tweetstream*. <https://github.com/tweetstream/tweetstream>. (Acessado em; 05/04/2017). Citado na página 4.

BRYNIELSSON, J.; JOHANSSON, F.; WESTLING, A. Learning to classify emotional content in crisis-related tweets. *Institute for Scientific Information*, v. 13, p. 33–38, 2013. Citado na página 2.

BUTLER, D. When google got flu wrong. *NATURE*, v. 494, p. 155–156, 2013. Citado na página 12.

CAMBRIA, E.; HUSSAIN, A. *Sentic Computing - A Common-Sense-Based Framework for Concept-Level Sentiment Analysis*. [S.l.]: Springer International Publishing, 2015. Citado 4 vezes nas páginas 13, 14, 15 e 16.

CAMBRIA, E.; OLSHER, D.; RAJAGOPAL, D. Senticnet 3: A common and common-sense sense knowledge base for cognition-driven sentiment analysis. In: . [S.l.: s.n.], 2014. Citado na página 19.

DESHWAL, A.; SHARMA, S. K. Twitter sentiment analysis using various classification algorithms. In: *Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2016 5th International Conference on*. Noida, India: IEEE, 2016. Citado na página 15.

HALL, M. et al. The weka data mining software: An update. *SIGKDD Explorations*, v. 11, n. 1, 2009. Citado na página 4.

HENDLER, J.; MULVEHILL, A. M. *Social Machines: The Coming Collision of Artificial Intelligence, Social Networking, and Humanity*. [S.l.]: Apress, 2016. Citado 2 vezes nas páginas 12 e 13.

HIGH, R. *The Era of Cognitive Systems: An Inside Look at IBM Watson and How it Works*. [S.l.], 2012. Disponível em: <http://www.redbooks.ibm.com/redpapers/pdfs/redp4955.pdf>. Citado na página 6.

KUMAR, A. S. *Knowledge Discovery Practices and Emerging Applications of Data Mining: Trends and New Domains.* [S.l.]: IGI Global, 2010. Citado 2 vezes nas páginas 10 e 12.

KURDI, M. Z. *Natural Language Processing and Computational Linguistics*. [S.l.]: John Wiley & Sons, 2016. Citado 3 vezes nas páginas 7, 8 e 9.

LIU, B. *Sentiment Analysis and Opinion Mining*. [S.l.]: Morgan & Claypool, 2012. Citado 2 vezes nas páginas 10 e 11.

LIU, B. et al. *Sentiment Analysis in Social Networks*. [S.l.]: Morgan Kaufmann, 2016. Citado na página 11.

MATHUR, I.; AL et. *Natural Language Processing: Python and NLTK*. [S.l.: s.n.], 2016. Citado 4 vezes nas páginas 7, 9, 10 e 11.

MUTHUTANTRIGE, S. R.; WEERASINGHE, A. Sentiment analysis in twitter messages using constrained and unconstrained data categories. *IEEE*, p. 304–310, 2016. Citado na página 2.

PORIA, S. et al. Sentiment data flow analysis by means of dynamic linguistic patterns. *IEEE Computational intelligence magazine*, v. 15, p. 1556–603, 2015. Citado 3 vezes nas páginas 2, 19 e 20.

PORIA, S. et al. Sentic patterns: Dependency - based rules for concept-level sentiment analysis. *Knowl.-Based Syst.*, 2014. Citado na página 19.

ROSENBERG, D.; STEPHENS, M. *Use Case Driven Object Modeling with UML*: Theory and practice. Nova Iorque: Apress, 2007. Citado 2 vezes nas páginas 17 e 18.

RUSSEL, S.; NORVIG, P. *Inteligência Artifical*. 2. ed. [S.l.]: Elsevier, 2013. Citado 2 vezes nas páginas 6 e 8.

SHARMA, Y.; MITTAL, E.; GARG, M. *Political Opinion Mining from Twitter. International Journal of Information Systems in the Service Sector (IJISSS)*. [S.l.: s.n.], 2016. Citado na página 12.

SOBRE o Twitter. <https://about.twitter.com/pt/company>. (Accessed on 04/10/2017). Citado na página 16.

TSUGAWA, S. et al. On estimating depressive tendencies of twitter users utilizing their tweet data. *IEEE Virtual Reality*, 2013. Citado na página 18.

TURING, A. Computing machinery and intelligence. *Mind*, v. 59, p. 433–460, 1950. Citado na página 7.

TWITTER Usage Statistics - Internet Live Stats. <http://www.internetlivestats.com/twitter-statistics/>. (Accessed on 04/10/2017). Citado na página 16.

VASUDEVAN, V. et al. Is twitter a good enough social sensor for sports tv? In: *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2013 IEEE International Conference on*. San Diego, CA, USA: IEEE, 2013. Citado na página 16.