

UNIVERSIDADE FEDERAL DO PIAUÍ – UFPI
CENTRO DE CIÊNCIAS DA EDUCAÇÃO - CCE
PROGRAMA DE PÓS-GRADUAÇÃO EM COMUNICAÇÃO – PPGCOM

PEDRO ALEXANDRE CABRAL

**O USO DO DATA MINING NA DESCOBERTA DE MARCAS IDENTITÁRIAS DO
PIAUÍ: UM ESTUDO DE CASO DA REDE SOCIAL TWITTER**

TERESINA – PI

2015

UNIVERSIDADE FEDERAL DO PIAUÍ – UFPI
CENTRO DE CIÊNCIAS DA EDUCAÇÃO - CCE
PROGRAMA DE PÓS-GRADUAÇÃO EM COMUNICAÇÃO – PPGCOM

**O USO DO DATAMINING NA DESCOBERTA DE MARCAS IDENTITÁRIAS DO
PIAUÍ: UM ESTUDO DE CASO DA REDE SOCIAL TWITTER**

Dissertação apresentada por Pedro Alexandre Cabral ao Programa de Pós-Graduação em Comunicação, do Centro de Ciências da Educação, da Universidade Federal do Piauí, como requisito para a obtenção do grau de Mestre em Comunicação. O presente trabalho foi elaborado sob a orientação da Prof. Dr. Gustavo Fortes Said.

TERESINA – PI

2015

RESUMO

O armazenamento da informação é característica intrínseca ao ser humano. Desde as eras mais remotas o homem sempre se pautou pela eminente necessidade de registrar dados. Entretanto, nos dias atuais, com a profusão da internet, essa quantidade de informações alcançou um crescimento exponencial. Nesse contexto, o presente trabalho tem como objetivo compreender como as técnicas de *data mining* podem ajudar na análise da identidade cultural do Piauí no Twitter. Para tanto, será realizada uma contextualização acerca da “sociedade na era do big data”, de forma especial nos cenários que tangenciam a orquestração de criação de dados a partir das redes sociais. Além disso, será realizada uma incursão epistemológica acerca do tema, pontuando a aplicabilidade do *data mining* em meio a esses massivos volumes de dados e, assim, trazer à baila uma proposta transmetodológica, uma vez que, dentro do nosso entendimento, os métodos tradicionais de pesquisa em comunicação já não conseguem deslindar os questionamentos trazidos na análise de grande volumes de dados. Neste sentido, foram analisados 671.366 tweets realizados no período de 01.01.2014 a 31.12.2014 por usuários que pontuaram, na sua descrição no Twitter, que eram do Estado do Piauí. No que se refere aos aspectos metodológicos, nos inspiramos na visão de teóricos da computação para referendar o processo de *data mining*, sem, contudo, dissociar do no nosso lugar de produção científica, que é a Comunicação.

Palavras-Chave: Data mining. Identidade. Twitter. Redes sociais

ABSTRACT

The storage of information is intrinsic characteristic of human being. Since the earliest times man has always been guided by the perceived need to record data. However, nowadays, with the profusion of internet, this amount of information reached exponential growth. In this context, this study aims to understand how data mining techniques can help in the analysis of the cultural identity of Piauí on Twitter. Therefore, a contextualization about "society in the era of big data" will be held in a special way in the scenarios that are tangent the orchestration of data creation from the social networks also an epistemological foray will be held on the topic, pointing the applicability of data mining in the midst of these massive data volumes and thus bring up one trans-methodological proposal, since, in our understanding, the traditional methods of communication research can no longer unravel the questions brought on the analysis of large data volumes. In this sense, we analyzed 671,366 tweets carried out from 01.01.2014 to 31.12.2014 by users who scored in his description on Twitter, which were the State of Piauí. With regard to methodological aspects we look to theoretical computer vision to endorse the data mining process, without, however, dissociate the place in our scientific production, which is the communication.

Keywords: Data mining. Identity. Twitter. Social Networks

AGRADECIMENTOS

À Deus, pelo dom da vida e pelo livre arbítrio para decidir o que julgo mais conveniente para minha fatídica trajetória na terra.

Aos meus amados pais, Lidia e João pelo exemplo e amor incondicional.

A minha irmã Kelly Vanessa pelo apoio e carinho.

Ao meu sobrinho, afilhado e herói, João Pedro, que mesmo sem entender, ainda, o que significa um mestrado me deixava estudar e entendia as minhas ausências.

Ao meu orientador, professor Dr. Gustavo Said, por me guiar nesta trajetória, sempre muito atencioso se mostrou um exemplo de mestre que quero ser quando crescer.

A Professora Dra. Jan Alyne com quem tive o prazer de conviver e que me mostrou os desafios do *data mining*.

À Thamirys Viana, pelo companheirismo e carinho ao longo deste desafio.

Ao amigo Jeferson Henrique, mente prodigiosa que muito contribuiu na criação dos algoritmos para esta pesquisa.

À Infoway empresa que investiu no meu crescimento profissional e pessoal.

À Adalton Sena e Ney Paranaguá, pelo exemplo de profissionalismo.

Ao friend the wine, Adriana, Ruthete e Vinicius que ajudaram a transformar essa trajetória mais amena e em muitos momentos um pouco mais ébria.

À Michelle Janaina que sempre tinha uma palavra de carinho para acalantar meu coração

Aos queridos companheiros de turma: Nina, Thamirys, Lisiane, Sarah, Fábio, Dani, Eveline e Islândia pessoas abnegadas que sempre me ajudaram com palavras de carinho.

E a você, caro leitor, que paga muitos impostos e que de certo modo ajudou a “custear” meu mestrado na UFPI.

LISTA DE FIGURAS

FIGURA 1 - DENSIDADE DO CELULAR BRASIL (FEV/14).....	19
FIGURA 4 – EXPLOÇÃO DE TWEETS NO 1º GOL DA COPA DO MUNDO FIFA.....	33
FIGURA 6 - A TEMPERATURA NO TWITTER.....	42
FIGURA 7 - ETAPAS DO PROCESSO <i>KNOWLEDGE DISCOVERY IN DATABASES</i>	48
FIGURA 8 - ANÁLISE DE CLUSTER FACEBOOK.....	56
FIGURA 9 - EMPREGO DA TÉCNICA DE ASSOCIAÇÃO.....	57
FIGURA 10 - FASES DATA MINING.....	59
FIGURA 11 - PROCESSO DE COLETA DE DADOS.....	61
FIGURA 12 - FERRAMENTA DE INDEXAÇÃO DE CONTEÚDO SOLR.....	64
FIGURA 13– INTERFACE WEKA.....	65
FIGURA 14 - CONTA DO USUÁRIO DO TWITTER SEM DESCRIÇÃO DA BIO.....	67
FIGURA 15 – CLOUD TAG DOS TERMOS UTILIZADOS EM TODOS OS MESES DE 2014.....	73
FIGURA 16 - TWITTE DE PERFIL ORGANIZACIONAL.....	73
FIGURA 17 - EXEMPLO DE TWITTE SOBRE TEMÁTICA NACIONAL.....	74
FIGURE 18 - TWITTE SOBRE O #CURSO.....	75
FIGURE 19 - TWITTE SOBRE O #SOSUESPI.....	76
FIGURA 20 - TWITTE SOBRE O #TEAMWILSONPI.....	77
FIGURA 21 - TWITTE SOBRE O #THEAMO.....	77
FIGURA 22 - TWITTE SOBRE O #SALVERAINHA.....	78
FIGURE 23 - TWITTE SOBRE O #DIADOPIAUI.....	79
FIGURA 24 - TWITTE SOBRE O #RIVER.....	80
FIGURA 25 - TWITTE SOBRE O #SALIPI.....	80
FIGURA 26 - EXEMPLO DE ARQUIVO ATTRIBUTE-RELATION FILE FORMAT (.ARFF).....	81

LISTA DE TABELAS

TABELA 1 - TOTAL MINUTES SPENT SOCIAL NETWORKING	20
TABELA 2 - AS PRINCIPAIS CARACTERÍSTICAS DAS ABORDAGENS PARA A ANÁLISE DE TEXTOS	51
TABELA 3 - PRINCIPAIS FERRAMENTAS DE DATA MINING DISPONÍVEIS NO MERCADO	58
TABELA 4 QUANTIDADE DE SEGUIDORES POR PROFISSÃO	68
TABELA 5 - QUANTIDADE DE ADESÃO DE USUÁRIOS NA REDE SOCIAL TWITTER, POR TEMPO	71
TABELA 6 QUANTIDADE DE HASHTAG, POR MÊS.....	72

SUMÁRIO

1. INTRODUÇÃO	9
2. O USO DA TECNOLOGIA EM PESQUISA EM COMUNICAÇÃO: UMA PROPOSTA TRANSMETODOLÓGICA PARA TRABALHAR COM GRANDE VOLUME DE DADOS	14
3. INTERNET: CONCEITOS E NOVAS FORMAS DE SOCIABILIDADE.....	18
2.1 COMUNICAÇÃO MEDIADA POR COMPUTADOR E AS COMUNIDADES VIRTUAIS	24
2.2 REDES SOCIAIS.....	27
2.3 TWITTER.....	31
4. IDENTIDADE COLETIVA DENTRO DO CIBERESPAÇO.....	35
4.1 A IDENTIDADE COLETIVA NAS REDES SOCIAIS	38
4.2 AS DIFICULDADES METODOLÓGICAS PARA PESQUISA DE IDENTIDADE COLETIVA NAS REDES SOCIAIS	43
5. KNOWLEDGE DISCOVERY IN DATABASES	47
5.1 SELEÇÃO.....	48
5.2 PRÉ-PROCESSAMENTO	48
5.3 FORMATAÇÃO	49
5.4 MINERAÇÃO DE DADOS (DATA MINING)	49
5.5 INTERPRETAÇÃO E ANÁLISE	49
5.6 TIPOS DE ABORDAGENS DE DADOS.....	49
5.7 ANÁLISE SEMÂNTICA	50
5.8 ANÁLISE ESTATÍSTICA.....	50
6. MÉTODOS DE DATA MINING.....	52
6.1 CLASSIFICAÇÃO	54
6.2 ANÁLISE DE CLUSTERING	55
6.3 ASSOCIAÇÕES	56
7. METODOLOGIA.....	59
7.1 COLETA	59
7.2 PRÉ-PROCESSAMENTO	61
7.3 INDEXAÇÃO.....	62
7.4 MINERAÇÃO.....	64
7.5 ANÁLISE	65
8. ANÁLISE DOS RESULTADOS	67
9. CONSIDERAÇÕES FINAIS.....	85
REFERÊNCIAS.....	88

1. INTRODUÇÃO

O armazenamento da informação é característica intrínseca ao ser humano. Ao analisar a história, percebemos que desde os primórdios o homem carece da necessidade de guardar informações: figuras rupestres em cavernas ou escritas em paredes de pedras encravadas nas pirâmides egípcias são apenas alguns exemplos. Isto também é visto nas culturas que somente dominam a linguagem oral, na busca de perpetuar o antigo por intermédio da oralidade (LIMA JUNIOR, 2011).

Hoje, a penetração da internet e das tecnologias digitais no cotidiano do cidadão comum alavancou incrível mobilidade e ubiquidade comunicacional e informacional entre os indivíduos, não mais apenas no nível das organizações, catalisando assim tanto o controle e a transparência, quanto as possibilidades de auto exposição em níveis inéditos na nossa história (GABRIEL, 2010).

O valor mais acessível de máquinas computacionais (processamento e memória) e dos dispositivos de captura e armazenagem de dados (sensores, câmeras fotográficas e de vídeo, celulares, pen-drives, *flash memory*, discos rígidos externos, etc.) criaram inimaginável quantidade de dados, que estão sendo disponibilizados na Web, proporcionando a formação da “Era do Big Data” (LIMA JUNIOR 2011).

No trabalho intitulado de Big Data, Jornalismo Computacional e Data Journalism: estrutura, pensamento e prática profissional na Web de dados, Lima Junior (2012, p. 211) define big data (BD) como sendo:

[...] conjunto de dados (dataset) cujo tamanho está além da habilidade de ferramentas típicas de banco de dados em capturar, gerenciar e analisar. A definição é intencionalmente subjetiva e incorpora uma definição que se move como um grande conjunto de dados necessita ser para ser considerado um big data (LIMA JUNIOR, 2012, p. 211).

Como o autor propõe, não existe uma definição precisa acerca do big data, entretanto autores como Zikopoulos et alii (2012) advogam que BD caracteriza-se pela presença de quatro aspectos: volume, velocidade, variedade e veracidade. Volume refere-se, como o próprio nome sugere, a quantidade de dados disponível na internet e que nos últimos anos

vem crescendo de forma exponencial. A velocidade diz respeito à rapidez com que os dados podem ser capturados e processados. Variedade, por sua vez, pauta-se nas diversas fontes de dados em que estas informações podem ser encontradas e, o último aspecto, veracidade abrange com característica o fato que os dados não apresentam uma verdade absoluta. Ou seja, certa incerteza onde se deve observá-los com muita parcimônia para que os mesmos possam gerar informações úteis e oportunas.

Isto posto, o massivo crescimento de dados na internet traz consigo grandes óbices no que tange a disponibilização da informação. Ao digitarmos a palavra “identidade” no Google, por exemplo, encontramos aproximadamente 13.500.000 resultados¹, dificultando o processo de seleção e interpretação das informações por parte do usuário. Outro grande problema, segundo Sprink, Wolfram et alii (2001) é a dificuldade encontrada pelos usuários em expressar a necessidade de informação por meio de palavras-chave, visto que aproximadamente cinquenta e dois por cento das buscas realizadas nas máquinas são reformuladas.

Dentro dessa nova perspectiva de utilização das bases de dados, Machado e Palacios (2007), em estudo sobre as competências dos profissionais de Comunicação, citam a existência de pesquisas que tratam sobre a prática dos profissionais da área, em especial a necessidade destes se adaptarem às novas exigências do mercado, tendo o domínio dos processos de digitalização da informação.

Dentre as competências digitais compiladas no estudo citado, eis algumas: uso básico do computador como ferramenta para busca, avaliação e classificação de informações; “cultura de internet” (MACHADO e PALACIOS, 2007, p.79); conhecimentos básicos e utilização de programas de edição de texto, tratamento de imagem, áudio, programação visual; conhecimento teórico sobre redes e seu funcionamento e alta capacidade de aprendizagem de uso de novos programas (MACHADO e PALACIOS, 2007, p.79).

Para Lima Junior (2011), a capacidade e facilidade em reunir e armazenar informações em banco de dados, assim como sua utilização, cresce a cada dia e na mesma proporção que novas tecnologias são desenvolvidas e propagadas para facilitar o trabalho do consumidor. Com a popularização da rede mundial de computadores, quase todo e qualquer conteúdo produzido passa a ser colocado em espaço considerado até então infinito, reflexo também da

¹ Pesquisa realizada em 05/07/2014.

atualização de produtos comunicacionais palpáveis, como revistas, jornais e principalmente livros.

Nesse contexto, organizações de toda ordem, bem como usuários comuns, estão migrando seus conteúdos para os discos rígidos dos computadores. Logo, a atividade em reunir a maior quantidade de informações e disponibilizar de forma organizada, simples e objetiva, começa a se tornar uma tarefa obrigatória no mundo da comunicação, mas concomitantemente árdua e repleta de ruído. O modo mais clássico de armazenamento de informação é através da palavra escrita, impressa. O acesso à informação estocada dessa forma é lento, difícil e de pouco rendimento. Para todas as etapas da manipulação da informação é necessária a presença do ser humano, e suas limitações na capacidade de aquisição de dados e processamento de grande volume constituem o principal gargalo do processo (MANDEL, SIMON, & DELYRA, 1997).

Inobstante a este excesso de dados, a quantidade de informações produzidas encontra-se, em sua maioria disposta em base de dados não estruturadas, ou seja, base de documentos textuais, cujo formato está adequado ao homem que, somente através da leitura, é capaz de decodificar a informação contida no texto e aprendê-la (SCHIESSL, 2007). Frente a este contexto de enormes mananciais de dados, a fim de prover fluidez e agilidade no manuseio da informação, existe uma necessidade premente de associar novas tecnologias ao contexto atual em que, em nosso entendimento, as pesquisas atuais em comunicação já não conseguem resolver suas inquietações em face deste grande volume de dados. Assim sendo, sugere-se que exista uma interdisciplinaridade maior com outras áreas, sobretudo as que advêm das ciências da computação.

Em face desta conjuntura, o presente trabalho justifica-se pela necessidade de extrair informações até então desconhecidas, utilizando como base técnicas de *data mining* e assim melhorar a qualidade da informação pesquisada em bancos de dados e na obtenção de relações ‘invisíveis’ de temas e contextos que compõem a formação da identidade cultural do Piauí no Twitter. Deste modo, a presente pesquisa traz em seu bojo aspectos para criação de uma interdisciplinaridade e, de forma mais acentuada, uma proposta de transmetodologia, uma vez que para a sua execução serão utilizados métodos que não são oriundos do campo da comunicação, pois acreditamos que as abordagens atuais já não conseguem resolver as inquietações trazidas neste trabalho, no que diz respeito a trabalhar com grandes volumes de dados.

Ademais, a problemática aqui apresentada é ainda pouco explorada no contexto acadêmico dos estudos sobre identidade. A partir de então, há possibilidades de estudos inovadores sobre descoberta de identidade através da mineração de dados.

Do apresentado, esta pesquisa inclina-se a compreender a utilização do data mining com o propósito de analisar a incidência das principais palavras-chave bem como os elementos, temas e categorias que compõem a construção da identidade do Piauí no Twitter e assim tentar responder o seguinte questionamento: Como o *data mining* é útil para analisar a identidade cultural do Piauí no Twitter?

A partir dessas informações iniciais, colhidas através de técnicas de data mining, vislumbra-se montar um quadro onde seja possível evidenciar quais as palavras e temáticas abordadas no twitter pelos piauienses, durante todo o ano de 2014. Para detectar a formação de padrões nos twittes veiculados, utilizaremos a ferramenta Solr e um algoritmo de mineração desenvolvido, especificamente, para este fim.

No que concerne aos aspectos metodológicos desenvolvidos neste trabalho, partimos de uma proposta transmetodológica, em que buscou-se trazer conhecimentos nativos de outras áreas a fim de verificar a eficácia de tais métodos para a realização de pesquisas na área da comunicação.

Para tanto, partimos da análise de 671.366 twittes coletados durante o período de janeiro a dezembro de 2014. Estes twittes, por sua vez, foram coletados a partir dos usuários que colocaram em suas respectivas descrições que eram do estado Piauí. Por considerar mais oportuno para a realização desta pesquisa, optamos por aqueles usuários que tinham uma quantidade de seguidores acima de 500.

A presente dissertação é composta por três capítulos, os quais apresentam uma visão acerca dos mecanismos basilares para se trabalhar com grandes volumes de dados tomando como vertente o estudo de identidade nas redes sociais.

No primeiro capítulo, realizou-se uma incursão sobre o uso da tecnologia em pesquisa em comunicação tendo como base a realização de uma proposta transmetodológica para trabalhar com grande volume de dados gerados a partir das interações dentro das redes sociais. Além disso, foi suscitado o comportamento da sociedade atual tendo como prerrogativas a internet como agente catalisador de novas formas de sociabilidade. Ainda no primeiro capítulo, pontuou-se a relação entre comunicação mediada por computador e as

comunidades virtuais e para finalizar foi apresentada uma contextualização sobre redes sociais dando um enfoque mais significativo para o twitter.

O segundo capítulo teve como temática norteadora o entendimento sobre identidade, entretanto o enfoque principal foi relacionado às questões que perpassam o ciberespaço e, de forma mais especial, os contornos criados dentro da interatividade gerada pelas redes sociais. Os aspectos relacionados às dificuldades metodológicas para se realizar pesquisas de identidade nas redes sociais também foram pontuados neste capítulo.

Knowledge Discovery in Databases, ou descoberta de conhecimento em base de dados foi a tônica do terceiro capítulo. Neste sentido, apresentou-se detalhadamente quais os processos necessários para se trabalhar com grandes volumes de dados, o referido capítulo buscou por descrever de forma sistemática como é construída cada etapa e associado a isto como as mesmas se relacionam. Além disso, pontuou-se um quadro resumo com as principais ferramentas disponíveis para realização de data mining.

2. O USO DA TECNOLOGIA EM PESQUISA EM COMUNICAÇÃO: UMA PROPOSTA TRANSMETODOLÓGICA PARA TRABALHAR COM GRANDE VOLUME DE DADOS

A epistemologia do campo da comunicação é interdisciplinar desde a sua gênese, pois para sua concepção foi necessária a junção de várias manifestações, como, por exemplo, publicidade e propaganda, marketing, telejornalismo, fotografia, relações públicas, comunicação digital dentre outros (GRINS, 2008.) Sendo assim, temos que as pesquisas em comunicação não podem ser pautadas em um hermetismo epistemológico, uma vez que a cada dia o campo da comunicação vem se expandindo, principalmente com os avanços tecnológicos. Em face desta expansão e avanço é latente a necessidade de uma transmetodologia cada vez mais presente, onde se tem métodos e práticas que não necessariamente tiveram sua gênese na comunicação.

Para Grins (2008, p 107), a utilização da transmetodologia funciona como “um meio essencial na busca por uma melhor compreensão da problemática, por atravessar as disciplinas, pelas disciplinas, e as transcender na busca de um entendimento mais amplo e concreto”. A autora afirma que a utilização de métodos de outras áreas propõe uma “quebra de paradigmas, não apenas confluindo teorias, metodologias e disciplinas, mas questionando, desconstruindo/reconstruindo e refletindo as mesmas na busca da compreensão do complexo campo da comunicação” (GRINS, 2007, p. 107).

Como dito anteriormente, temos um desenvolvimento tecnológico cada vez mais pujante, o qual passou a imprimir desafios às rotinas das pesquisas em comunicação. Neste cenário, onde temos um campo comunicacional relativamente recente e em formação também não é diferente. Todos os dias surgem novas tecnologias comunicacionais e para entender o seu funcionamento faz-se necessário fomentar o desenvolvimento de estudos interdisciplinares, a partir do envolvimento de outros saberes. Defendemos que essa conexão deve ocorrer de forma sinérgica e, sobretudo, sem perpassar o processo comunicacional em si, a tecnologia aparece apenas como suporte e não como objeto principal.

Seguindo essa linha de raciocínio, Werthein (2004, p. 3) traz um pensamento que traduz de forma sucinta como a tecnologia está intimamente ligada, não só às rotinas profissionais do comunicador, mas também às de todas as pessoas. A saber:

A era da nova sociedade informatizada, que trabalha quase que em tempo real, nos coloca muitos desafios. Traz um tipo diferente de viver: um viver instantâneo que afeta profundamente as relações entre as pessoas. E, em especial, a relação entre o serviço público e os cidadãos. É isso que estamos vivendo hoje e, fatalmente, uma relação diferente nos será apresentada amanhã, tamanha é a velocidade com que os avanços tecnológicos acontecem. Nosso principal desafio é não apenas conviver bem com a era atual, e tudo o que a tecnologia já nos oferece - o que já é muito! -, mas também mantermos uma postura aberta e um olhar de aprendiz a tudo o que é novo que nos surge a cada instante. Porque é esse novo que, ao longo do tempo, afeta nossos parâmetros, nossos conceitos, enfim, toda a nossa vida e nosso relacionamento com os outros. É a permanente mutação da realidade individual e coletiva no planeta (WERTHEIN, 2004, p.3).

A partir desta perspectiva, é válido trazer para este trabalho as discussões em torno da interdisciplinaridade, visto que a essência das discussões aqui presentes é a inserção de novos recursos tecnológicos, em especial da área da informática, com o intuito de maximizar o rendimento do trabalho dos pesquisadores em comunicação.

Dentro deste bojo de interdisciplinaridade, a introdução das novas tecnologias da informação e da comunicação trouxe consigo novas perspectivas para os processos comunicacionais. Tal mudança, por sua vez, vem modificando certas convicções até então defendidas por vários teóricos da comunicação que, muitas vezes, ainda continuam presos a realidades ultrapassadas, as quais não condizem com os novos rumos que vêm sendo traçados pelo campo comunicacional. Nesse sentido, Signates (2012, p.32) defende a necessidade de explicar as abordagens comunicacionais tendo em vista os novos recursos oriundos do surgimento da internet e seus correlatos.

Essa emergência exigente da realidade comunicacional do mundo é o que, a nosso ver, principalmente exige que a abordagem comunicacional seja reexplicada, á luz de seus próprios termos. Ao tornar-se uma condição central para a vida, em seus variados aspectos, a comunicação exige seu lugar de objeto científico prioritário. (SIGNATES, 2012, p.32)

Nesse contexto, o aparecimento dessas novas tecnologias no âmbito da comunicação vem a contribuir com a soma de novos objetos de estudos, que vão se corporificar pelo viés de estudo dos meios de comunicação ou ainda dos processos comunicativos, apontados por França (2001) como sendo dois objetos próprios da área da Comunicação. Tal profusão de

saberes vai ser caracterizada por Martino (2004) como "diversidade de campo", refletindo as próprias complexidades existentes no campo comunicacional e no desenvolvimento de seus processos. A saber,

[...] é um efeito da complexidade dos processos; mas também do jogo das forças macro/micro políticas que se encontram sujeitas suas instituições; da heterogeneidade das lógicas sociais de seus agentes, de efeitos estruturais de sua organização burocrática, etc. E temo que os níveis de apreensão podem acabar sendo tão diversos quanto a diversidade encontrada em cada um deles (MARTINO, 2004, s/p).

Num contexto de busca pelo progresso científico, bem como de uma pretensa revolução tecnológica social, Maigret (2010) defende que a Comunicação passou a ser um terreno propício para o desenvolvimento de ideologias como o pós-modernismo e o tecnicismo proteico. Partindo desse pressuposto, o autor relata que a internet tem sido apontada neste século como uma espécie de salvação. Contudo, este defende que tal pensamento se configura como utopia daqueles que querem instaurar uma espécie de comunidade mundial, pautada em princípios como a liberdade, a instantaneidade e o regime de trocas sem fronteiras. Em outras palavras, os defensores dessa ideia veem na adoção das novas tecnologias uma forma possível de manter contato entre todos, por meio do fim das hierarquias e o advento do desenvolvimento da inteligência coletiva.

Por outro lado, também é pertinente ressaltar que, tanto a internet como as novas tecnologias da comunicação e da informação podem se configurar

[...] numa ideologia ingênua de progresso [...] com o contraponto das críticas sobre os perigos possíveis da pornografia em livre acesso, da violação das liberdades fundamentais pelo fichamento dos indivíduos, do risco de pulverização de uma sociedade que se comporia somente de indivíduos conectados, críticas que inquietam as camadas populares (MAIGRET, 2010, p. 405).

A internet e seus correlatos são vistos hoje como espécie de supermídia, de caráter agregador e, por isso, "superior" aos demais meios de comunicação tradicionais. No entanto, apesar de reunir em uma mesma plataforma todas as funcionalidades disponibilizadas em separado pelas demais mídias, Maigret (2010) traz a ideia de que não há motivo para tanto

entusiasmo com as funcionalidades oferecidas pela Web. Isto porque, se atentarmos para o grau de penetração desta mídia perante a sociedade como um todo, percebe-se ainda que a maior parte da população não tem acesso a ela ou não compreende de forma satisfatória as suas funcionalidades, além da incidência de altas taxas de analfabetismo e os custos de sua utilização, inviabilizando, pelo menos por hora, a sua disseminação. Tecendo várias críticas ao segmento, o autor ressalta que:

A internet é uma maravilhosa ferramenta de compartilhamento e de armazenamento de dados, mas não possui nenhuma virtude superior que explicaria o advento de um novo pensamento universal. Ela não fornece as chaves da compreensão dos dados que contém, assim como nenhuma fonte de informação jamais pôde abrir mão da interpretação e de reenquadramento contextual (MAIGRET, 2010, p. 410).

Em outras palavras, a internet assume um lugar entre os meios de comunicação de massa, mas sem excluir seus predecessores ou mesmo substituí-los. Partindo desse marco, há uma crescente corrente de autores que argumentam sobre a mutação das formas de fazer pesquisa em comunicação.

3. INTERNET: CONCEITOS E NOVAS FORMAS DE SOCIABILIDADE

A sociedade contemporânea está cada vez mais imersa em redes de conexões digitais, nas quais os fluxos de informações estão quase sempre associados a uma “inócua” conversa polissêmica nos mais variados canais. A conectividade dá lugar à construção de novos artefatos tecnológicos, mas pode-se perguntar se ela está, de fato, a serviço de uma necessidade básica da existência humana: a socialização.

Mesmo antes do surgimento da internet, essas interações sociais, via tecnologia, já eram percebidas. O processo se deu com o surgimento dos meios de transporte e de comunicação (MCLUHAN, 1964). Com o uso e acesso a novos serviços e produtos tecnológicos, sobretudo os advindos da internet e, de uma forma mais fecunda as redes sociais, temos percebido a articulação de plataformas tecnológicas para formação de novas comunidades e criação de mecanismos cada vez mais interacionais, onde, possivelmente, se renovam de forma intensa e reforçam a construção de discursos polifônicos e dialógicos.

Ainda dentro deste prisma, nota-se que a evolução tecnológica vem refletindo diretamente sobre as sociedades e principalmente sobre o comportamento humano. Percebemos que houve um crescimento exponencial de Tecnologias de Informação e Comunicação (TICs) nos últimos séculos. De modo geral, todas essas tecnologias impulsionaram uma nova lógica social e cultural que, dentro desta premissa, o homem passa a ser criador e usuário das ferramentas tecnológicas, apropriando-se das possibilidades técnicas e, concomitantemente, sendo afetado por elas em todos os aspectos de sua existência (TEIXEIRA, 2011).

Em meio a todo esse crescimento tecnológico percebe-se que desde os primórdios o homem, por apresentar-se como um animal gregário, precisou estar em grupos para sobreviver e, assim, com o passar dos anos passou a utilizar cada vez mais ferramentas de tecnologias de informação e comunicação para potencializar e diversificar as diversas maneiras de se comunicar. Uma parcela considerável deste avanço pauta-se na melhoria incessante de processos que permeiam a atividade de comunicação. Hoje, a comunicação toma outra extensão, pautada pelo uso de microchips, redes de internet sem fios e pelo excesso de informação.

Atualmente, adentramos em uma “era da conexão” (WEINBERGER, 2003) onde se

flagra de forma cada vez proeminente a necessidade de uma conexão ubíqua e uso cada vez mais frequente de *gadgets* que têm como objetivo tornar a *pervasive computing*² ainda mais mandatória (LEMOS, 2002). A partir da popularização do acesso a novas tecnologias percebe-se uma difusão e de certo modo uma “democratização” no acesso a internet. Em recente pesquisa elaborada pela Nielsen, no Brasil, a classe C consome mais internet que a classe A e B, enquanto na classe A e B o consumo de médio por pessoa de páginas era no máximo 821 a classe C abriu, em média, 972 páginas no mês. Trata-se assim de transformações da práxis social e na forma de produzir e consumir informação (LEMOS, 2005).

Para Lemos (2005), a era da conexão é a era da mobilidade, onde se visualiza uma densidade *mobile*, maior inclusive que a quantidade de computadores. Dados da Anatel (Agência Nacional de Telecomunicações) revelam que o Brasil terminou o mês de março de 2014 com 273,6 milhões de celulares, o que representa uma teledensidade de 135,3 dispositivos móveis para cada 100 habitantes. Em linhas gerais, este comportamento é presente em todas as unidades da Federação. Apenas o estado do Maranhão apresenta uma densidade de celular menor que 100 aparelhos para cada 100 habitantes, conforme demonstra a figura I.

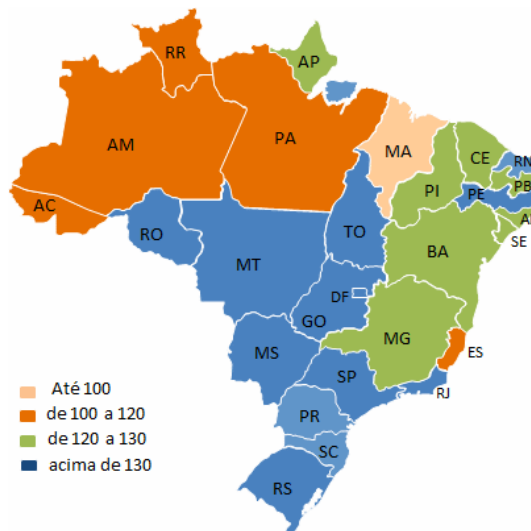


FIGURA 1 - Densidade do Celular Brasil (Fev/14)

Fonte: TELECO, 2014.

² Disseminação dos computadores em todos os lugares.

Dentro desta acepção, Lemos (2005) destaca que a popularização de celulares e das redes de acesso à internet vem proporcionando mudanças na produção e consumo de informação, uma vez que o acesso está imbricado de forma cada vez mais pervasiva, ou seja, conectividade em todos os lugares. Cooper, Green, Murtagh e Harper (2002) partem do entendimento de que a era da conexão é a era da mobilidade. Essa nova era pressupõe a ideia que os dispositivos móveis irão se tornar a principal forma de conexão à rede mundial de computadores. Em linhas gerais, percebe-se a que o acesso mobile, nos últimos anos, vem crescendo e ficando cada vez mais presente no cotidiano das pessoas, que de certo modo, geram mais dados, uma vez que o *device* está sempre ao alcance da mão.

De acordo com projeções da CISCO®, empresa estadunidense especializada em soluções em redes de comunicação e dados, o tráfego na internet via dispositivos móveis em 2017 será 13 vezes maior do que em 2012. Entretanto, é oportuno salientar que boa parte deste acesso é destinada para interação nas redes sociais. A Nielsen publicou em 2012 um estudo intitulado *State of the media*³: *the social media report* no qual flagrou um aumento considerável no tempo gasto nas redes sociais, via smartphone, em relação ao ano de 2011 e 2012 conforme tabela abaixo:

Social Networking	YOY% Change
Facebook	61%
Twitter	48%
foursquare	154%
Pinterest	6,06%

Tabela 1 - Total Minutes Spent Social Networking

Fonte: Nielsen, 2012

Dentro dessa premissa, estas redes dão lugar a um ambiente de enorme interação e geração de informação. Ao analisar, de forma pontual, o *buzz* gerado em torno da morte do CEO da Apple, Steve Jobs, a empresa australiana de monitoramento de mídia social, SR7, estima que foram gerados 10 mil *tweets* por segundo citando o ocorrido. O exemplo citado permite inferir que:

³ Disponível em <http://www.nielsen.com/content/dam/corporate/us/en/reports-downloads/2012-Reports/The-Social-Media-Report-2012.pdf> acesso em 14/05/14

Na sociedade mediatizada, as instituições, as práticas sociais e culturais articulam-se diretamente com os meios de comunicação, de tal maneira que a mídia se torna progressivamente o lugar por excelência da produção social do sentido, modificando a ontologia tradicional dos fatos sociais (SODRÉ, 1999, p. 27).

Esta modificação dá-se, sobretudo, na forma como as pessoas estão buscando e criando informação. Há algumas décadas, basicamente, só tínhamos a TV e o rádio como principais difusores de notícias. Hoje as redes sociais funcionam também como agentes de comunicação de massa. Essas novas interações, por sua vez, geram cada vez mais enormes quantidades de dados, em outro estudo realizado pela CISCO®, o tráfego na internet cresceu 45% em um ano (entre 2009 e 2010), chegando a 15 exabytes⁴ por mês, e as projeções são de que o fluxo de informação na internet em 2015 será quatro vezes maior, chegando a 767 exabytes no ano. Recentemente, o PennyStocks Lab⁵ desenvolveu um infográfico interativo, onde mostra o que está acontecendo na internet em tempo real, dentre os inúmeros *insights* a ferramenta mostra, por exemplo, que em um minuto, 27.780 posts são publicados no Tumblr, 204.166.680 são e-mails enviados, bem como 138.840 horas de conteúdo são assistidas no YouTube. Ao analisar os dados oriundos das redes sociais percebe-se que esse tráfego torna-se ainda mais expressivo, em apenas um minuto temos:

- a) WhatsApp: 11.088 contas criadas, 2.08.251.929 mensagens são enviadas;
- b) Facebook: 3.549.328 likes, 60.088.768 posts e 6.612 Gb de dados são gerados;
- c) Twitter: 6.617.700 tuítes são enviados e 12.771 novas contas.

⁴ Um exabyte corresponde a 1 bilhão de gigabytes.

⁵ Disponível em <http://pennystocks.la/> acesso em 19/05/14

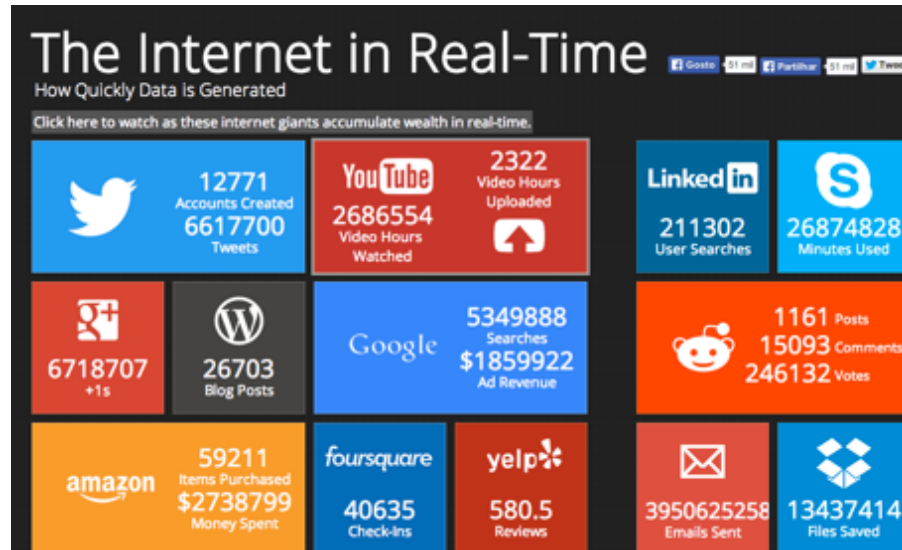


FIGURA 2 – The Internet In Real Time

Fonte: PennyStocks

É perceptível que nas últimas décadas, a tecnologia, sobretudo as utilizadas na produção e difusão de conteúdo na área da comunicação, sofreram grande inovação. “Diversas plataformas digitais conectadas ficaram mais acessíveis do ponto de vista econômico e também foram reconfiguradas para serem acessadas de forma mais amigável por profissionais e amadores” (LIMA JUNIOR, 2012).

Entretanto, é oportuno destacar que as pesquisas em comunicação outrora utilizadas já não conseguem mais resolver as inquietações trazidas por essas novas formas de “sociabilização digital”. Neste ímpeto, nasce a necessidade cada vez premente de buscar uma interdisciplinaridade com outras áreas do conhecimento, principalmente as de base tecnológica. Lima Junior (2012, p.4) parte do entendimento que

A expansão tecnológica, além de ampliar e baratear os custos das possibilidades de produção e distribuição de conteúdos digitais, também abriu novas frentes na área da pesquisa acadêmica (...). Para acompanhar todo o processo de evolução tecnológica, é premente que os pesquisadores da área da comunicação social ampliem seus ferramentais metodológicos, adaptando-os aos instrumentos de verificação que são desenvolvidos em outras áreas do conhecimento (LIMA JUNIOR, 2012, p.4).

Antes os estudos relacionados ao comportamento humano citados nas pesquisas em comunicação eram alicerçados em metodologias behavioristas, agora tem-se a possibilidade de usar aparatos tecnológicos (LIMA JUNIOR, 2012). O objeto deste estudo, por exemplo,

sem a utilização de tecnologia seria impraticável a sua execução, uma vez que existe um enorme volume de dados a ser analisado e caso optássemos por pesquisas tradicionais como a análise de conteúdo, ainda assim seria necessário desprender uma quantidade de horas infinitamente maior para realizá-los. Imaginemos analisar meio milhão de tuítes usando apenas planilhas em Excel, como exortado anteriormente, sem mecanismos específicos para realização desta tarefa, o tempo e o esforço hercúleo inviabilizaria a realização deste ou qualquer outro trabalho que se proponha a analisar um volume de dados considerável.

Dentro dessa conjectura, a ideia de utilizar tecnologia na construção de novas pesquisas em comunicação engloba cada vez mais uma importância capital no alcance dos objetos. Isto posto, tem-se percebido que a sociedade, de modo geral, tem visualizado um olhar diferente sobre temáticas tecnológicas que antes eram utilizadas por experts em computação (LIMA JUNIOR, 2012).

O pesquisador canadense em filosofia da tecnologia da escola de comunicação Simon Fraser University, Andrew Feenberg, pontua no seu estudo *What Is Philosophy of Technology*⁶ que o olhar da sociedade está cada vez imbricado em assuntos de cunho tecnológicos:

The public sphere appears to be opening slowly to encompass technical issues that were formerly viewed as the exclusive preserve of experts. Can this trend continue to the point where citizenship will involve the exercise of human control over the technical framework of our lives? We must hope so for the alternative appears to be certain destruction. Of course the problems are not only technological. Democracy is in bad shape today on all fronts, but no one has come up with a better alternative. If people are able to conceive and pursue their intrinsic interest in peace and fulfillment through the political process, they will inevitably address the question of technology along with many other questions that hang in suspense today. We can only hope this will happen sooner rather than later⁷ (FEENBERG, 2003, online).

Dessa forma, o pressuposto de um “pensamento computacional” defendido por Jeannete M. Wing no qual pontua como temática principal a resolubilidade trazida pelos métodos

⁶ Disponível em <http://www.sfu.ca/~andrewf/komaba.htm> acesso em 19/06/2014,

⁷ A esfera pública parece estar se abrindo lentamente para abranger assuntos técnicos que antes eram vistos como da esfera exclusiva dos especialistas. Pode esta tendência continuar a ponto de a cidadania envolver o exercício de controle humano sobre a estrutura técnica de nossas vidas? Esperemos que sim, pois a alternativa parece levar a certa destruição. Naturalmente os problemas não são apenas tecnológicos. A democracia está em má forma hoje em todas as frentes, mas ninguém propôs uma alternativa melhor. Se as pessoas puderem conceber e perseguir seus interesses intrínsecos em paz e plenitude por meio do processo político, elas inevitavelmente abordarão a questão da tecnologia, juntamente com muitas outras questões que hoje estão em suspense. Restam apenas esperar que isso aconteça mais cedo que mais tarde (FEENBERG, 2003; tradução nossa).

computacionais referenda nosso entendimento acerca da utilização de tecnologia para resolver problemas de pesquisas em comunicação, já que esta “nova sociedade” caracteriza-se pela utilização e produção de grandes volumes de dados e sem algoritmos e métodos de pesquisa específicos para esse fim seria praticamente impossível resolvê-los, a saber:

Métodos e modelos computacionais nos encorajam a resolver problemas e desenhar sistemas que nenhum de nós seria capaz de desenvolver sozinho. O pensamento computacional envolve resolver problemas, desenhar sistemas e entender o comportamento humano, inspirados nos conceitos fundamentais das Ciências da Computação (WING, 2006, p. 33).

2.1 Comunicação mediada por computador e as comunidades virtuais

A comunicação mediada por computador – CMC - trouxe consigo diversas modificações no contexto sócio cultural. Diante do exposto,

A Comunicação Mediada por Computador está afetando a sociedade e influenciando a vida das pessoas e a noção de comunidade. Por isso, muitos autores optaram por definir as novas comunidades, surgidas no seio da CMC por “comunidades virtuais” (RECUERO, 2009, p.32).

É latente que a CMC propiciou diversas mudanças no cotidiano das pessoas, a criação de redes de relacionamentos cada vez mais dinâmicas, orquestradas prioritariamente pelo contexto digital, no qual se percebe mais uma vez o desprendimento do real em detrimento do virtual. De modo geral, as pessoas estão passando mais tempo conectadas, utilizando assim, espaços de conversas online e buscando estabelecer laços que outrora era impossibilitado em virtude das grandes distancias. Para tanto, basta analisar o tempo médio despendido nestas redes. Em recente reportagem veiculada no portal olhar digital⁸, os brasileiros passam em média 12 horas por mês no Facebook, isso equivale a 1,68% do mês, este tempo corrobora com a aceção do termo “comunidade virtual” que possui como um dos seus principais percussores o autor Rheingold (1998, p.20). Para ele, este termo pode ser definido como:

⁸ Disponível em <http://olhardigital.uol.com.br/noticia/40875/40875> acesso em 14/05/14

As comunidades virtuais são agregados sociais que surgem da Rede [Internet], quando uma quantidade suficiente de gente leva adiante essas discussões públicas durante um tempo suficiente, com suficientes sentimentos humanos, para formar redes de relações pessoais no espaço cibernético [ciberespaço] (Ibidem).

Levy (2007) entende que para existir uma comunidade virtual precede a ideia de construção de afinidades independentemente das proximidades geográficas. Ou seja, uma desterritorização, onde não importa de onde o sujeito esteja, pois para existir uma comunidade virtual basta que se tenha processo de cooperação ou trocas entre os usuários. Esse ciberespaço, na acepção de Levy (2010), não pode ser construído apenas pela rede mundial de computadores, mas sim pelo agrupamento de “redes independentes de empresas, de associações, de universidades, sem esquecer as mídias clássicas (bibliotecas, museus, jornais, televisão etc.)” (LÉVY, 2010, p.128). Entretanto o www pode ser considerado um arquétipo de “construção cooperativa internacional, a expressão técnica de um movimento que começou por baixo, constantemente alimentado por uma multiplicidade de iniciativas locais” (LÉVY, 2010, p. 128).

Ainda de acordo com as conjecturas postuladas por Rheingold (1998), percebemos como agentes formadores desta comunidade virtual essas discussões públicas, que dentro deste âmbito configuram encontros, articulações, interações e acima de tudo aspectos ligados à linguagem, onde a tecnologia, sobretudo a internet, funciona como plataforma que subsidia toda essa articulação para formação destas comunidades. Entretanto é oportuno salientar que as redes sociais complexas já existiam, para além dos dispositivos tecnológicos e conformidade com a atividade de produção e uso de linguagem. A visão de Wellman (1999, p.2) aponta que:

Redes sociais complexas sempre existiram, mas os desenvolvimentos tecnológicos recentes permitiram sua emergência como uma forma dominante de organização social. Exatamente como uma rede de computadores conecta máquinas, uma rede social conecta pessoas, instituições e suporta redes sociais. (Ibidem).

Como pontuado anteriormente, o homem sempre careceu se comunicar, entretanto, para que essa comunicação seja exteriorizada é latente a presença de meios que possibilitem a articulação entre os diversos sujeitos envolvidos neste processo. Neste novo cenário, a internet flagrou-se como principal meio articulador para esta forma de comunicar mediada

pelas tecnologias. Pierre Lévy (2000) pontua que a internet vem trazer em seu bojo uma desterritorialização e uma virtualização, nas quais se percebe uma sociedade caracterizada pela velocidade, multimodalidade e universalidade.

Ainda sobre essa ótica, Lévy (2008) analisa como a tecnologia vem sendo utilizada ao longo dos anos, sem se restringir meramente a artefatos técnicos.

A presente mutação antropológica somente pode ser comparada à revolução neolítica que viu surgirem, em poucos séculos, a agricultura, a criação de animais, a cidade, o Estado e a escrita. Dentre todas as transformações fundamentais que afetaram os países desenvolvidos na época atual, ressaltamos o desaparecimento do mundo agrícola, o apagamento da distinção cidade/campo e conseqüentemente surgimento de uma **rede urbana onipresente**, um novo imaginário do espaço e do tempo sob a influência dos meios de transporte rápidos e da organização industrial do trabalho, deslocamento das atividades econômicas para o terciário e a influência cada vez mais direta da pesquisa científica sobre as atividades produtivas e os meios de vida. As conseqüências a longo prazo do sucesso fulminante dos instrumentos de comunicação audiovisuais (a partir do fim da Segunda Guerra Mundial) e dos computadores (a partir do fim dos anos setenta) ainda não foram suficientemente analisadas. Uma coisa é certa: vivemos hoje em uma destas épocas limítrofes na qual toda a antiga ordem das representações e dos saberes oscila para dar lugar a imaginários, modos de conhecimento e estilos de regulação social ainda pouco estabilizados. Vivemos um destes raros momentos em que, a partir de uma nova configuração técnica, quer dizer, de uma nova relação com o cosmos, **um novo estilo de humanidade é inventado** (LÉVY, 2008, p.16-17) [grifo do autor].

Em meio a este pensamento, entendemos que neste “novo estilo de humanidade” (LEVY, 2008, p. 17) houve uma relativização da presença física e do convívio social. As interações sociais agora são pautadas através de um formato digital, eminentemente diferentes de formas outrora utilizadas. Estas novas interações são retratadas através de *cutucadas*, *curtidas*, *shares*, (Facebook), *mentions*, *RTs*, *unfollows* (Twitter).

Para Lima (2013), altera-se o uso, mas não as características fulcrais da linguagem enquanto elemento chave, fundador do pensamento humano e marcada pela noção de dialogismo preconizada por Bakhtin (2003). “A vida é dialógica por natureza. Viver significa participar de um diálogo” (BAKHTIN, 1961, p. 293), ou seja, mesmo com a evolução dos meios de comunicações, a ideia do diálogo sempre a precede.

2.2 Redes Sociais

Ao adentrar pontualmente nas chamadas Redes Sociais (RS), nota-se que os conceitos advogados por Bakhtin estão postulados de uma maneira muito atual, na medida em que para se constituir as redes precisam estar pautadas na interação dialógica entre seus membros. Dentro destas redes Recuero (2009, p109) destaca o seguinte significado para RS:

[...] sites de redes sociais propriamente ditos são aqueles que compreendem a categoria dos sistemas focados em publicar e expor redes sociais de atores. São sites cujo foco principal está na exposição pública das redes conectadas aos atores, ou seja, cuja finalidade está relacionada a publicização dessas redes. (RECUERO, 2009, p109).

Ademais, é oportuno destacar que a rede social não nasceu com o advento da internet. Há muito tempo a sociedade já estava segregada em grupos que interagiam e compartilhavam interesses comuns e lembranças comuns. Nazistas, hippies, EMOS são exemplos de redes que se agruparam para partilhar uma memória e uma identidade similar. Na percepção de Recuero (2009), a rede veio para facilitar as interações *off-line*, de modo que as interações que ocorriam no seio de uma presença física, agora com a internet puderam ter uma característica eminentemente virtual. Seguindo essa linha de raciocínio, Levy (1999) corrobora com Recuero quando afirma que:

[...] o desenvolvimento das comunidades virtuais acompanha, em geral, contatos e interações de todos os tipos. A imagem do indivíduo ‘isolado em frente à sua tela’ é muito mais próxima do fantasma do que da pesquisa sociológica. Na realidade, os assinantes da Internet (estudantes, pesquisadores, universitários, executivos sempre em deslocamento, trabalhadores intelectuais independentes etc.) provavelmente viajam mais do que a média da população [...] as comunidades virtuais são os motores, os atores, a vida diversa e surpreendente do universal por contato (LEVY, 1999, p.130-131).

No entendimento de Castells (2005, p. 431), a rede mundial de computadores é “a espinha dorsal da comunicação global mediada por computadores (CMC): é a rede que liga a maior parte das redes”. De maneira singela, os usuários se apropriam dessa rede com determinado objetivo (buscar, compartilhar informações, por exemplo) e para formar conexões, constituindo uma rede social.

Dentro deste prisma, Recuero (2010, p. 25), advoga que os atores são percebidos como

“representações dos atores sociais ou como construções identitárias do ciberespaço”. Desta forma compreendemos que

Essas apropriações funcionam como uma presença do “eu” no ciberespaço, um espaço privado, e, ao mesmo tempo, público. Essa individualização dessa expressão, de alguém “que fala” através desse espaço é que permite que as redes sociais sejam expressas na Internet (RECUERO, 2010, p.27).

Não é objeto deste estudo analisar se houve uma exacerbação dos contatos virtuais em detrimento do contato face a face. Nosso interesse reside apenas em tentar esclarecer alguns conceitos acerca deste tema. Otimizando a aceção sobre rede social na internet (RSI) julgamos conveniente destacar, ainda que de maneira breve, que a análise acerca de RSI exorta idiosincrasias pautadas nas bases da sociologia tradicional, uma vez que suas “articulações entraram na agenda da investigações sociológicas sobre as ações individuais e coletivas (SANTAELLA, LEMOS 2010). Castells (2003) compartilha o pensamento que a rede social desempenha o papel de “organizar a interação” entre os mais diversos relacionamentos, sejam grupais ou individuais:

As comunidades, ao menos na tradição da pesquisa sociológica, baseavam-se no compartilhamento de valores e organização social. As redes são montadas pelas escolhas e estratégias de atores sociais, sejam indivíduos, famílias ou grupos sociais. Dessa forma, a grande transformação da sociedade em sociabilidades complexas ocorreu com a substituição de comunidades espaciais por redes como formas fundamentais de sociabilidade (CASTELLS, 2003, p. 106-107).

Deste modo, mais uma vez recorremos à ideia de independência geográfica, sendo que a rede, neste cenário, passa a desempenhar um papel de contato entre as pessoas, não sendo obrigado para isso uma presença física. Por outro lado, esse novo “estilo de sociabilização” gera certo individualismo, no qual temos que as pessoas são autoras da definição do seu próprio saber. É válido ressaltar que o individualismo em rede é

[...] um padrão social, não um acúmulo de indivíduos isolados. O que ocorre é antes que indivíduos montam suas redes, on-line e off-line, com base em seus interesses, valores, afinidades e projetos. Por causa da flexibilidade e do

poder de comunicação da Internet, a interação social on-line desempenha crescente papel na organização social como um todo. As redes on-line, quando se estabilizam em sua prática, podem formar comunidades, comunidades virtuais, diferentes das físicas, mas não necessariamente menos intensas ou menos eficazes na criação de laços e na mobilização (CASTELLS, 2003, p.109).

Feito esta breve contextualização acerca desse novo processo de sociabilização trazido pela rede social é oportuno destacar também o papel que essas articulações digitais trazem para o contexto econômico, “a economia global é constituída pelas trocas e fluxos quase instantâneos de informação, capital e comunicação cultural” (SANTAELLA; LEMOS, 2010 p.16). Ainda para estas autoras, a “sociedade permanece capitalista, mas a base dos meios tecnológicos com os quais ela age saltou da energia para informação” (SANTAELLA; LEMOS, 2010 p.16), ou seja, a busca por informações contidas nestas redes passa a desempenhar papel capital na construção dessa economia do conhecimento. Neste âmbito, temos que a informação é o fator fulcral para o desenvolvimento desta nova economia do conhecimento.

Assim, Santaella; Lemos (2010, p.16) destacam que “o que existe de mais novo nesse circuito é a virada informacional, a manipulação da informação ela mesma, ou seja, ação do conhecimento sobre conhecimento”. Ainda dentro desta linha de entendimento, as autoras destacam características dessa sociedade onde a sua lógica organizacional independe da localização geográfica, a saber:

- a) Globalização de atividades estrategicamente decisivas da economia;
- b) Forma de organização em rede;
- c) Instabilidade do trabalho e individualização do emprego;
- d) Cultura da virtualidade real, construída por um sistema pervasivo, interconectado e diversidade de sistemas de mídia;
- e) Transformação das condições do material da vida, do espaço e do tempo, devido aos espaços de fluxos e do tempo sem tempo.

Inobstante a estas conjecturas e aos comportamentos advindos dessa sociedade em rede, é válido destacar que as mudanças na construção das redes sociais na internet aconteceram de forma paulatina, contemplando, assim, o período que vai sendo alicerçado pela construção própria da www.

Buscando sedimentar melhor a evolução das redes sociais na internet (RSIs), Santaella; Lemos (2010) esclarecem que existem diferentes abordagens, entretanto é possível analisar a partir da relação com as mídias massivas. Para as autoras, “o prisma mais interessante da sua análise reside nas modalidades diferenciadas de interação que evoluem em compasso com a penetração e apropriação social dessas redes (Ibid 2010, p.55).

Ao longo dos anos, esta modalidade vem sofrendo grandes alterações, se analisarmos o modelo interacional da década de 90 temos, na acepção de Santaella, Lemos (2010), que a característica fulcral é o nexu monomodal, ou seja, uma experiência individual e unidirecional, onde tem-se um ponto de acesso na rede e, a partir desse ponto, o usuário tem acesso a outras páginas. Hornik (2005) definiu esta fase como Redes 1.0.

Ainda dentro da premissa evolutiva, Hornik (2005) postula que o passo seguinte foram os contextos de uma rede social 2.0, onde se tem o compartilhamento dos mais variados tipos de arquivos como foto, vídeos, links e interesses afins. Em outras palavras,

As redes sociais 2.0 representaram a transição entre os dois modelos por trazerem múltiplas monomodalidades de interação integradas em uma mesma plataforma. Essa foi uma etapa fundamental na evolução das RSIs, responsável pelo amadurecimento da sociabilidade em rede (SANTAELLA; LEMOS, 2010, p.58).

Orkut, mySpace, LinkedIn são para Hornik (2005) arquétipos de redes que atendem essa definição. Já para as redes sociais 3.0, o que vem à baila como característica definidora é a inserção da mobilidade acompanhada pelo desenvolvimento de aplicativos além da “integração com outras redes e pelo uso generalizado de jogos sociais como Farmville e Mafiawars” (SANTAELLA, LEMOS 2010, p.58).

As RSIs 3.0 trazem um acesso ubíquo e tem modelo multiplataforma, a interação com essas redes pode ser feita de diversos *devices*, smartphone, tablets, pcs. Dentro deste prisma, Santaella, Lemos (2010, p.59) exorta que nas redes de modelo 3.0, o acesso é nômade e mutante, uma vez que “existem tantas vias de acesso quanto vias de integração entre as diversas redes”.

Hornik (2005) propõe uma síntese, em que traz três fases distintas da evolução das redes sociais:

- a) Redes 1.0: coordenação em *real time* entre os usuários, ICQ, MSN são exemplos desta fase
- b) Redes 2.0: caracterizada pela busca do entretenimento, contatos profissionais, marketing social (Orkut, LinkedIn)
- c) Redes 3.0: uso de aplicativos específicos e mobilidade, por exemplo, Facebook, foursquare e Twitter.

2.3 Twitter

Como o objeto deste trabalho reside no entendimento acerca das peculiaridades relacionadas ao Twitter, a ideia de fazer uma incursão sobre o seu funcionamento se mostra muito auspiciosa, uma vez que o entendimento das especificidades irá facilitar a dinâmica proposta por esse estudo. Entretanto, não iremos adentrar em aspectos basais acerca do seu funcionamento, tais como *retweets*, *Direct Message* (DM), *#hashtag*, *Trending Topics* (TT's), etc, pois julgamos que existe vasto material epistemológico à disposição tanto na blogosfera como no mercado editorial.

Criado em 2006 por Jack Dorsey, Biz Stone e Evan Williams, em meio a um projeto de uma empresa de podcasts chamada Odeo (RECUERO, 2010), o Twitter tinha a missão de ser um “SMS na web”. Hoje o Twitter pode ser entendido como um serviço de *microbloggings* onde os seus usuários tem no máximo 140 caracteres para responder a pergunta “*What’s happening?*”, entretanto suas funcionalidades extrapolaram a miríade desse simples questionamento. A versatilidade da API (Application Programming Interface) do Twitter, associada a adesão de celebridades, justificam a “explosão” do *microblogging*. A facilidade do uso da API permitiu que programadores pudessem difundir a sua utilização nas mais diversas plataformas (tablets, navegadores, celulares) e assim tornar o Twitter um serviço multiplataforma, onde as pessoas pudessem acessar de qualquer dispositivo. As personalidades também desempenharam papel capital no crescimento do serviço. Por exemplo, em abril de 2009, o ator Ashton Kutcher⁹ foi a primeiro astro a conseguir um milhão de seguidores no site. Atualmente, o Twitter conta com a aproximadamente duzentos milhões de usuários ativos¹⁰ espalhados por todo o mundo, conforme demonstra pesquisa realizada pelo Emarketer.

⁹ Disponível em <http://infograficos.oglobo.com.br/timeline/8/cronologia-do-twitter#7> acesso em 29/06/2014.

¹⁰ Disponível em <https://blog.twitter.com/2013/experiments-twitter> acesso em 29/06/2014.

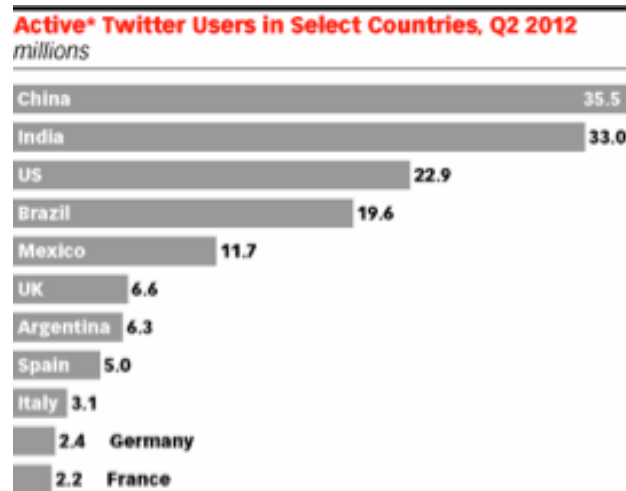


FIGURA 3 - Quantidade de usuários ativos, segregado por país

Fonte: All Techie News

Santaella; Lemos (2010, p.66) definem o Twitter como “mídia social que unindo mobilidade do acesso à temporalidade Always on das RSIs 3.0 possibilita o entrelaçamento de fluxos informacionais e o design colaborativo de ideias em tempo real”. Dentro da égide de informação em tempo real podemos mais uma vez recorrer a grandes volumes de dados criados a partir do Twitter para corroborar com acepção advogada pelas referidas autoras. No primeiro jogo da copa do mundo da Fifa (Brasil e Croácia), realizado no dia 12 de junho de 2014, foram realizados mais de 12,2 milhões de tweets citando o evento. O jogador Neymar Jr, por sua vez, foi o mais mencionado por torcedores de 150 países em torno do mundo. As figuras 4 e 5 evidenciam que a explosão de tweets ocorre justamente no clímax do jogo, os gols.



Figura 4 – Explosão de Tuites no 1º Gol da copa do mundo FIFA
 Fonte: Proxima.com



Figura 5 - Explosão de Tuites no 2º Gol da copa do mundo FIFA

Fonte: Proxima.com

Como se pode observar nas figuras 4 e 5 existe uma representatividade dos fatos em tempo real. Na medida em que os gols vão acontecendo as manifestações no *microblogging* vão acontecendo no mesmo instante, ou seja, é informação *always on*.

No que se refere à sua utilidade do Twitter, Santaella, Lemos (2010, p.66) partem do seguinte entendimento:

O Twitter serve como um meio multidirecional de captação de informações personalizadas; um veículo de difusão contínua de ideias, um espaço colaborativo no qual questões, que surgem a partir de interesses dos mais microscópicos aos mais macroscópicos.

Em resumo, o que as autoras propõem é que essa mídia funciona como agora digital, ou seja, um espaço onde são debatidos os mais variados contextos, desde assuntos de cunho políticos a frivolidades acerca da vida de personalidades. Outro ponto de extrema relevância é que, de certo modo, faz com o que o Twitter se sobressaia em relação as outras RSIs é o fato que enquanto nas outras redes sociais como Facebook a interação social é consubstanciada através de contatos pessoais entre os usuários, no Twitter o foco reside na qualidade e no tipo de conteúdo veiculado por um usuário específico (SANTAELLA; LEMOS, 2010).

Ainda dentro dessa dicotomia: Twitter versus outras RIs as pesquisadoras trazem mais um aspecto para nossa reflexão. O microblog apresenta uma tônica da interação e de formação diferente, uma vez que, os seus laços sociais são construídos em fluxos coletivos de ideias compartilhadas em tempo real que, neste arcabouço, pautam-se por um movimento frequentemente contínuo, já nas demais RSIs é baseado em vínculos preexistentes (SANTAELLA, LEMOS 2010). No Facebook, por exemplo, tem-se a predominância de laços sociais que já existem, são basicamente os amigos, colegas de trabalho ou pessoas conhecidas do mundo off-line. Dentro dessa premissa, as pesquisadoras entendem que no Twitter é mais difícil identificar os padrões de formação de laços sociais. “Muitas vezes, é possível desenvolver laços próximos com um usuário que não esteja nos seguindo de volta, ou com alguém que seja nosso seguidor e não desejamos seguir.” (SANTAELLA; LEMOS 2010, p.91).

Este enfoque faz com que tenhamos o entendimento de que o Twitter funciona, também, como um catalizador na construção de uma inteligência coletiva, sendo que neste contexto nos apropriamos de um viés conceitual concebido por Levy (1998, p.29), onde postula que toda inteligência coletiva tem pressuposto “pensamento com ideias, línguas, tecnologias cognitivas recebidas em uma comunidade.” E a partir dessa inteligência coletiva pode-se analisar, inclusive, como se dá o processo de formação de identidade. Dentro dessa perspectiva, o pensamento cunhado por Castells (2002, p. 499) corrobora com o nosso entendimento:

redes são estruturas abertas capazes de expandir de forma ilimitada, integrando novos nós desde que consigam comunicar-se dentro da rede, ou seja, desde que compartilhem os mesmos códigos de comunicação (por exemplo, valores ou objetivos de desempenho). Uma estrutura social com

base em redes é um sistema aberto altamente dinâmico, suscetível de inovação sem ameaças ao seu equilíbrio.

Assim sendo, essa característica pontual do Twitter, proporciona uma inquietação, para esta pesquisa, ainda mais especial, uma vez que a construção desses laços não se limitam apenas por proximidade ou afinidade de relacionamentos. Entender como são formuladas a construção de marcas identitárias dentro da “twitolândia” a partir dos nós gerados pela conectividade dessas RSIs e o grande volume de dados gerado por essas articulações trazem à tona o questionamento suscitado na gênese dessa pesquisa: será que os tradicionais métodos em pesquisa em comunicação conseguem deslindar tais questionamentos? Acreditamos que não, pois o manancial de dados produzido e o dinamismo que sedimenta essa análise necessitam de uma interdisciplinaridade metodológica como a de outros saberes, sobretudo os que advêm das ciências da computação.

4. IDENTIDADE COLETIVA DENTRO DO CIBERESPAÇO

A sociedade tem acompanhando nas últimas décadas diversas modificações no que diz respeito aos processos de interação. Hoje, as novas tecnologias digitais recriaram a forma de compartilhar comportamentos, pensamentos e ideias, além do uso da internet está impulsionando cada vez mais novos relacionamentos que desprezam o conceito de espaço e tempo.

Essa nova tônica circunscreve uma forma de sociabilidade cada vez mais paradoxal, visto que as pessoas estão mais conectadas, mas ao mesmo tempo fisicamente distantes. Criar esse paradoxo “conectado, mas distante” só foi possível graças à profusão da internet e dos contornos criados pela definição do que é ciberespaço, que neste contexto pode ser definido como: “espaço não físico ou territorial, que se compõe de um conjunto de redes de computadores através das quais todas as informações (...) circulam”. (LEVY, 1999, p. 87).

Em face deste conjunto de computadores conectados em rede, a instantaneidade da circulação da informação apresenta-se como outro aspecto definidor do ciberespaço. Em sendo assim, Lemos (2003) advoga que o potencial do ciberespaço estaria em fornecer uma comunicação ágil e democrática a todos os cidadãos.

Outrossim, no ciberespaço, as pessoas estão passando mais tempo conectadas, utilizando assim espaços de conversas online e buscando estabelecer laços que outrora eram impossibilitados em virtude das grandes distâncias. Para tanto, basta analisar o tempo médio despendido nas redes sociais virtuais. Em recente reportagem veiculada no portal Olhar Digital¹¹, os brasileiros passam em média 12 horas por mês no Facebook, isso equivale a 1,68% do mês.

Deste modo, o ciberespaço oportuniza ao sujeito pós-moderno, através da sua interface digital, novas facetas para um cenário cada vez mais dinâmico e centrado em um ambiente que favorece a exposição de ideias e das atividades do cotidiano. Neste sentido, Lévy (2010, p.148, grifo nosso) entende que todos esses processos de interação funcionam como “depósito de mensagens, contexto dinâmico acessível a todos e memória coletiva alimentada em *tempo real*”.

Essa voracidade pela instantaneidade, no nosso entendimento, alicerça o conceito acerca de identidade do sujeito pós-moderno cunhado por Hall (2006). Para o autor, a identidade pós-moderna é caracterizada por um contexto efêmero, no qual os sujeitos deixaram de ter uma identidade rígida, única, pautada em um espaço eminentemente territorial. Agora, por conta dos aspectos oportunizados pelo ciberespaço, tais como fluidez na comunicação, acesso as mais diversas fontes de informação, os sujeitos passaram a “experimentar” várias identidades, conforme seus interesses e objetivos.

Se antes a identidade era vista como algo hermético e estaque, nos dias de hoje ela passa a ser virtualizada e dinâmica. Hoje, a sociedade adentra, na visão de Bauman (2007), em uma vida fluída, onde viver é fruto de constantes mudanças que nos são demonstradas nestes novos tempos. Essas mudanças, por sua vez, tiveram sua sedimentação pautada com o surgimento da internet. As tecnologias de interconexão proporcionaram um novo reagrupamento social, onde as barreiras geográficas foram desprezadas e a relativização do espaço e tempo deu lugar a um novo ambiente de interação (SILVA, 2008).

Neste espaço virtual, temos uma espécie de “versões de identidades” nas quais os sujeitos tem a opção de construir uma identidade diferente da sua versão “off line”. Ou seja, dentro do ciberespaço, o indivíduo pode modular a sua construção identitária de modo que a mesma se adeque a realidade que deseja se apresentar.

¹¹ Disponível em <http://olhardigital.uol.com.br/noticia/40875/40875> acesso em 14/05/14

Dentro deste prisma, Turkle (1997, p.15) sintetiza o comportamento da identidade virtualizada como “in my computer mediated words, the self is multiple, fluid, and constituted in interaction with machine connections”¹². Essa identidade no ciberespaço pode apresentar também uma variante idílica, uma vez que o sujeito pode projetar a sua identidade sendo algo totalmente diferente da vida real: um super herói, um milionário, um intelectual. À guisa do pensamento de Baumam (2005), essa identidade fluida passa a operar sobre o cenário mais conveniente para o indivíduo, assumindo assim um caráter de “identidade mutante”.

Essa mutação é fomentada pelo fato de que o ciberespaço oferece um “pseudo véu” de anonimato. A respeito disso é válido ressaltar que “a sensação de liberdade fomenta o caráter performático do fenômeno pós-moderno e lhe assegura um tratamento um tanto quanto lúdico, por parte dos sujeitos da cibercultura” (SILVA 2008).

Igualmente, as identidades construídas no ciberespaço circunscrevem características próprias, como a onipresença e onisciência. No ambiente digital, como pontuado anteriormente, não existe a barreira espaço-temporal, o indivíduo pode em qualquer lugar, a qualquer tempo, se comunicar com vários usuários ao redor do mundo. Essa comunicação virtual permite novos contatos e conseqüentemente novas experiências de modo a proporcionar a criação de uma “aldeia global”. Deste modo, temos que o “indivíduo contemporâneo”, antes limitado às suas raízes e a sua tradição (modelo de identidade clássica), se liberta na transcendência possibilitada pela comunicação através de um novo *modus vivendi*, manifestando sem as amarras espaciais seu desejo de participação, de troca e interação. (BEZERRA, SOBRINHO 2011, p.125).

Essa capacidade de compartilhar conhecimentos individuais no ciberespaço muda mais uma vez o fluxo de construção de identidades, uma vez que o indivíduo deixa de ser apenas um lugar e passa a ser global, assimilando saberes e culturas.

¹² Em meus mundos mediados por computador, o eu é múltiplo, fluido e construído na interação com os mecanismos de conexão (tradução nossa).

4.1 A Identidade coletiva nas redes sociais

Com o advento da internet e a criação das mais diversas redes sociais, como Facebook, Twitter, Instagram, dentre outras similares, a temática da identidade vem sendo pontuada de uma forma mais exacerbada. Tanto que no momento do cadastro nas referidas redes, o primeiro questionamento com que nos deparamos é “*quem sou eu?*”. Este questionamento vem transcorrendo do mundo *offline* para o mundo *online* de uma forma tão pujante que muitas vezes passa despercebido.

Dentro deste ambiente digital, onde os bancos de dados destas redes sociais (RS) carecem de dados que descrevam uma identidade cada vez mais detalhada, percebemos que as palavras postuladas por Calhoun *apud* Casttels (1999) estão presentes de uma forma muito patente, uma vez que os sites que se configuram como RS exigem os atributos que caracterizam um processo de descrição de individuação.

Neste sentido, é de suma importância destacar que as redes sociais assumiram, dentro desse contexto, um vetor catalisador nas representações de uma identidade fluída, visto que se tem percebido uma necessidade de atender várias identidades para fazer parte do grupo. Estas redes sociais propõem a construção de uma gama de significantes e assim possibilitam fatores que influenciam a formação de identidade (AZEVEDO, 2012).

Isto posto, percebe-se a existência constante de um processo de interação, onde os atores são pautados em um processo de negociação e que o fruto final desse processo é a aceitação por parte do grupo. Deste modo, entendemos que a construção do “sujeito digital” vem muito arraigada ao processo de alteridade propriamente dito, ou seja, a construção do eu, depende do outro.

[...] a partir uma rede de agentes e agências sociais, como seus fluxos e interações, e não como uma realidade dada e neutralizada, mas como um processo de permanente construção e desconstrução, podemos perceber o quanto a posição dos agentes dentro dessa rede [...] é claramente constitutiva de identidades individuais e coletivas (ENNE, 2004, p. 106).

Ana Luiza Enne (2004) vai propor que a construção da identidade é concebida a partir do reconhecimento do outro, seja ele em um processo de conflitos ou semelhanças. Assim, a

‘apreensão dos mecanismos de identificação’ seriam fundamentais “porque eles refletem a identidade em processo, como é assumida por indivíduos e grupos em diferentes situações concretas” (ENNE, 2004, p.107).

Como mencionado em outros momentos, toda essa interatividade faculta a criação de pontos de encontros e dentro deste cenário, a linguagem floresce como papel modal nestas articulações. Entretanto, o processo de assimilação de uma ideia ou consentimento de uma ideologia acontece em um aspecto meramente simbólico, representado por um simples *curtir* ou um *rt*.

Neste sentido, a identidade em rede também abarca características simbólicas, uma vez que a representação do sujeito, neste ambiente, pauta-se por uma fluidez ainda mais acentuada. Isso se dá pela facilidade que a internet oportuniza ao sujeito de utilizar “máscaras identitárias”, ou seja, o indivíduo pode se apresentar como bem entender dentro destas redes: alternando o seu “avatar”, mudando seu gênero, ou até mesmo modificando seu discurso.

Outro ponto de destaque no processo de construção de identidade na rede é o fato que a interatividade proporciona a troca/criação de experiências e dentro deste contexto o sujeito adentra por vários ambientes, assimilado diversos contornos sociais, criando, assim, um ambiente de degustação de papéis sociais.

Além disso, esse invólucro de papéis sociais oportuniza a criação de uma inteligência coletiva, que é definida por Lévy (2003, p. 28), como “[...] uma inteligência distribuída por toda parte, incessantemente valorizada, coordenada em tempo real, que resulta em uma mobilização efetiva das competências”. Nesta acepção, concordamos com Santaella; Lemos (2010, p.25) em achar mais conveniente o nome de “ecologia cognitiva” visto que o valor semântico do termo inteligência está muito atrelado à racionalidade, ao passo que “ecologia cognitiva” sugere “diversidade, mistura entre razão, afeto e o impulso para a participação” (Ibidem).

Isto posto, esta “ecologia cognitiva” é imbricada pela participação de inúmeros canais globais, organizados dentro de uma rede, que interagem em uma velocidade surpreendente, possibilitando assim vários nexos onde é possível assimilar aspectos culturais, ideologias ou simplesmente conhecer os contornos que estão sendo debatidos dentro dessa rede. Um exemplo que endossa essa “ecologia cognitiva” é o uso de *hashtags* no Twitter, que além de funcionar como um indexador de conteúdo, facilitando a busca por determinadas temáticas,

pode ser utilizado também como um agente que favorece de troca de ideias e consequentemente uma “mobilização efetiva de competências” (LÉVY 2003, p. 28). Dentro dessa conjectura, Montilla (2011) ventila como o uso das *hashtags* pode prover um ambiente de conversão, a saber:

La comunicación sin intermediarios, entre las cuentas de usuarios comunes y las oficiales de políticos, gobernantes y medios de comunicación permite un verdadero intercambio de ideas entre los actores políticos de una sociedad. Con La utilización de los “*hashtags*” o “palabras clave”, es posible unirse a conversaciones sobre un tema específico, independientemente de que dichos usuarios sigan o no SUS respectivas cuentas¹³ (MONTILLA, 2011, grifo do autor).

Temos que toda essa interatividade das redes passa a ser, também, presença marcante no processo de construção de identidade, visto que o sujeito pode apresentar uma pluralidade de identidades na mesma velocidade que a rede sugere, ou seja, este sujeito pode adequar seu *status* a realidade que lhe for mais apropriada de forma instantânea. Neste contexto, é oportuno destacar que toda essa interatividade gera um espaço maior de debates, e consequentemente um grande fluxo de informação, o que traz certos gargalos quando se pretende realizar pesquisas de identidades nas redes sociais.

Outro aspecto que deve ser evidenciado é que as redes sociais na internet podem funcionar como anteparo para o processo de manutenção de uma memória coletiva. Partindo da premissa que memória é “[...] a constituição gigantesca e vertiginosa do estoque material daquilo que nos é impossível lembrar, repertório insondável daquilo que poderíamos ter necessidade de nos lembrar” (NORA, 1993, p. 15) e que os lugares de memória são espaços para armazenar as memórias que não são espontâneas, legitimando a memória a nível coletivo (ENNE, 2004) entendemos que as RSI podem ser consideradas como objetos de armazenamento, verdadeiros banco de dados que irão marcar os acontecimentos de um povo ao longo da sua existência. Em outras palavras, um lugar de memória digital.

Nesta conjuntura, as redes sociais funcionam também como memória social, visto que os indivíduos podem aglutinar seus depoimentos de modo a tornar o processo de construção de memória mais dinâmico. Através do compartilhamento de informação realizadas ao longo

¹³ A comunicação sem intermediários, entre contas de usuário comum e as oficiais dos políticos, governantes e dos meios de comunicação, permite uma verdadeira troca de ideias entre os atores políticos de uma sociedade. Com o uso de “*hashtags*” ou “palavras-chave”, é possível participar de conversas sobre um tópico específico, independentemente de esses usuários seguirem ou não suas respectivas contas (MONTILLA, 2011, tradução nossa).

do tempo, lembranças, fatos e contextos sociais podem ser recuperados. Neste sentido, baseado em Marcondes Filho (1995, p.315), partimos do entendimento de que as redes sociais favorecem a disseminação de um arcabouço no qual é possível gerar um ambiente de rememoração. Visto que para o referido autor “rememorar é sempre uma reconstrução, permeada de artifícios pessoais e sociais, nos quais as tecnologias (novas e antigas) exercem presença significativa”.

Eliza Bacheга Casadei, na obra *Os Novos Lugares de Memória na Internet: as práticas representacionais do passado em um ambiente online*, propõe que estes espaços tecnológicos favorecem a criação de um elo com o passado, um túnel largo no qual facilmente pode-se atravessar e conhecer como foi construído o passado.

Os entrecruzamentos e as reestruturações de linguagens proporcionados pelas novas tecnologias de informação e comunicação parecem pôr em operação um redesenho dos modos tradicionais de transmissão da memória e do passado, a partir do estabelecimento de novos modos de sociabilização e de interação com os tradicionais “lugares de memória”. Além de transformar a configuração destes espaços em seu cerne (através da reconfiguração de sua linguagem), esses novos espaços promovem uma nova forma de relacionamento das pessoas com o passado, a partir da abertura da possibilidade de interação e participação ativa na construção desta memória relacionada à construção das identidades coletivas (CASA DEI, 2009, p.3).

Diversos veículos de comunicação usam estes espaços tecnológicos buscando realçar e perceber como foram construídas determinadas temáticas. O portal G1.com, por exemplo, ao propor uma análise sobre o debate para eleições presidenciais do ano de 2014, criou através de tecnologia um ambiente onde foi possível mensurar a “temperatura no Twitter¹⁴”. A ferramenta analisa, dentre outras coisas, a quantidade de menções relacionadas a cada candidato, de modo que é possível ponderar qual foi a aceitação de cada presidencial durante o debate. Esta ferramenta descreve de forma pontual como o uso das redes sociais pode funcionar como espaço de rememoração, visto que sempre que desejar, os usuários poderão fazer uso e assim inferir como era a aceitação dos candidatos naquele ano.

¹⁴ Disponível em <http://g1.globo.com/politica/eleicoes/2014/debate-presidencial-a-temperatura-no-twitter.html>



Figura 6 - A TEMPERATURA NO TWITTER

Outro recorte em que se percebe a utilização das redes sociais funcionando como lugar de memória foi o trabalho desenvolvido pelos pesquisadores Said, Magalhães (2012) sobre a hashtag #grifesvivas. O referido trabalho visava analisar, na rede social Twitter, quais eram as grifes vivas na cidade de Teresina, capital do Piauí. Dividido em 3 categorias, sendo elas: *personalidades* (referente a pessoas ou personagens), *cenários* (lugares memorados, como espaços públicos, lojas e restaurantes) os autores analisaram, durante o período de 03.03.2011 a 10.03.2011, 625 tweets, nos quais buscaram perceber como era construído o imaginário do piauiense acerca das pessoas e acontecimentos que marcaram a cidade. Dentre outros aspectos, a pesquisa demonstrou que:

[...] os registros deixados pelos twitteiros no Grifes Vivas THE tem uma função social. Eles ajudam a reorganizar o passado, reconfigurando a dimensão da temporalidade, na qual novos sentidos são produzidos pela relação entre eventos passados e projeção de futuro. (SAID; MAGALHÃES, 2012, p.42)

Como se pode perceber, a pesquisa expressou como as redes sociais podem funcionar

como repositório de resgate de fatos passados, um ambiente de busca onde se pode conhecer, através de um processo iterativo, como a história foi construída.

Contudo, analisar essa dinamicidade oriunda nas redes sociais não é uma tarefa simples. Na pesquisa das #grifesviva she notou-se certa dificuldade e extrair informações, visto que os próprios pesquisadores pontuaram que não utilizaram nenhuma ferramenta para analisar as postagens e talvez por esse motivo os mesmos optaram por averiguar apenas o que ocorreu durante 7 dias.

4.2 As dificuldades metodológicas para pesquisa de identidade coletiva nas redes sociais

Atualmente, com a profusão da internet, o excesso de dados gerado a partir das interações digitais passou a ser um ativo para muitas organizações. Informações pessoais, *hobbies* e interesses revelam a identidade do sujeito que, deste modo, tornaram-se pauta recorrente desta “indústria de dados”.

A rede social Facebook, por exemplo, coleta aproximadamente 70 informações diferentes sobre seus usuários, tais como informações sobre idade, cidade natal, páginas visitadas, visões religiosas e políticas, atividades recentes, metadados de fotos (hora e local em que foram feitas), configurações faciais, número de telefone, endereço de IP, número de cartão de crédito, o que se olha na linha do tempo de outras pessoas, as mensagens trocadas e páginas que visitam, entre outros (AGÊNCIA BRASIL, 2015¹⁵). Todas essas informações constituem um padrão identitário que é comercializado para os anunciantes da rede social. Sob essa lógica de identificação de grupos de personas são oferecidos produtos, serviços, além de notícias cada vez mais correlatas aos seus usuários.

Neste sentido, as exacerbações criadas em torno da superexposição nas redes sociais, associadas aos interesses por parte das empresas em colher mais informações sobre a identidade dos seus usuários tem oportunizado um ambiente de imersão, onde se busca de forma cada vez mais acurada perceber tanto o contexto individual como o coletivo.

Em recente reportagem concedida à revista VEJA, o diretor do MIT (Massachusetts Institute of Technology) Alex Pentland pontuou que os dados pessoais dos indivíduos são “o

¹⁵ Disponível em <http://agenciabrasil.abc.com.br/geral/noticia/2015-01/facebook-vai-usar-novas-regras-de-privacidade-e-anuncios-partir-deste-mes> acesso realizado em 06/01/15

novo petróleo da internet e a nova moeda do mundo digital” (PENTLAND, 2015, p.17) e continua sua reflexão advogando que:

Nos últimos 300 anos, quase nada mudou no estudo do comportamento do indivíduo em sociedade. Ainda estamos presos a ideias que vêm do século XVII [...] Ocorre que até 90% do nosso comportamento é influenciado por relações. Mas isso não é propriamente novo. O que é novo é que agora somos capazes de observar, entender e mesmo influenciar certos comportamentos humanos graças à gigantesca quantidade de dados disponível no mundo (PENTLAND 2015, p.17).

Contudo, mesmo com todas as informações disponíveis no ciberespaço, realizar uma pesquisa diante desses mananciais de dados não é uma tarefa trivial. O que se tem percebido são grandes *players* no mercado oportunizando esse tipo de análise. Empresas como e.life¹⁶ e Scup¹⁷ tem investido cada vez mais em algoritmos especializados para realização destas pesquisas. Tal investimento referenda a necessidade de se conhecer os padrões identitários dentro das redes sociais.

A fluidez dessas redes, como pontuado neste trabalho, gera uma grande quantidade de dados e um valor incomensurável. Entretanto, essa dinâmica de excesso informacional só se faz pertinente se houver uma razoável competência acerca das informações que estas redes podem gerar. Na visão de Bruno (2006, p. 14),

Os dados não são, em si mesmos, nem muito reveladores nem facilmente acessíveis aos sentidos nus, pois, além de serem extremamente numerosos, são fragmentados e não compõem um indivíduo a ser apreendido pelo olhar; estes indivíduos só emergem num segundo momento graças às técnicas de composição de perfis computacionais.

Dentro dessa conjectura, percebe-se um paradoxo: um crescimento de dados em escala exponencial, ao passo que a análise destas informações é feita de forma morosa e custosa. Neste sentido, BROWN; DUGUID (2001, p. 10) propõem que houve inversão nas pesquisas na web. Outrora se partia da premissa de busca pela informação, ao passo que nos dias atuais essa pesquisa tem como cerne a abstração das informações a favor do conhecimento implícito a essas informações. A saber,

¹⁶ <http://www.elife.com.br/>

¹⁷ <http://www.scup.com/>

a preocupação com relação ao acesso à informação deu lugar à preocupação sobre como lidar com o volume de dados que podemos acessar [...] mal se abre o fluxo de informações – o efeito mais parece o estouro de uma represa do que abertura de uma torneira - , controlar essa torrente tornou-se rapidamente um problema crucial BROWN; DUGUID (2001, p. 10).

Essa nova preocupação em trabalhar com sobrecarga de informação é alicerçada pelo fato de que “estamos afogados em informação, mas sedentos de conhecimento” (NAISBITT *apud* FARIA; QUONIAM, 2002, p. 14). Nesta sociedade pontuada como “sociedade do big data”, a dinâmica pela busca de conhecimento torna-se cada vez mais patente. Adentramos em um cenário onde não precisamos mais de informação, mas de formas para procurar fragmentar o conhecimento. Brown, Duguid (2001, p. 106) trazem uma apropriação oportuna quando afirmam que “...precisamos não simplesmente de mais informações, mas de pessoas para assimilar, compreender e dar sentido a tudo isso”. Na mesma linha de raciocínio, Levet (2001, p. 38) corrobora com este pensamento quando pontua que “não é mais o acesso à informação que é a mola do crescimento e do emprego, mas a aptidão dos atores em transformar, compreender, interpretar e utilizar a informação”.

Contraopondo esse pensamento, Castells (1999) afirma que o crescente volume de informações é proporcional à capacidade do ser humano de selecionar e compreender o que está sendo exposto. No nosso entendimento, corroboramos com a ideia de Levet (2001), tendo em vista que sem a ajuda de mecanismos tecnológicos, o volume de dados imposto pela web inviabiliza uma pesquisa mais detida nos objetivos de quem a realiza. Ou seja, diante do volume informacional, o grande gargalo é justamente a mineração dessas informações, criando conhecimentos úteis e oportunos a partir delas. Sobre a técnica de mineração de dados e seus benefícios para a pesquisa de identidade nas redes sociais falaremos de forma mais aprofundada a diante.

Do apresentado, temos que a nova tônica deste *boom* informacional é prover técnicas que proporcionem análise rápida e acurada destes enormes volumes de dados. Neste cenário, percebemos um hiato nas pesquisas científicas no campo da comunicação, uma vez que dentro da nossa compreensão, as técnicas atuais já não conseguem resolver as inquietações advindas do bojo dessa era informacional.

Tomemos como exemplo a análise de identidades no Twitter, uma mídia que pode

atingir 143.199 tweets por segundo¹⁸. Diante de uma marca tão expressiva, qual método tradicional em pesquisa em comunicação poderia responder de forma rápida, por exemplo, a qual o lugar de fala destes usuários? Qual o percentual por gênero? Certamente, a coleta manual dos dados acarretaria uma demanda expressiva de tempo, dificultando a realização e exploração de maior contingente de dados.

Em face deste óbice, nossa proposta é sugerir um método que possa deslindar esses questionamentos no mesmo dinamismo em que eles ocorrem. Para tanto, optamos por utilizar Data Mining, uma técnica que não nasceu no seio da comunicação, mas que através de um modelo transmetodológico pode facilitar a pesquisa de identidade nas redes sociais. Tal técnica é muito utilizada em outras áreas, tais como a Administração e o Marketing, e pela sua eficácia trouxe ganhos significativos a ambos.

¹⁸ Quantidade de Tweets realizada durante a exibição do anime Castle in the Sky no Japão, disponível em <https://blog.twitter.com/2013/new-tweets-per-second-record-and-how> acesso em 10/01/2014

5. KNOWLEDGE DISCOVERY IN DATABASES

Os constantes avanços na área da Tecnologia da Informação (TIC) têm viabilizado o armazenamento de grandes e múltiplas bases de dados (GOLDSCHMIDT & PASSOS, 2005). Associado a isto, temos a redução dos custos dos dispositivos de armazenamento que, por sua vez, despertam na sociedade o desejo mais agudo de manter informações. Muito embora essas novas tecnologias ganhem proporção cada vez mais acentuada, constata-se em estudos recentes que 85 % (oitenta e cinco por cento) de toda a informação disponibilizada no mundo ainda está em formato textual (IBM, 2008).

No entanto, a análise destas informações, muitas vezes desestruturadas, é de difícil codificação, não é uma tarefa fácil, métodos tradicionais de análise de dados, baseados principalmente no manuseio direto das informações pelo homem, simplesmente, não permitem a manipulação de conjuntos volumosos de dados (SINGH, 2001). Neste sentido, é premente o emprego de ferramentas de descoberta de conhecimento em base de dados (Knowledge Discovery in Databases - KDD), que permitam melhor aproveitamento das informações disponíveis, podendo as mesmas serem processadas segundo o desejo do pesquisador, otimizando assim a leitura, bem como possibilitando a descoberta de novas informações a partir da interpretação de dados similares ou ainda repetitivos.

Por definição temos que KDD é um processo composto de várias etapas, não trivial, interativo, para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados (FAYYAD et al, 1996). A utilização do termo processo sugere que existem passos a serem executados e, neste sentido, no KDD a sequência correta de execução dos passos é primordial para o alcance da descoberta de conhecimento. Face a este contexto, as etapas do processo de *Knowledge Discovery in Databases* são: criação de um banco de dados alvo (dados de interesse), pré-processamento, transformação de dados (formatação), data mining (mineração de dados), interpretação/avaliação, conforme demonstra a figura 7:

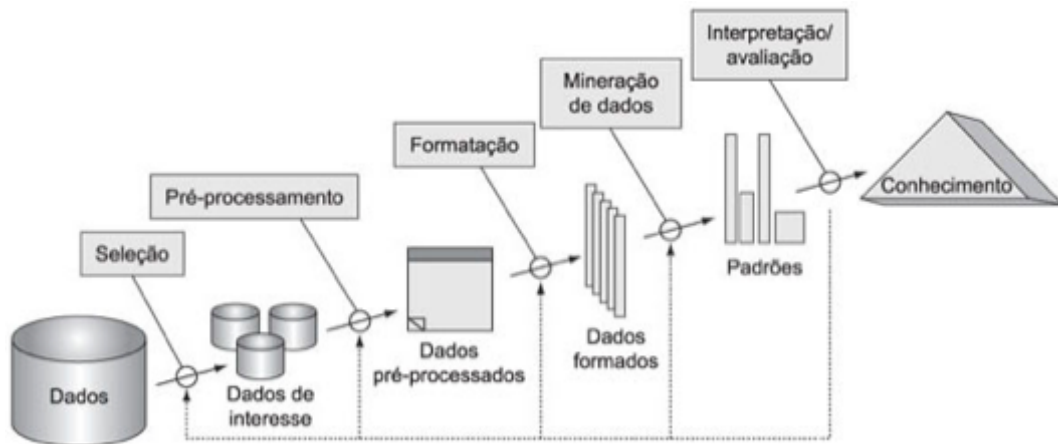


Figura 7 - Etapas do processo *Knowledge Discovery in Databases*

Fonte: FAYYAD et al, 1996.

5.1 Seleção

A primeira fase do processo de KDD consiste na busca dos dados para iniciar o processo de investigação de conhecimento e pode ser realizada basicamente de duas formas: estática ou automática. A estática é a forma mais simples, porém é a mais trabalhosa, tendo em vista que é feita de forma manual por quem está manuseando as informações. A forma automática, por sua vez, precede a utilização de mecanismos automáticos para busca dos dados, esses mecanismos são robôs autômatos que navegam nos sites coletando as informações e armazenando em um banco de dados próprio.

5.2 Pré-processamento

O pré-processamento consiste na fase imediatamente após a seleção de dados. Baseia-se na “limpeza” dos dados, adequando-os correntemente para as próximas etapas. De forma mais ampla, esta fase compreende as seguintes etapas: identificar, compactar e tratar dados corrompidos, atributos irrelevantes e valores desconhecidos. (BATISTA, G. E. A. P. A., 2003). Frequentemente, os dados são encontrados com diversas inconsistências, tais como registros incompletos, valores imprecisos ou mesmo dados passíveis de questionamento. A etapa de limpeza dos dados visa, justamente, eliminar estes problemas de modo que eles não

influem no resultado dos algoritmos usados.

5.3 Formatação

Após o pré-processamento, os dados devem ser submetidos a uma adequação em relação ao enquadramento dos formatos. Nesta etapa, é verificado, por exemplo, se as datas estão com a mesma codificação (mm/dd/yy ou dd/mm/yy). Esse tipo de formatação visa corrigir possíveis inconsistências na base de dados.

5.4 Mineração de dados (data mining)

Esta etapa é o cerne do processo de descoberta de conhecimento em base de dados. Para tanto, consiste na busca por padrões informacionais até então desconhecidas. Face a sua importância dentro do processo de busca de conhecimento em base de dados, esta etapa será aprofundada nos próximos capítulos desse trabalho.

5.5 Interpretação e análise

Consiste na observação dos resultados obtidos visando prover inferências acerca de todo o processo de KDD. Apesar desta fase ser viabilizada por meio dos dispositivos tecnológicos, vai ser justamente nela em que o leitor (pesquisador) vai entrar em ação, cabendo a ele o entendimento surgido segundo o seu ponto de vista.

Feita essa discussão a respeito do processo de KDD, sentimos a necessidade de exortar, ainda que de maneira sucinta, como se dá a abordagem de dados ao se trabalhar com textos puros, uma vez que dentro da análise pontuada neste trabalho também iremos trabalhar com análise de dados na forma de texto.

5.6 Tipos de Abordagens de Dados

Antes de iniciarmos a aplicação das técnicas de data mining na comunicação, percebeu-se a necessidade de definir como as informações são tratadas em um texto. Para Ebecken et al.

(2003) existem duas abordagens que podem ser utilizadas: análise semântica e análise estatística. A saber,

O estudo dessas abordagens tem por objetivo melhorar a qualidade dos resultados no processo de mineração de textos, através do entendimento do funcionamento da linguagem natural ou da importância de determinados termos no texto (através da frequência), permitindo a produção de melhores resultados. (SILVA, 2007 p. 11)

5.7 Análise Semântica

É pautada na abordagem natural de leitura, ou seja, como os seres humanos interpretam um texto, fazendo uso do significado das palavras, características morfológicas, sintáticas e do contexto em geral, tendo como foco a funcionalidade dos termos do texto. (CARRILHO, 2007).

Na Análise Semântica, há a utilização da rica informação semântica, presente em qualquer linguagem, em proveito do processo de obtenção de conhecimento a partir de dados textuais. Mais do que considerar apenas aspectos estatísticos no tratamento de textos, a abordagem por Análise Semântica considera com grande centralidade a linguagem natural nos processos de Mineração de Textos. (SOARES, 2007, p. 42).

Na visão de Silva (2007), a aplicabilidade desse tipo de análise justifica-se pela melhoria em qualidade da Mineração de Textos quando incrementado a partir de um processamento linguístico mais complexo.

5.8 Análise Estatística

Esta análise pauta-se na frequência com que cada termo aparece no texto. No entanto, “informações sobre contextualização, precedência ou sucessão de outros termos não são consideradas” (SOARES, 2007, p. 43). A principal vantagem desta metodologia é que pode ser utilizada em qualquer idioma. A contextualização do termo, bem como em que parágrafo está inserindo, que termos o antecedem ou que estão diretamente relacionados são irrelevantes para este tipo de análise.

Ebecken et al. (2003) propõem que o aprendizado estatístico ou estimativa de dados perpassa os seguintes passos:

- i. **Codificação dos Dados:** é feita a escolha de um modelo de codificação, indicado ou não por um especialista, capaz de identificar características relevantes dentro de um texto e descartar as irrelevantes, buscando sempre manter as principais propriedades dos dados;
- ii. **Modelos de Representação de Documentos:** o principal modelo de representação de documentos utilizado na tarefa de mineração de textos é conhecido por *bag of words*. Neste tipo de codificação, considera-se uma coleção de documentos como uma espécie de *container* de palavras, ignorando a ordem das palavras dentro do texto assim como os caracteres de pontuação ou estrutura. O número de vezes que uma palavra aparece é mantido. Desta maneira, é possível obter um resumo de todas as informações expressas por um documento.

O quadro 2, proposto por Carrilho (2007), pontua as principais características das duas abordagens:

ANÁLISE ESTATÍSTICA	ANÁLISE SEMÂNTICA
<ul style="list-style-type: none"> ▪ Utilizável em qualquer idioma. ▪ Modelos com simples implementação e conhecidos na literatura. ▪ Descarta qualquer valor semântico presente nos textos. 	<ul style="list-style-type: none"> ▪ Necessita conhecimento específico do idioma que será objeto de análise. ▪ Utiliza a informação semântica dos textos, tal como humanos.

Quadro 2 - As principais características de cada uma das abordagens para a Análise de Textos

6. MÉTODOS DE DATA MINING

Como pontuado anteriormente, a análise de grandes volumes de dados apresenta-se como uma das grandes dificuldades dessa sociedade imergida em excesso de informação. Neste sentido, Hearst (1999) esclarece que *data mining (DM)* é um método surgido para apoiar pesquisadores a derivar novas e relevantes informações a partir de grande coletânea de dados. É um processo parcialmente automatizado, no qual o pesquisador ainda evolue-se e interage com o sistema, ou seja, para que o processo exista, é fundamental a interação entre homem e máquina. Nesta “simbiose” entre ambos em busca de informações em base dados, Weiss (2007, p.22) pontua que a

Busca de informação valiosa em grandes volumes de dados. Data mining é o esforço desenvolvido por homens e máquinas. Os homens desenham os bancos de dados, descrevem os problemas e setam os objetivos. As máquinas mineram os dados, em busca de padrões que atendam a estes objetivos (ibidem).

Dento dessa acepção, é oportuno salientar que dificilmente a técnica de DM vai atingir seus objetivos de forma isolada, sem a existência de analista de dados para prover as inferências corretas sobre os dados pesquisados. Neste cenário, o data mining funciona como uma espécie de lupa, produzindo uma visão macroscópica dos dados disponíveis, onde otimiza a percepção do pesquisador e traz à tona informações que, provavelmente, sem o uso da ferramenta seria impossível descobrir.

Na visão de Han e Kamber (2000, p.8), data mining é “uma etapa na descoberta do conhecimento em bancos de dados que consiste no processo de analisar grandes volumes de dados sob diferentes perspectivas, a fim de descobrir informações úteis que normalmente não estão sendo visíveis”. Ou seja, o DM é capaz de desvendar “informações escondidas” nos grandes mananciais de dados. Desta forma, grande predicado que a ferramenta pode oferecer é velocidade e robustez em analisar um volume expressivo de dados.

Ao longo dos anos, o data mining vem se tornando uma técnica multidisciplinar, visto que existem vários conhecimentos agregados na técnica propriamente dita, tais como: computação, estatística, recuperação de informação e visualização de dados. (MACHADO et al., 2007). Esta multidisciplinaridade e heterogeneidade de saberes aguçou nosso desejo em trazer a utilização do DM para estudo de identidades, uma vez a sua aplicação perpassa as

mais diversas áreas do conhecimento e tem de certo modo alcançando sucesso.

São vários os setores que trabalham com informação que utiliza a técnica do data mining para obter padrões válidos e potencialmente úteis em suas atividades. Na década de 90, por exemplo, ao procurar eventuais relações entre o volume de vendas e os dias da semana, um software de data mining apontou que, às sextas-feiras, as vendas de cerveja na rede Wal-Mart cresciam na mesma proporção que as de fraldas. Uma investigação mais detalhada revelou que, ao comprar fraldas para seus bebês, os pais aproveitavam para abastecer o estoque de cerveja para o fim de semana.

Já o Bank of América usou essas técnicas para selecionar entre seus 36 milhões de clientes aqueles com menor risco de dar “calote” no pagamento de um empréstimo. A partir desses relatórios, enviaram-se cartas oferecendo linhas de crédito para os correntistas cujos filhos tivessem entre 18 e 21 anos e que, portanto, precisassem de dinheiro para ajudar esses filhos a comprar o próprio carro, uma casa ou arcar com os gastos da faculdade. Resultado: em três anos, o banco lucrou 30 milhões de dólares (LIMA JUNIOR, 2006).

Na área da comunicação, de forma mais especial o jornalismo, tem iniciado a utilização de DM, ainda que de forma muito tímida. Neste sentido, Barbosa et. al (2007) esclarece que embora a mineração de dados seja amplamente discutida no campo da ciência da computação, o esforço de relacioná-la em aplicações no jornalismo é recente. Apesar de hodierno, já existem esforços pertinentes em prover aproximação de DM com a temática comunicacional. Em 2014, Cabral e Viana elaboraram um estudo sobre o uso do Data Mining e a descoberta de novos enquadramentos noticiosos: uma análise sobre o Programa Mais Médicos no G1.com e Estadão.com. A partir de técnicas de DM foi possível analisar em um curto intervalo de tempo aproximadamente trinta e duas mil notícias, das quais, com base nos dados coletados, foi possível inferir, por exemplo, qual o veículo teve uma maior cobertura, como se deu o caráter da notícia, ou seja, através da presença de certas palavras chaves tais como “Despreparados”, “Não Habilitados”, “Desqualificados”, foi possível perceber que os dois portais tinham certo “descontentamento” em relação à presença do médicos cubanos no país, uma vez que na grande parte da suas notícias, os referidos termos aparecem com muito frequência. Além disso, a mineração de dados proporcionou aos pesquisadores uma visão profunda e rápida acerca das quase trinta e duas mil notícias coletadas.

Analisar quais as temáticas mais comentadas no twitter dos piauienses; quem são os agentes que conduzem essas discussões?; Quais tweets tiveram maior repercussão? Como está

agrupado o perfil netnográfico destes agentes. Responder todos esses questionamentos com as ferramentas atuais de pesquisa em comunicação seria uma tarefa extremamente onerosa ou quiçá inexecuível em um intervalo curto de tempo.

Neste sentido, o data mining apresenta diversas vantagens acerca dos métodos tradicionais de pesquisa, a principal delas, na nossa percepção, é defendida por Thearling (1999 *apud* Castanheira, 2003, p. 12) quando pontua que a “mineração de dados difere de técnicas estatísticas porque ao invés de verificar padrões hipotéticos utiliza os próprios dados para descobrir tais padrões”. Ou seja, o DM busca identificar relações dispersas dentro da base de dados, que dificilmente seriam percebidas. O exemplo das fraldas e das cervejas, citado anteriormente, endossa esse diferencial, uma vez que na análise estatística este comportamento seria imperceptível. Ainda neste contexto, Thearling (1999 *apud* CASTANHEIRA, 2003) destaca que aproximadamente 5% de todas as relações podem ser encontradas por esses métodos estatísticos. O DM, por sua vez, pode descobrir outras relações anteriormente desconhecidas: os 95% restantes.

Dentro dessa acepção, onde se tem a análise estatística como fator indissociável no processo de compreensão de grandes volumes de dados, a mineração de dados apresenta outra vantagem significativa que é a facilidade na leitura das informações. Com o avanço das técnicas de DM, os algoritmos já disponibilizam os resultados de forma de fácil compreensão, desobrigando o pesquisador de possuir conhecimentos avançados na área de estatística.

Além disso, o refinamento na utilização de DM pode ainda oferecer ao pesquisador uma gama de técnicas que possibilitam análise mais acurada, oportunizando assim um leque maior de possibilidades de interpretações. Essas podem ser utilizadas em tarefas de classificação, *clustering*, associação, sumarização.

6.1 Classificação

A classificação consiste em analisar um conjunto de dados, com base em outros registros fornecidos, de maneira que a partir desse “aprendizado” seja possível classificar um novo registro como, por exemplo, em uma base de dados com tweets onde estes já estão classificados por temáticas: lugares, gírias, política, saúde. A técnica analisa os registros e em seguida é capaz de dizer qual temática o novo tweet pode se enquadrar. Em outras palavras, essa técnica é capaz de prever, através da análise dos registros antigos, tweets que ainda não

foram classificados. Outros exemplos que podem justificar a utilização da técnica de classificação são: determinar se uma transação em cartão de crédito é fraude, com base no comportamento de compra do cliente; identificar a existência de determinada enfermidade com base no histórico de doenças em uma região ou analisar se um grupo de pessoas pode desenvolver doenças com base no perfil epidemiológico da região.

6.2 Análise de Clustering

A análise de *clustering* visa identificar e aproximar registros. A referida técnica consiste em segmentar uma população com características homogêneas e pode ser aplicada mesmo sem conhecer como os dados estão distribuídos e, ainda assim, é possível descobrir estruturas escondidas. Esta técnica difere da classificação por não necessitar que os dados sejam previamente categorizados.

Na figura a seguir, temos a análise de *clustering* de determinado usuário do Facebook, o processo de clusterização foi construído a partir de duas características: quantidade de amigos em comum (representado pelo tamanho da circunferência, quanto mais amigos em comum, maior o diâmetro) e o local de trabalho/estudo dessas pessoas (representado pela cor). Como pode-se observar, a técnica segmentou em grupos trazendo resultados que possivelmente não seriam perceptíveis a “olho nu”.

análise, com base no perfil de compras, de quais seriam os produtos que estão relacionados aos seus hábitos de consumo. Na figura 9, temos um exemplo onde evidencia a utilização da referida técnica.

quem viu este produto acabou comprando ocultar ^

<p>50% comprou</p>  <p>iPhone 5c 8GB Branco Desbloqueado IOS 8 4G e Wi-Fi Câmera...</p> <p>★★★★★ (2)</p> <p>R\$ 1.499,00 10x de R\$ 149,90 sem juros</p> <p>cartão americanas.com até 12x de R\$ 124,92</p> <p>sul e sudeste 10% no boleto</p> <p>+ Apple</p>	<p>13% comprou</p>  <p>iPhone 4S Branco 8GB - Apple</p> <p>★★★★★ (419)</p> <p>de R\$ 1.699,00 por R\$ 1.099,00 10x de R\$ 109,90 sem juros</p> <p>cartão americanas.com até 12x de R\$ 91,58</p> <p>sul e sudeste 10% no boleto</p> <p>+ Apple</p>	<p>6% comprou</p>  <p>iPhone 5s 16GB Dourado Desbloqueado Câmera 8MP 4G e Wi-Fi...</p> <p>★★★★★ (99)</p> <p>de R\$ 2.799,00 por R\$ 2.499,00 10x de R\$ 249,90 sem juros</p> <p>cartão americanas.com até 12x de R\$ 208,25</p> <p>sul e sudeste 10% no boleto</p> <p>+ Apple</p>	<p>4% comprou</p>  <p>Smartphone Motorola Novo Moto G DTV Colors Dual Chip XT 1069...</p> <p>★★★★★ (1080)</p> <p>de R\$ 899,00 por R\$ 849,00 10x de R\$ 84,90 sem juros</p> <p>cartão americanas.com até 12x de R\$ 70,75</p> <p>sul e sudeste 10% no boleto</p> <p>+ Motorola</p>
---	--	--	---

Figura 9 - Emprego da técnica de associação
Fonte: Site americanas.com

Como podemos observar na figura 9, o site da americanas.com utiliza-se da técnica de associação para demonstrar aos seus clientes quais os produtos estão relacionados. No exemplo acima se tem que das pessoas que viram a oferta no site, 50% compraram uma iPhone 5c, 13% compraram um iPhone 4s, 6% compraram um iPhone 5s e apenas 4% compraram o aparelho moto G. Deste modo, a definição proposta por Berry (1997) foi confirmada uma vez que se percebeu a relação “se A, então B”, onde o “A” seriam as pessoas que visualizaram a oferta e “B” seria o resultado desta ação - a compra do aparelho.

No contexto dessa pesquisa (análise de identidades na rede social twitter), o pesquisador, utilizando apenas essa técnica, pode ser capaz de descobrir, por exemplo, qual o percentual de pessoas que reclamaram do calor do estado do Piauí e demonstraram o interesse em tomar cerveja e, de forma mais profunda, pontuar até mesmo o lugar preferido para se fazer o consumo desta bebida. Em notação matemática, o exemplo acima poderia ser representando da seguinte forma: “dos 12% de pessoas que reclamaram do calor no twitter, 30% demonstraram desejo de tomar uma cerveja, sendo que destes 5% preferem o bar

Cantinho do Jambo”.

Isto posto, percebe-se que o emprego da mineração de dados dá azo a uma nova dimensão em se trabalhar com dados em grande escala. A análise de excesso de informação deixa de ser um óbice para os pesquisadores, uma vez que o DM irá proporcionar novas perspectivas na realização de estudos, além de diminuir, de forma significativa, o tempo gasto no levantamento das análises.

Ademais, o avanço das técnicas e algoritmos do data mining tem sido pauta constante de diversas empresas que visam a todo instante desenvolver ferramentas cada vez mais intuitivas e que possibilitam aos profissionais que não são “nativos” da área ciência da computação a sua utilização. Na quadro abaixo, apresentamos um quadro resumo das principais ferramentas disponíveis hoje no mercado:

Ferramenta	Descrição	Site
IBM SPSS Modele	Ferramenta desenvolvida pela IBM, muito utilizada para desenvolver modelos preditivos além de outras análises	http://www-01.ibm.com/software/analytics/spss/products/modeler/
Oracle Data Mining (ODM)	Ferramenta de mineração de dados desenvolvida pela Oracle, para uso exclusivo em banco de dados Oracle	http://www.oracle.com/technetwork/database/options/advanced-analytics/odm/index.html
PIMIENTO	Platform Independent Text Mining Engine Tool. Ferramenta open source para mineração de textos	http://erabaki.ehu.es/jjga/pimiento
SAS Text Miner	Ferramenta de mineração de texto desenvolvido pela empresa SAAS	http://www.sas.com/en_us/software/analytics/text-miner.html
Solr	ferramenta de código aberto, usada para buscas em textos, cluster e integração com banco de dados	http://lucene.apache.org/solr/
WEKA	Ferramenta open source que oferece diversas possibilidades de análises, tais como: classificação, regressão, agrupamento, regra de associação e visualização	http://www.cs.waikato.ac.nz/ml/weka/

Quadro 3 - Principais ferramentas de data mining disponíveis no mercado

Fonte: Adaptado pelo autor

7. METODOLOGIA

Dada a proposta transmetodológica desta pesquisa, optou-se em verificar como ferramentas advindas de outras áreas do conhecimento poderiam ser aplicadas para realizar pesquisas em comunicação. Nesse sentido, adotamos técnicas de Data Mining na tentativa de se trabalhar com grandes volumes de dados, culminando na descoberta de padrões de marcas identitárias na rede social Twitter.

Deste modo, o presente estudo sintetiza os preceitos de uma pesquisa quali-quantitativa, visto que se pretende numerar, medir unidades e estabelecer relações. O trabalho também se configura como pesquisa fundamental, uma vez que se propõe a “aumentar a soma de saberes disponíveis, saberes esses que, em algum momento, [...] serão utilizados para a solução de problemas empíricos” (SANTAELLA, 2010, p. 89). Trata-se, ainda, de uma pesquisa analítica, pois se propõe a analisar dados e extrair deles conclusões, conforme explica Santaella (2010, p. 92).

Para atender aos objetivos propostos nesta pesquisa, utilizaremos o processo de DM estabelecido por Aranha (2007), onde o mesmo pontua os seguintes passos: coleta de dados, pré-processamento, indexação, mineração e análise, conforme descrito a seguir:



Figura 10 - Fases Data Mining

Fonte: Aranha, 2007

7.1 Coleta

A primeira fase da pesquisa consiste na coleta dos dados, que visa à montagem de um banco de dados a respeito do tema central da pesquisa, configurando-se como elemento básico

de qualquer processo que envolva a aplicação do data mining. O processo de coleta iniciou-se com a criação de um algoritmo para coletar quem seriam os participantes da análise. Para tanto, foi desenvolvido um mecanismo que pudesse acessar diretamente a base de dados do Twitter, onde a partir deste acesso foi possível extrair informações pessoais contidas na conta de cada usuário.

Este acesso deu-se através de uma API (*Application Programming Interface*), que na percepção de Freire (2013) tem como função basilar proporcionar interatividade quando do desenvolvimento de softwares através de códigos adaptáveis, uma espécie de *plugin* por meio do qual é possível enviar e receber requisições, facilitando a troca de dados.

Com acesso aos dados da base do Twitter, buscamos todos os usuários que preencheram na sua descrição que eram do Estado do Piauí, contabilizando 8.112 (oito mil cento e doze) usuários atenderam a esta especificação. Desse montante, requisitamos os seguintes dados: nome, descrição da bio¹⁹, quantidade de tweets, quantidade de seguidores, número de usuários que o seguem, além da data da criação da conta na referida rede.

Identificados quem eram os usuários, achamos prudente realizar um filtro elegendo como atores da pesquisa as contas do Twitter que tinham acima de 500 seguidores. Este filtro se deu pelo fato de partirmos do entendimento de que estas contas tinham uma audiência representativa, tendo maior abrangência na rede social citada. Partindo dessa triagem obtivemos 825 contas a serem analisadas.

Em face destas 825 contas, iniciou-se o processo de armazenamento dos tweets propriamente dito, no qual foram coletados todos os tweets, do ano de 2014. Assim conseguimos estruturar uma base de dados com 671.366 tweets. Neste contexto, é oportuno destacar que além do tweets, também foi armazenada a data de realização do mesmo, bem como a quantidade de retweets que a referida postagem sofreu no micro blog. Por se tratar de um manancial de dados razoável, todas as informações foram armazenados em um sistema de gerencialmente de banco de dados (SGBD). Dentre os inúmeros SGBD disponíveis no mercado optamos pelo mysql²⁰, que é gratuito. A figura abaixo resume todo o processo de coleta de dados.

¹⁹ Campo onde o usuário descreve a sua biografia

²⁰ sistema de gerenciamento de banco de dados gratuito



Figura 11 - Processo de Coleta de dados

Fonte: Criado pelo autor

7.2 Pré-processamento

Uma vez realizada a coleta de dados, o próximo passo foi a preparação dos tweets para que os mesmos pudessem ser manipulados pelos algoritmos de mineração de dados. Sistemas de *data mining* não submetem aos seus algoritmos de descoberta de conhecimento coleções de textos despreparadas (GOMES, 2008), neste sentido fez-se necessário um processamento prévio da base de dados.

Nesta etapa, os dados armazenados no banco de dados foram submetidos a inúmeras operações capazes de obter uma forma de representá-los estruturadamente. Para Soares (2009, p. 89),

O primeiro passo de uma operação de Pré-processamento é a *tokenização* ou atomização e sua execução tem como finalidade seccionar um documento textual em unidades mínimas, mas, que exprimam a mesma semântica original do texto. O termo token é utilizado para designar estas unidades, que em muitas vezes correspondem a somente uma palavra do texto, porém, nem sempre estas unidades textuais não podem ser consideradas palavras ou apresentam mais de uma palavra: “21/10/2007”, “PM”, “R\$100,00” e “couve-flor”.

O processo de tokenização é auxiliado pelo fato de as palavras serem separadas por caracteres de controle de arquivo ou de formatação, tais como espaços ou sinais de pontuação, que em alguns casos podem ser considerados tokens delimitadores (FELDMAN & SANGER, 2007).

Buscando sempre tornar possível o processamento computacional de textos, uma vez realizado o processo de tokenização, o passo seguinte é a identificação do que pode ser desconsiderado nos passos posteriores do processamento dos dados. É a tentativa de retirar tudo que não constitui conhecimento nos textos (SOARES, 2009, p. 92).

Em um documento, existem muitos *tokens* que não possuem valor semântico, sendo úteis apenas para o entendimento e compreensão geral do texto. Estes tokens são palavras classificadas como *stop words* e fazem parte do que é chamado de *stoplist* de um sistema de Mineração de Textos (BASTOS, 2006). Geralmente, fazem parte de uma *stoplist* elementos como conjunções, preposições, pronomes e artigos, pois são considerados termos de menor relevância, ou seja, sua presença pouco contribui para a determinação do valor semântico de um documento.

Uma *stoplist* bem elaborada permite a eliminação de muitos termos irrelevantes, tornando mais eficiente o resultado obtido pelo processo de DM. Normalmente, 40 a 50% do total de palavras de um texto são removidas com uma *stoplist* (SILVA A. A., 2007).

7.3 Indexação

A indexação foi o processo responsável pela criação de estruturas auxiliares denominadas índices e que garantiram rapidez e agilidade na recuperação dos twittes. De acordo com Soares (2009, p. 99),

Técnicas de indexação de documentos foram bastante difundidas pela demanda e crescimento da área de Recuperação de Informação desde a década de sessenta. Contudo, muitas pessoas acreditam que esta é uma área nova. Esta ideia talvez tenha surgido com a grande popularização das máquinas de buscas que tornaram possível a pesquisa do conteúdo de páginas web, ou seja, documentos textuais.

Com a indexação foi possível estabelecer ligações entre termos que não estavam necessariamente bem definidas dentro da estruturação dos twittes. Para realizar o processo de classificação dos agentes da pesquisa, a indexação foi primordial, uma vez que não se tinha de forma clara qual o perfil dos usuários do Twitter. Como exemplo, podemos citar um determinado usuário que coloca na sua descrição os seguintes termos: “35 anos, evangélica fervorosa, fiel e temente a Deus”. Em face destes termos, pudemos classificar esse usuário na categoria religiosa, entretanto, para fazê-lo de forma automática foi necessária uma consulta a um dicionário de termos, sendo que a sua função é simples: após receber um conjunto de palavras encontradas no twitte, realiza-se uma consulta em uma base de dados e indica ao

indexador o termo correto a ser utilizado na indexação da palavra recebida. Ou seja, com a presença dos termos “evangélica” e “temente a Deus” foi possível indexá-la na esfera religiosa.

Outro exemplo do emprego da indexação foi para mapear as principais profissões dos usuários. A partir de um dicionário colhido com base na classificação brasileira de ocupação (CBO) foi possível inferir qual a ocupação do usuário, sendo realizado da seguinte forma: tinha-se a seguinte descrição “ginecologista, casado e pai do João”, todos estes termos foram submetidos a uma consulta no dicionário da CBO e o resultado foi a equivalência do termo “ginecologista”. Deste modo, foi factível perceber, com base na descrição da bio, que o referido usuário era médico. Neste contexto, Soares (2009, p. 101) advoga que

A utilização desta técnica, além de tornar o índice mais compacto, permite a localização de documentos grafados de forma diferente, mas que apresentam mesmo valor semântico. Porém, a maior dificuldade para a utilização deste mecanismo reside na criação do próprio dicionário.

Ademais, esta técnica oportunizou agilidade na busca dos dados, visto que com os índices criados, o processo de mineração fica mais prática. Por conseguinte, a ferramenta tecnológica utilizada para realizar o processo de indexação foi a Solr, um projeto *Open Source* de um servidor de buscas de alta performance para indexação e busca.

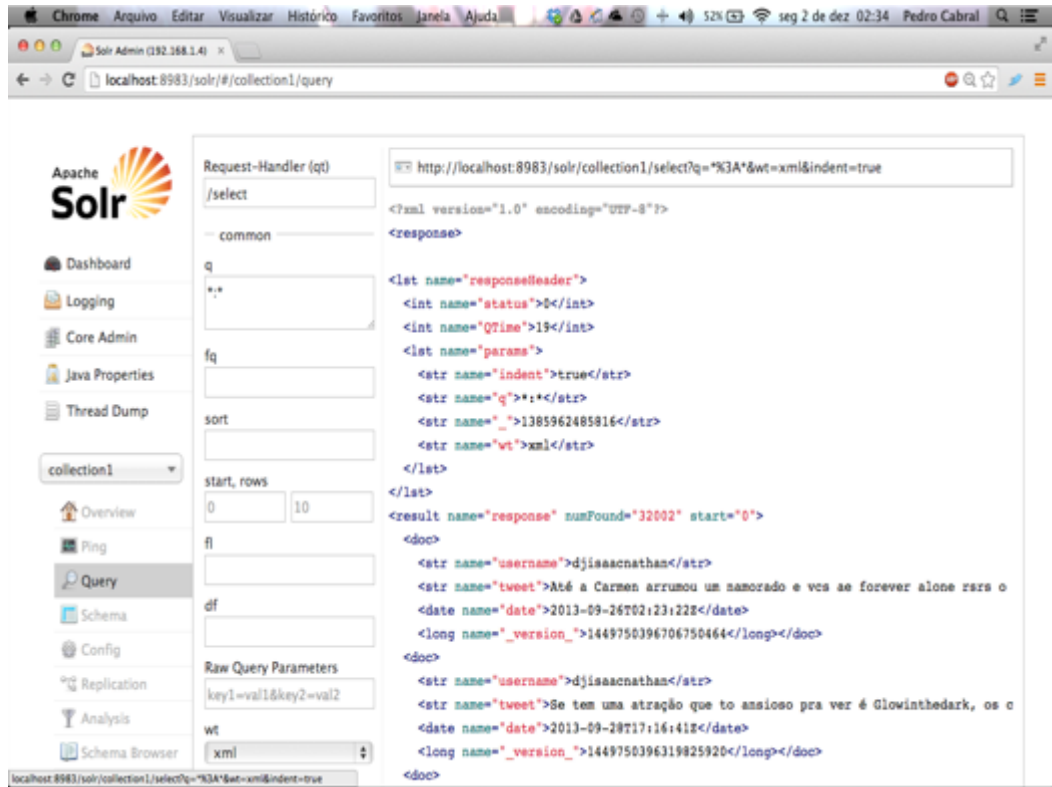


Figura 12 - Ferramenta de indexação de conteúdo Solr

Fonte: Criado pelo autor

7.4 Mineração

A quarta etapa foi a mineração de dados, que teve como finalidade buscar conhecimentos novos e úteis a partir dos dados coletados. Esta fase compreendeu basicamente a aplicação de algoritmos específicos e amplamente utilizados dentro do processo de DM. Diante das várias técnicas para se realizar mineração de dados, optamos por empregar classificação, *cluster* e associação, técnicas estas já abordadas anteriormente.

Nesse ímpeto, é apropriado pontuar que a escolha dessas técnicas deu-se pela tentativa de dar azo à proposta trazida neste trabalho, que é analisar marcas identitárias na rede social Twitter.

Portanto, para realizar a aplicação das técnicas supracitadas, adotamos a ferramenta Weka²¹, por acreditar que a mesma possui um arcabouço com boa parte das técnicas necessárias para realizar data mining, além disso, o Weka apresenta, na nossa concepção, um design fácil e intuitivo.

²¹ <http://www.cs.waikato.ac.nz/ml/weka/>.

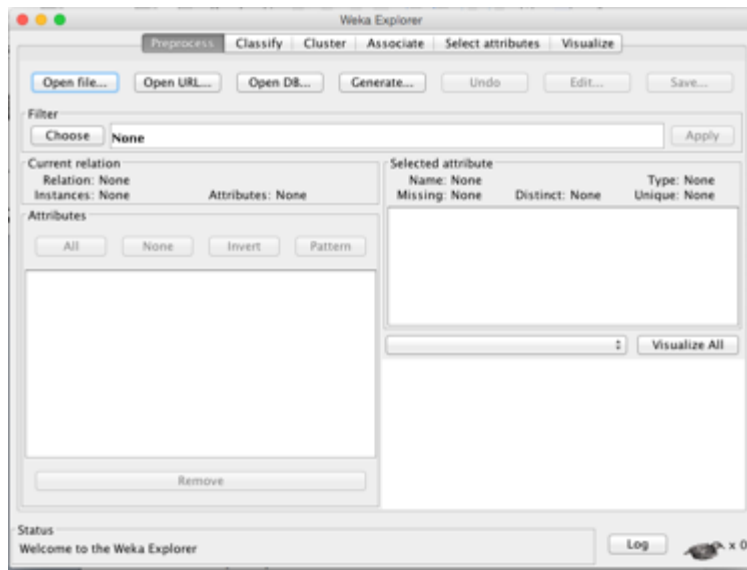


Figura 13– Interface Weka

Fonte: Weka

7.5 Análise

Análise é a última etapa do processo e também uma das mais trabalhosas. A análise consistiu em extrair, a partir das informações geradas com o data mining, *insights* que pudessem permitir entendimento mais acurado acerca do manancial de dados coletados.

Para a concepção do processo de análise deste trabalho vislumbramos quatro vertentes, a primeira foi entender quem eram os 825 agentes formadores da pesquisa. Esse entendimento deu-se por perceber quais eram as suas ocupações, quanto tempo estavam na rede social, quantos seguidores tinham, além da quantidade de retwites que suas postagens recebiam. Neste sentido, construímos uma análise estatística, que foi oportunizada graças as técnicas de classificação e clusterização.

A segunda parte da análise centrou-se em investigar quais eram as temáticas mais recorrentes, para tanto levou-se em consideração a presença das *hashtags*. A partir desta vertente, foi possível perceber como foi segregada a propagação destes termos ao longo do ano de 2014. A cada mês, foram levantados quais eram os termos mais comentados no Twitter e se existia uniformidade na quantidade dos termos ao longo dos meses. Por exemplo, o termo *#calor* é presente durante todo o ano ou só nos meses mais quentes do estado do Piauí? E o fato da sua presença ser marcante pode-se inferir que o piauiense é um povo que reclama recorrentemente dos fatores climáticos?

A terceira vertente da análise foi perceber através do emprego das técnicas de associação, se existe alguma relação entre estes termos, ou seja, uma *hashtag* pode ser associada a outra, sob um viés estatístico. Isto é, das pessoas que reclamam do calor quantos por centos associam essa reclamação com uma vontade de tomar uma cerveja gelada? Para a elaboração desta investigação, todas as *hashtags* foram submetidas a uma técnica específica de DM que a partir do seu emprego foi possível estabelecer se existe ou não temáticas relacionadas no Twitter.

A quarta e última vertente analisou se estas temáticas estão associadas especificamente a determinado grupo de usuários, questões como quem comenta mais sobre cervejas no Twitter? Médicos, professores, jornalistas? Para construção dessa análise recorreremos mais uma vez a técnicas de cluster.

8. ANÁLISE DOS RESULTADOS

De posse dos dados, iniciamos a análise na tentativa de ponderar como era constituído o perfil dos 825 usuários selecionados para elaboração deste trabalho. Ao longo da prospecção dos insumos percebeu-se que em 48% dos casos não foi possível visualizar informações mais detalhadas sobre estes usuários. Essa imprecisão na análise deu-se, basicamente, por dois motivos: ou o usuário deixava o campo bio vazio ou colocava expressões que eram impossíveis de descobrir qual era a sua identidade. A figura abaixo, apresenta um exemplo onde não foi possível identificar o proprietário da conta.



Figura 14 - Conta do usuário do Twitter sem descrição da bio

Fonte: Twitter

Como se pode observar na figura 14, a usuária @danusesantiago fez questão de não pontuar algo que pudesse identificá-la, essa atitude precede uma das facetas suscitadas pelo ciberespaço: a possibilidade do anonimato. Neste contexto, a internet se consubstancia como um ambiente em que se pode acompanhar várias interações e em muitos casos até participar das mesmas, contudo a prerrogativa do anonimato pode ser mantida. Neste sentido, tal prerrogativa não é tão comum em redes sociais na internet, haja vista que tais redes pressupõem a ideia de criação de novos laços e esses laços, na maioria, são construídos a partir de um processo mútuo de auto revelação, ou seja, as RS favorecem a criação de um ambiente onde os indivíduos tentam a todo instante desenvolver mecanismos que possam definir seus *hobbies*, preferências pessoais, seja através de fotografias ou postagens sobre o seu cotidiano.

Esta atenção ao ato de se definir pode ser encarado como traço de diferenciação face a pluralidade de identidades que as redes sociais aventam, ao passo que, no nosso entendimento, quanto mais singular for essa identificação, maior a probabilidade de se estabelecer novas

conexões. Tal elucubração deu-se por perceber, dentro do objeto pesquisado, que os indivíduos que descrevem a sua biografia apresentam, em média, uma incidência de 94% a mais no número de seguidores em comparação àqueles usuários que optaram por deixar o campo bio vazio.

Dentro do prisma no qual a identificação é relacionada às ocupações, flagrou-se que as profissões que apareceram de forma mais proeminente tinham como característica o aspecto de serem formadores de opinião. Tal premissa pode ser alicerçada pelo fato da quantidade de seguidores que cada profissão apresenta. Jornalistas, professores, advogados, médicos, economistas foram as 5 ocupações que obtiveram quantidade mais expressiva na amostra utilizada na presente pesquisa, destacando-se os profissionais da Comunicação, tendo em vista que o Twitter passou a figurar na rotina produtiva desses profissionais como fonte de pautas, bem como meio de interlocução com entrevistados, entre outros fatores. O fato desses profissionais debaterem temas relevantes do cotidiano e, por isso, de interesse público, acaba por atrair maior número de seguidores, como atesta a tabela abaixo.

Profissão	Quantidade de seguidores
Jornalista	51.161
Professor	47.399
Advogado	15.241
Médico	13.625
Publicitário	7.585
Fotografo	7.432
Contador	6.163
Administrador	6.030

Tabela 4 Quantidade de seguidores por profissão

Fonte: Adaptação do próprio autor

Em face desta conjuntura, é oportuno salientar que graças ao emprego do processo de DM, em poucos segundos foi possível estabelecer um “censo” a respeito das profissões e aliado a isso a quantidade de seguidores para cada segmento. Face a este contexto, onde necessita-se a realização de estudos partindo do agrupamento de grandes volumes de dados, o data mining atendeu a contento esta especificidade. Outrossim, foi possível perceber, também, uma característica que não estava muito óbvia, estando implícita entre os dados: os profissionais liberais apresentam grande audiência dentro do Twitter dos piauienses.

Na tentativa de elucidar quais eram as temáticas que estes profissionais pautaram e assim inferir quais termos eram afetos a estes usuários do microblog, mais uma vez recorreu-

se às técnicas de mineração de dados para ponderar a incidência das *hashtags*. Neste sentido, optamos por levantar, através de estudo quantitativo, quais as *hashtags* eram mais utilizadas por estes profissionais e se as mesmas eram empregadas de forma recorrente ao longo do ano. Como resultado percebemos que os advogados utilizaram 6.906 *hashtags* ao longo de 2014, esse excesso de termos coaduna, no nosso entendimento, com um traço identitário reservado aos advogados: a forma verbosa de se comunicar.

Os jornalistas, por sua vez, ocuparam o segundo lugar em quantidade de *hashtag*, tendo aproximadamente 50% a menos de termos utilizados em relação aos advogados. Estes termos, em sua maioria, faziam alusão a pautas políticas ou menções aos debates. Das 3.638 *hashtag* 18% foram relacionadas aos debates realizados pelas emissoras de TV.

Com 1.701 *hashtags*, os publicitários atingiram o terceiro lugar em suas postagens e também foi possível perceber marcas identitárias acerca da sua ocupação. A utilização do termo #ad, fazendo alusão a *advertising* (atividade de atrair atenção do público), esteve presente em 20% das postagens.

Os médicos, por conseguinte, não utilizaram termos associados ao seu ofício, das 1.490 postagens nenhuma fazia alusão a medidas de profilaxia ou de cuidados com a saúde. Seus comentários na rede social Twitter eram mais relacionados às temáticas sobre o pleito eleitoral. Assim como os médicos, os contadores, administradores e professores não explicitaram em suas contas no Twitter questões sobre a suas rotinas de trabalho, grande parte de suas *hashtags* eram associadas a política.

Na quinta posição, os fotógrafos imprimiram 1.098 *hashtags*, nas quais pontuavam expressões relacionadas às suas atribuições. #Ensaio, #art, #photography foram recorrentes em suas postagens. O gráfico abaixo resume a quantidade de termos usada por cada profissional:

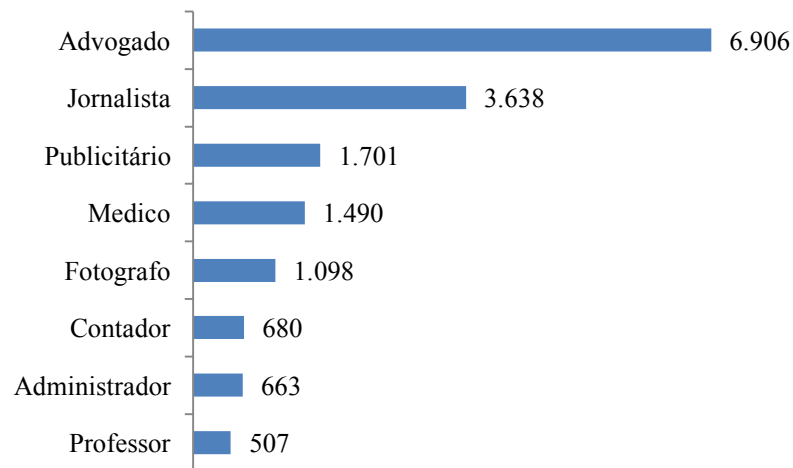


Gráfico 1 - Quantidade de hashtag, por profissão

Isto posto, com a utilização do DM foi possível inferir que muitas das *hashtags* utilizadas apresentavam certo valor identitário, sendo possível estabelecer a partir do seu contexto de uso quem eram os agentes que as utilizavam. Nesta acepção, buscamos sedimentar essa afirmação mapeando estas *hashtags* e procurando visualizar se as mesmas eram utilizadas por diferentes profissionais, ou seja, inferir se um dado termo era comum a todas as profissões. Como produto deste mapeamento, percebemos que nenhuma *hashtags* foi comum a todas as profissões ao longo do ano.

No que tange à aplicação de data mining para prover análise estatística acerca de um conjunto de dados com características temporais, a técnica de cluster mostrou-se a mais apropriada para responder possíveis questionamentos relacionados a esta temática. Visto que, dentro da presente pesquisa procurou-se por detalhar como estava organizada a adesão dos usuários do microblog, tendo como base a variável tempo, ou seja, o estudo pautou-se por detalhar quantas adesões foram realizadas por ano. Com o desfecho deste experimento foi possível perceber que, da amostra pesquisada, 34% dos usuários realizaram sua adesão no ano 2009, enquanto no ano anterior, 2008, a taxa de adesão foi de apenas 2%, conforme demonstra tabela abaixo:

Ano	Qtde de Adesões	%
2008	15	2%
2009	277	34%
2010	211	26%
2011	163	20%
2012	74	9%
2013	53	6%
2014	32	4%

Tabela 5 - Quantidade de adesão de usuários na rede social Twitter, por tempo

Fonte: Adaptação do autor

A referida técnica ainda possibilitou descobrir que nos anos de 2010 a 2013 o mês com menor taxa de adesão foi novembro. Entretanto, não foi possível perceber razões concretas que justificassem tal fato. Diante desta premissa, algumas vezes o emprego de DM pode propiciar a descoberta de padrões em que o pesquisador não consegue fundamentar seus achados, todavia não se deve abandonar tais evidências, visto que os mesmos possuem relevância estatística.

Diametralmente oposto a outras redes sociais, o Twitter não requisita dados mais pessoais sobre seus usuários, atributos como gênero, data de nascimento, bem como preferências do proprietário da conta. Face a esta particularidade da rede social em questão não foi possível estabelecer um aprofundamento destes dados e assim dispor de estudo mais acurado, tendo como base a construção de um perfil netnográfico.

No que se refere à análise orquestrada a partir do emprego de *hashtags*, notou-se que durante todo o ano de 2014 os proprietários de contas do Twitter usaram 36.873 *hashtags*, tendo as mesmas diversas temáticas. A média mensal de utilização dos indexadores de conteúdo foi de 3.073, tendo o mês de setembro um ligeiro aumento face aos demais meses. Nesta acepção, o referido acréscimo foi decorrente dos debates que antecederam o pleito eleitoral, que ocorreu no início do mês de outubro de 2014. A tabela abaixo assinala a quantidade de hashtag utilizada ao longo do ano.

Mês	Qtde. de hashtag ²²	%
janeiro	3.375	9,15%
fevereiro	3.351	9,09%
março	3.235	8,77%
abril	3.038	8,24%
maio	2.832	7,68%
junho	2.879	7,81%
julho	3.197	8,67%
agosto	3.238	8,78%
setembro	3.379	9,16%
outubro	3.151	8,55%
novembro	2.607	7,07%
dezembro	2.591	7,03%
Total	36.873	100,00%

Tabela 6 Quantidade de hashtag, por mês

Inobstante a esta quantidade de termos, é interessante ressaltar que uma *hashtag* pode aparecer mais de uma vez no mesmo mês, ou seja, esta quantidade não se refere a termos únicos, sem repetição. Neste contexto, é oportuno frisar que diante do montante de usuários dispostos na rede, o seu uso não é exclusivo, podendo assim, mais de um usuário indexar o mesmo conteúdo através da mesma *hashtag*.

Entretanto, em análise mais ampla, os dados demonstraram que estas *hashtags* não foram recorrentes ao longo do ano, ou seja, poucos termos se mostraram frequentes no período de janeiro a dezembro. Sendo que destas, apenas 56 apareceram nos doze meses do ano, conforme demonstra a *cloud tag*²³ abaixo:

²² Para efeito de cálculo foi levado em consideração a quantidade de vezes que cada termo aparecia na pesquisa, desprezando-se, assim, se os mesmos eram repetidos ou não.

²³ Ferramenta de visualização de dados que tem como premissa facilitar a análise de dados através de uma disposição gráfica, no exemplo acima tem-se que o os termos #ad, #timbeta e #rt dispostos com maior fonte. Isso significa que os mesmos tiveram frequência mais acentuada.

Contudo, em uma análise mais ampla, sem levar em consideração a incidência dos temas em todos os meses percebeu-se que muito embora os termos encontrados indicassem alguns acontecimentos específicos ao Estado, os piauienses que utilizaram a rede social, deram prioridade de forma significativa às discussões em torno de acontecimentos nacionais, com destaque para programas inseridos na grade televisiva da Rede Globo de Televisão, tais como as telenovelas Amor à Vida e Em Família, além do reality show Big Brother Brasil.

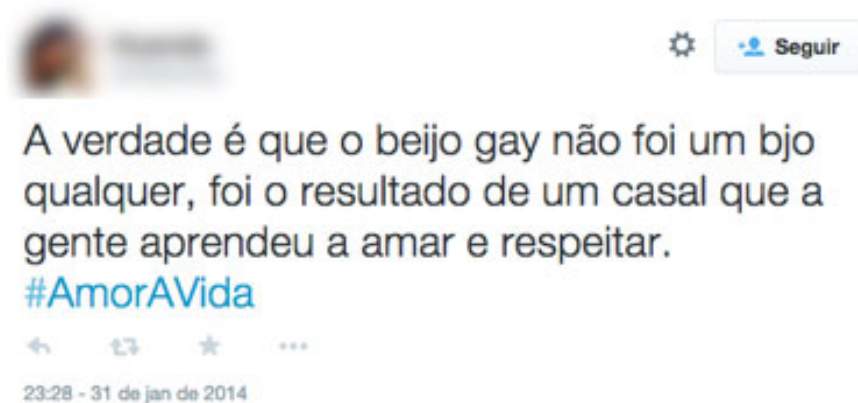


Figura 17 - Exemplo de twitte sobre temática nacional

Fonte: Twitter, 2014

Ações promocionais promovidas por grandes empresas, em especial a operadora de telefonia móvel TIM, repercutem em todos os meses devido ao interesse dos usuários em adquirir um chip promocional denominado de Tim Beta. A referência a artistas nacionais e internacionais também faz parte da massa de tweets coletados, que em geral trazem menções à estreia de filmes e espetáculos, muito embora também haja um grande volume de críticas e sátiras em relação às personalidades da mídia.

Visando pontuar um entendimento dos principais acontecimentos que pautaram as discussões no Twitter em relação a acontecimentos no Estado bem como temas circunscritos à cultura local, as *hashtags* serão explanadas de forma individual, corroborando com o contexto em questão. Após a análise efetuada em cada mês do ano, pudemos perceber que alguns eventos de caráter local ganham destaque, tais como o Corso de Teresina e as festividades pelos 162 anos da capital piauiense, assim como as ações de marketing político em decorrência das eleições realizadas em 2014 nos perfis de candidatos a cargos eletivos.

#curso

Criado por populares na capital piauiense, o Corso de Teresina, que consiste num desfile de carros enfeitados, bem como na reunião de foliões fantasiados numa das principais avenidas da cidade de Teresina, foi ao longo dos anos se tornando um evento de grande magnitude, atraindo foliões de várias partes do país, que acabaram por contribuir para que o evento se tornasse o maior do mundo, segundo os parâmetros do Guinness Book ainda em 2012. A *hashtag* #curso foi mencionada 201 vezes ao longo do mês de fevereiro, com maior incidência no dia da realização do evento, tendo em vista a divulgação de fotos dos participantes em sua rede social.

Desde o ano de 2012, a manifestação carnavalesca se tornou tema recorrente no mês de fevereiro, fazendo alusão à grandiosidade da festa, a reunião de foliões, fantasias irreverentes, além de outras peculiaridades inerentes à manifestação cultural. Devido o evento ter ganhado grandes proporções, é possível perceber entre os usuários da rede um ‘apego’, além de certo orgulho em relação à festa. É válido ressaltar que boa parte das discussões em torno dos acontecimentos que envolveram o Corso partiram dos perfis de usuários que se identificaram como jornalistas.

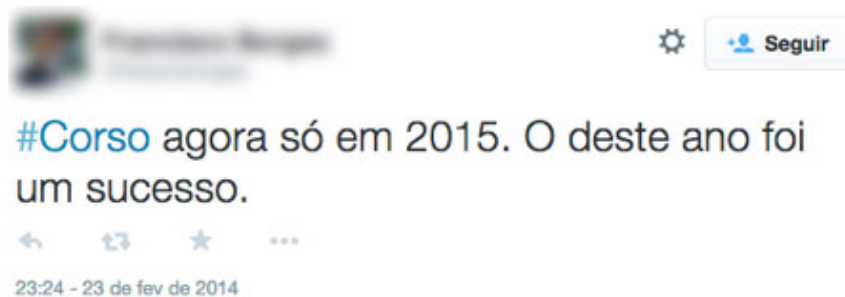


Figure 18 - Twitte sobre o #curso

Fonte: Twitter

#sosuespi

O movimento SOS UESPI fez alusão às manifestações de professores e alunos piauienses em relação às condições estruturais dos polos da Universidade Estadual do Piauí, bem como à falta de professores para o completo andamento do ano letivo de 2014. Além de ações presenciais, na maioria das vezes em frente a órgãos públicos, os manifestantes

buscaram visibilidade por meio das redes sociais, em especial o Twitter, chegando a constar nos Trend Toppics, que são os assuntos mais comentados do dia na rede social citada. Nas postagens com a *hashtag* #sosuespi, os usuários fazem alusão ao descaso do Governo do Estado com a educação superior, reivindicando melhorias em vários segmentos da universidade, com destaque para a convocação de novos professores. As discussões em torno da manifestação foram encontradas de forma mais acentuada no mês de maio, muito embora seja discutido no decorrer do ano.



Figure 19 - Twitte sobre o #SOSUESPI

Fonte: Twitter

#teamwilsonpi #piauinocoracao

Ano de eleição, 2014 foi um período movimentado para os candidatos a cargos eletivos nas redes sociais, tendo em vista que a Web passou a ser mais um cenário de embates políticos ente militantes e cidadãos comuns. Muitas *hashtags* podem ser vistas, em especial aquelas que fazem menção ao nome e ao número dos candidatos. Em alguns casos, como o candidato a senador Wilson Martins (PSB), ao seu nome atrelava-se as iniciais PI (Piauí), assim como o nome da coligação da qual fazia parte, intitulada de Piauí no Coração.

Tais manifestações reforçam a ideia de valorização do Estado, bem como o orgulho de ser desta terra, reforçando a identidade piauiense atrelada ao candidato. Proeminente entre os meses de junho até setembro, mês que antecedeu a realização do pleito eleitoral, as menções de candidatos a municípios piauienses, bem como ao Piauí foram bastante significativas. No entanto, muito embora haja o uso expressivo de palavras que denotem a piauiensidade dos usuários, classificamos tais manifestações como estratégia dos

candidatos para chamar a atenção dos eleitores, não se configurando, portanto, como uma ação espontânea do grupo político na referida época.



Figura 20 - Twitte sobre o #teamwilsonpi

Fonte: Twitter, 2014

#teresina162 #theamo #162anos

As comemorações pelos 162 anos de Teresina, no mês de agosto, figuram entre os temas debatidos pelos usuários piauienses do Twitter, ganhando densidade por meio de três *hashtags* distintas, sendo elas: #teresina162, #theamo e #162anos. Nas postagens, destaca-se o amor dos cidadãos pela cidade, os cartões postais mais conhecidos da capital piauiense, muitas vezes acrescidos de fotografias, além de mensagens sobre a realização de eventos nos mais diversos bairros de Teresina, na sua maioria promovidos pela Prefeitura e destacados por ela no perfil organizacional do órgão. A *hashtag* #theamo, por exemplo, é uma alusão ao termo THE, tendo em vista que o nome Teresina, em outras épocas, já utilizou estas iniciais. Além do mais, o TheAmo também se tornou um monumento local, sendo reproduzido em um letreiro no parque Potycabana – um dos pontos turísticos da capital do estado.

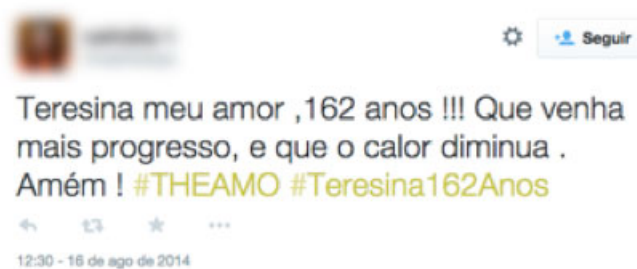


Figura 21 - Twitte sobre o #THEAMO

#Salverainha

O movimento cultural Salve Rainha surgiu no último trimestre de 2014, com o intuito de proporcionar aos piauienses espaço gratuito, reservado a manifestações artísticas de diferentes naturezas, tais como pinturas, música e outras artes plásticas, que geralmente não figuram nos espaços tradicionais voltados às expressões artísticas. Realizado na região central de Teresina, sempre aos domingos, o evento encontra nas redes sociais uma via para divulgação das edições realizadas, tendo em vista o seu caráter alternativo.

Dentre as 33 citações com a hastag #salverainha coletadas pela presente pesquisa, pode-se perceber que os usuários faziam menção à organização do evento e às atrações reservadas para o dia, além de fazer conexões com amigos para que esses se fizessem presentes no local.



Figura 22 - Twitte sobre o #salverainha

Fonte: Twitter

#diadopiaui

O dia 19 de outubro marca, oficialmente, as comemorações pelo Dia do Piauí, sendo lembrado pelo Governo do Estado com uma série de ações nas cidades de Teresina, Parnaíba e Oeiras, tendo em vista que as duas últimas figuraram como importantes campos de batalha em prol da independência local. A data é lembrada pelos usuários piauienses do Twitter, que manifestam o seu apreço pelo Estado, desejando felicitações, bem como relatando suas preferências em relação a locais para visitaç o. Na oportunidade, o Piauí estava completando 192 anos. Em momentos comemorativos, percebe-se que os sentimentos dos usu rios afloram em rela o   sua terra natal, destacando-se no teor das mensagens o v nculo identit rio dos usu rios.



Figure 23 - Twitte sobre o #diadoPiauí

Fonte: Twitter

#river

No m s de maio de 2014, o River Atl tico Clube se tornou campe o do segundo turno do campeonato piauiense e voltou a disputar uma decis o do estadual. Para o jogo, o est dio de futebol Albert o, localizado em Teresina, foi reaberto ap s dois anos sem a realiza o de partidas oficiais. O Galo Tricolor (como   conhecido pelos piauienses e torcedores do clube) encerrou um jejum de sete anos e voltou a comemorar um t tulo contra o Pia i Esporte Clube. Nas postagens coletadas para esta pesquisa, constatou-se que o clube representa um dos s mbolos identit rios do Estado, sendo exaltado pelos usu rios em suas postagens. O fato de o

estádio estar cheio para uma partida do campeonato piauiense, bem como fatos marcantes ao embate entre os times pontuaram as postagens analisadas.



Figura 24 - Twitte sobre o #river

Fonte: Twitter

#salipi

O Salão do Livro do Piauí (SALIPI) é realizado anualmente desde 2003, integrando o circuito cultural das principais feiras e bienais de livros do Brasil. O evento acontece sempre no mês de junho, em Teresina, e tem duração de uma semana, reunindo publicações variadas oriundas de diversas partes do mundo. Com a *hashtag* #12salipi, em alusão à décima segunda edição do evento no Estado, o salão do livro é mencionado pelos usuários do Twitter fazendo referência à compra de livros ou mesmo à participação no evento.



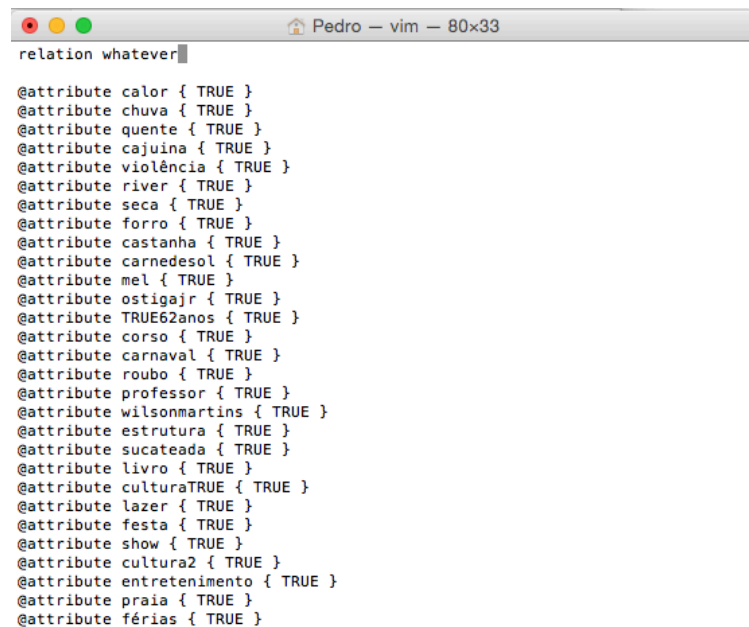
Figura 25 - Twitte sobre o #salipi

Fonte: Twitter

Feita esta breve explanação sobre as *hashtags* que apresentavam marca identitária relacionada ao Piauí, a outra parte do nosso estudo visou experimentar a aplicação da técnica

de associação. Técnica esta que, como explicitado no capítulo sobre data mining, consiste em procurar correlações entre os termos através de um viés estatístico, ou seja, mensurar de forma percentual a associação de uma expressão à outra.

Um grande óbice enfrentado nesta análise foi a estruturação do layout com as especificações necessárias para rodar o processo de DM. Em geral, ferramentas de data mining têm padrão no que diz respeito ao modelo de arquivo a ser empregado, neste caso específico, usamos a ferramenta WEKA, que necessita de uma modelagem do tipo Attribute-Relation File Format (ARFF).



```

relation whatever

@attribute calor { TRUE }
@attribute chuva { TRUE }
@attribute quente { TRUE }
@attribute cajuina { TRUE }
@attribute violência { TRUE }
@attribute river { TRUE }
@attribute seca { TRUE }
@attribute forro { TRUE }
@attribute castanha { TRUE }
@attribute carnedesol { TRUE }
@attribute mel { TRUE }
@attribute ostigajr { TRUE }
@attribute TRUE62anos { TRUE }
@attribute corso { TRUE }
@attribute carnaval { TRUE }
@attribute roubo { TRUE }
@attribute professor { TRUE }
@attribute wilsonmartins { TRUE }
@attribute estrutura { TRUE }
@attribute sucateada { TRUE }
@attribute livro { TRUE }
@attribute culturaTRUE { TRUE }
@attribute lazer { TRUE }
@attribute festa { TRUE }
@attribute show { TRUE }
@attribute cultura2 { TRUE }
@attribute entretenimento { TRUE }
@attribute praia { TRUE }
@attribute férias { TRUE }

```

Figura 26 - Exemplo de arquivo Attribute-Relation File Format (ARFF)

Fonte: Reproduzido pelo autor

Para sanar esse obstáculo e otimizar o tempo na geração do arquivo .arff, buscamos sites na internet que pudessem automatizar esse processo. Das várias opções encontradas, optamos por usar o site CSV2ARFF²⁴. O site gerou o arquivo no formato ideal e assim foi possível iniciar o processo de associação.

Entretanto, este experimento foi muito moroso e requisitou do computador uma capacidade razoável de processamento, visto que, para executar a análise, o algoritmo percorreu todas as 36.873 *hashtags* buscando calcular se existiam correlações entre elas. Dos resultados encontrados, notou-se que entre as *hashtags* havia baixa correlação, ou seja, os usuários dificilmente utilizavam na mesma postagem duas *hashtags* e quando o faziam a sua

²⁴ Disponível em <http://ikuz.eu/csv2arff/>

associação tinha baixo valor estatístico. Diante deste contexto, nos deparamos com uma dúvida acerca da eficiência da mineração de dados, visto que, neste momento, não obtivemos uma amostra significativa sobre a associação das *hashtags*.

Face a esse embate, resolvemos aplicar mais uma vez a técnica de associação, porém dessa vez optamos por selecionar algumas expressões contidas dentro dos twittes e não as *hashtags* propriamente ditas. Além disso, segregamos as associações por profissões, isto é, dimensionamos a análise de modo a perceber como determinados profissionais associam os termos nas suas postagens. O resultado encontrado foi:

Para os administradores:

- Todas as vezes em que o termo “festa” apareceu, em 33% o termo “estrutura” estava associado;
- Todas as vezes em que o termo “festa” apareceu, em 11% o termo “violência” estava associado;

Para os deputados:

- Todas as vezes em que o termo “chuva” apareceu, em 50% o termo “festa” estava associado;
- Todas as vezes em que o termo “livro” apareceu, em 33% o termo “cultura” estava associado;
- Todas as vezes em que o termo “carnaval” apareceu, em 17% o termo “cultura” estava associado;

Para os engenheiros:

- Todas as vezes em que o termo “entretenimento” apareceu, em 25% o termo “cultura” estava associado;
- Todas as vezes em que o termo “cultura” apareceu, em 25% o termo “lazer” estava associado;
- Todas as vezes em que o termo “calor” apareceu, em 17% o termo “chuva” estava

associado;

Para os escritores:

- Todas as vezes em que o termo “cultura” apareceu, em 100% o termo “livro” estava associado;

Para os estudantes:

- Todas as vezes em que o termo “cultura” apareceu, em 5% o termo “livro” estava associado;
- Todas as vezes em que o termo “162 anos²⁵” apareceu, em 32% o termo “festa” estava associado;
- Todas as vezes em que o termo “calor” apareceu, em 4% o termo “chuva” estava associado;
- Todas as vezes em que o termo “river” apareceu, em 15% o termo “show” estava associado;

Para os professores:

- Todas as vezes em que o termo “cajuína” apareceu, em 25% o termo “cultura” estava associado;
- Todas as vezes em que o termo “carnaval” apareceu, em 50% o termo “festa” estava associado;
- Todas as vezes em que o termo “praia” apareceu, em 14% o termo “festa” estava associado;

De posse destes novos resultados, foi possível perceber um viés identitário nas associações, uma representação de certos posicionamentos fecundos a determinadas profissões, por exemplo, a relação existente nas postagens dos escritores “*Todas as vezes em que o termo “cultura” apareceu, em 100% o termo “livro” estava associado*”. Tal relação

²⁵ Alusão ao aniversário de Teresina, capital do Piauí.

denota um sentimento muito pessoal, que foi possível perceber mesmo em grande volume de dados. Neste sentido, o data mining é útil para analisar mananciais de informações e a partir deles extrair conhecimento, sobretudo quando se propõe realizar pesquisas em comunicação, um campo de certo modo recente, e no qual os métodos de pesquisa atuais já não conseguem abarcar questionamentos dispostos em um nível de informação muito acentuado.

Ademais, com esses novos resultados, flagrou-se que o emprego das técnicas de associações trazem informações relevantes, desde que o dado esteja com uma qualidade aceitável, ou seja, dentro do processo de data mining é imprescindível a estruturação correta dos dados. Em nossa primeira tentativa de figurar o emprego do DM, não obtivemos o êxito esperado em virtude de uma baixa correlação existente entre os dados analisados (*hashtag* com *hashtag*). Quando num segundo momento, procuramos estruturar o experimento com base nos termos, desprezando as *hashtags*, a quantidade de associações foi significativamente maior.

Outrossim, é importante mencionar que dentre todas as possibilidades de análises desta pesquisa, o DM atendeu a contento todas as prerrogativas suscitadas, obviamente que para cada experimento deve ser pontuado qual a melhor técnica e assim retirar maior proveito da ferramenta em questão.

9. CONSIDERAÇÕES FINAIS

Em meio a este novo cenário, em que as pessoas passaram de meras receptoras e assumiram também a condição de produtoras de informação, temos uma sociedade centrada na era do “big data”. Nesse contexto, a quantidade de *bytes* produzida na web, e de maneira mais pontual nas redes sociais, necessita de filtros cada vez mais dinâmicos para transformar esse manancial de dados em informação útil e relevante.

Diante deste cenário, o excesso de informação passou a ser pauta de análises nas mais variadas esferas do conhecimento e associado a isto um tempo cada vez mais diminuto para inferências e construções de pressupostos. Nessa perspectiva, os pesquisadores precisam realizar uma imersão em grandes volumes de dados num tempo relativamente curto faz-se necessário a utilização de técnicas que oportunizem tal análise. Contudo, muitas vezes, para superar este imbróglio, há a necessidade de quebrar um hermetismo epistemológico e buscar mais artefatos transmetodológicos. Ou seja, obter em outras áreas do conhecimento métodos que possam se relacionar e assim inferir novas descobertas.

No campo da comunicação, essa tônica também não poderia ser diferente, sobretudo pelo fato de entendermos que a própria concepção do campo deu-se pela junção de outros saberes, a exemplo do *marketing*, da fotografia, da relações públicas, entre outros. Neste sentido, trouxemos à baila uma proposta metodológica, pela qual procuramos responder o seguinte questionamento: como o data mining pode ser útil para analisar a identidade do Piauí no twitter?

Por conjecturarmos que os métodos atuais em pesquisa em comunicação já não conseguem deslindar os questionamentos aventados por grande volumes de dados, trouxemos um método nativo da ciências da computação e colocamos à prova sua eficiência na tentativa de pontuar seus contornos e como sua adesão poderia ser utilizada dentro do seio da comunicação.

Diante da problemática suscitada neste trabalho, onde o ponto fulcral era trabalhar

com um volume de dados considerável, contabilizando 671.366 tweets, e assim estabelecer um entendimento mais detalhado acerca da identidade dos piauienses no twitter, o data mining se mostrou muito eficiente em vários momentos, haja vista a rapidez na entrega de informações.

Entre as análises que foram propiciadas pelo emprego do DM, percebeu-se que em 48% da nossa amostra não foi possível pontuar características que pudessem identificar quem eram os usuários. Combinado a esse percentual, flagrou-se que esta característica de “anonimato” não coaduna com os preceitos impostos pelo ciberespaço e que a auto-revelação favorece o aumento de laços dentro da rede social twitter. Para tanto, basta analisar que os indivíduos que descrevem a sua biografia apresentam, em média, uma incidência de 94% a mais no número de seguidores em comparação àqueles usuários que optaram por deixar o campo bio vazio.

Outro ganho que se percebeu em trabalhar com mineração de dados foi a facilidade de realizar classificações de grande volume de dados. Em poucos segundos foi possível estabelecer uma espécie de censo sobre as principais ocupações descritas no micro blog. A partir da classificação dos agentes da pesquisa, combinado ao número de seguidores na referida rede social, constatamos que os profissionais liberais possuem grande audiência. Nesse caso, temos que para inferir descobertas onde necessitam-se classificar grandes quantidades de elementos, as técnicas de DM, em especial os algoritmos de classificação, atendem a contento essa necessidade.

No que diz respeito aos aspectos relacionados à identidade dos piauienses no twitter, a mineração de dados ofereceu uma série de descobertas, entre elas destacamos o fato que as *hashtags* utilizadas, durante os doze meses do ano, não representam um valor identitário significativo. Em outras palavras, as temáticas abordadas através de *hashtag* em 2014 não apresentaram características que pudessem imprimir uma “piauiensidade” consolidada, visto que dos 56 termos que apareceram em todos os meses, apenas três *hashtags* trouxeram aspectos que pudessem representar tal identificação: #teresina, #piaui e #ufpi.

Contudo, mesmo quando estes termos apareceram, na maioria da vezes, continham um apelo de *marketing*, tendo em vista que os estabelecimentos comerciais usavam #piaui ou #teresina apenas para identificar as suas respectivas localizações. Nesta acepção, as discussões pontuadas na amostra estudada davam prioridade a programas inseridos na grade televisiva da Rede Globo de Televisão, tais como as telenovelas Amor à Vida e Em Família, além do

reality show Big Brother Brasil.

Outra descoberta oportunizada pela utilização do data mining foi perceber, através de técnicas de associação, como eram construída as correlações entre as *hashtags*. Nesta análise, em especial, não obtivemos os resultados esperados em virtude da baixa correlação existente entre as *hashtags*. Isto é, os usuários dificilmente utilizavam na mesma postagem duas *hashtags* e, quando o faziam, a sua associação tinha um baixo valor estatístico e por conta deste comportamento a mineração de dados não trouxe informações relevantes. Diante deste cenário, percebemos que a utilização de técnicas de associação aplicando as *hashtags*, além de moroso, não foi assertiva.

No entanto, ao utilizar a mesma técnica com algumas expressões contidas dentro dos twittes e excluindo a análise de *hashtag* obtivemos informações mais relevantes. Nesse interim foi possível notar que a eficiência das técnicas de data mining estão intimamente relacionadas à forma como os dados estão dispostos, além disso, utilizar mineração de dados não implica necessariamente na descoberta de novas informações. Isto porque, o relacionamento entre os dados contido na base de dados é primordial para a inferência e o sucesso da pesquisa.

Outro aspecto que deve ser levado em consideração é que o data mining é uma ferramenta e como tal apenas facilita a busca de informações em grandes volumes de dados. Sem o domínio do pesquisador sobre o objeto pesquisado, as informações coletadas terão baixa significância. Neste sentido, acreditamos que o data mining é uma simbiose entre homens e máquina, onde a máquina minera os dados e o pesquisador, com base nos seus conhecimentos, ventila a assertividade dos objetivos propostos na pesquisa.

Por fim, no que circunscreve ao questionamento suscitando na gênese desta pesquisa percebeu-se que face ao grande volume de dados trazido neste trabalho, 671.366 twittes, sem a utilização de técnicas de data mining não seria possível realizá-lo em tempo hábil. Ademais, o emprego da mineração de dados foi muito proveitoso, pois apresentou vários arquétipos de análise (classificação, clusterização e associação) para o mesmo conjunto de dados e com cada modelo empregado foi possível estabelecer, sob vários enfoques, como a identidade dos piauienses era pontuada na rede social twitter.

REFERÊNCIAS

- AGÊNCIA BRASIL. Disponível em: < <http://agenciabrasil.etc.com.br/>>. Acesso em: 12 de ago. 2015.
- BAKHTIN, Mikhail. **Estética de Criação Verbal**. Trad. Paulo Bezerra. 4ª ed. São Paulo: Martins Fontes, 2003.
- _____, Mikhail. **Marxismo e Filosofia da Linguagem**. Trad. Michel Lahud e Yara Frateschi. 7. ed. São Paulo: Hucitec, 1995.
- BATISTA, G. E. A. P. A. (2003). **Pré-processamento de dados em aprendizado de máquina supervisionado**. Tese de Doutorado, ICMC-USP.
- BAUMAN, Zygmunt. **Identidade: entrevista a Benedetto Vecchi**. Rio de Janeiro, Jorge Zahar Ed., 2005.
- BERRY, M.J.A.; LINOFF, G. Data Mining Techniques: for Marketing, Sales and Customer Support. New York: John Wiley & Sons, 1997.
- BEZERRA, Francisco O.; SOBRINHO, Lemuel D. G. **Ciberespaço, Cotidiano e Identidades**: Novas Leituras Sobre Interações Mediadas. BARATARIA, Revista Castellano-Manchega de Ciências Sociais. Número 12, 2011.
- BROWN, J. S.; PAUL, D. **A Vida Social da Informação**. v. 1. Makron Books: São Paulo, 2001.
- BRUNO, Fernanda. **Dispositivos de vigilância no ciberespaço: duplos digitais e identidades simuladas**. Revista Fronteira, São Leopoldo/RS, v. VIII, p. 152-159, 2006.
- CASADEI, Eliza Bachega. Os Novos Lugares de Memória na Internet: as práticas representacionais do passado em um ambiente online. **BOCC - Biblioteca On-line de Ciências da Comunicação**, v. 1, p. 1-27, 2009.
- CASTELLS, Manuel. **A Galáxia da Internet: Reflexões sobre a Internet, os negócios e a sociedade**. Rio de Janeiro: Jorge Zahar Editor, 2003.
- _____, Manuel. **A Era da Informação: economia, sociedade e cultura**, vol. 1. Lisboa: Fundação Calouste Gulbenkian, 2002.
- _____, Manuel. **A sociedade em rede**. São Paulo: Paz e Terra, 1999.
- EBECKEN, N. F. F; LOPES, M. C. S.; COSTA, M. C. A. Mineração de Textos. In: Solange Oliveira Rezende. (Org.). **Sistemas Inteligentes - Fundamentos e Aplicações**. 1ed.Barueri - SP: Editora Manole Ltda., 2003, v. , p. 337-370.
- ENNE, Ana Lucia. Memória, identidade e imprensa em uma perspectiva relacional. **Revista Fronteiras: estudos midiáticos**, v. 6, n. 2, p. 101-116, jul/dez, 2004.

FAYYAD, U. M.; PIATETSKY-SHAPIO, G.; SMYTH, P. **From Data Mining to Knowledge Discovery: An Overview**. Knowledge Discovery and Data Mining, Menlo Park : AAAI Press, 1996.

FRANÇA, V. **Paradigmas da Comunicação: conhecer o quê?** Ciberlegenda, Niterói (RJ), n.05, 2001. Disponível em: <http://www.uff.br/ciberlegenda/ojs/index.php/revista/article/view/314/195>. Acesso em: 17 de maio de 2013.

FREIRE, Flávia. **APIs, Uma Questão Estratégica e de Inteligência**. Revista TI Digital, nº 5, p. 28-40, 2013.

GABRIEL, Marta. **Marketing de Otimização de Buscas na Web**. São Paulo: Esfera, 2008.

GRINGS. Cristiane **A pesquisa da pesquisa e a descoberta do transdisciplinar e do transmetodológica**. In Perspectivas Metodológicas em Comunicação: desafios na prática investigativa. João Pessoa:EUFP: 2008

HALL, Stuart. **A identidade cultural na pós-modernidade**. Tradução de Tomás Tadeu da Silva, Guaracira Lopes Louro. 11 ed. Rio de Janeiro: DP&A, 2006.

HAN, J.; KAMBER, M. **Data mining: concepts and techniques**. New York: Morgan Kaufmann, 2000.

IBM. Disponível em: < <http://www.ibm.com/br/pt/>>. Acesso em: 6 de mar. 2015.

HORNIK, David. **Social Networks 3.0 Ventureblog, 2005**. Disponível em www.ventureblog.com/article/2005/12/social_networks.php. Acesso em 25/04/2014.

LE MOS, André. **Cibercultura e Mobilidade: a Era da Conexão**. Razón Palabra, México, out.-nov.2004

_____, André. **Cibercultura, Tecnologia e Vida Social na Cultura**. Contemporânea. Porto Alegre, Sulina. 2010

Lévy, Pierre **Cibercultura**. Tradução de Carlos Irineu da Costa. 3 ed. São Paulo: Ed. 34, 2010.

_____, Pierre. **A Inteligência Coletiva**. São Paulo: Loyola, 2007.

_____, Pierre. **O que é o virtual?** São Paulo: Editora 34, 1996.

LIMA JR. Walter Teixeira. **Jornalismo Inteligente na era do data mining**. Publicado na Revista do Programa de Pós-graduação da Faculdade Cásper Líbero, ano IX – no.18, p. 121-126, 2011.

LIMA JUNIOR, W. T. . **Big Data, Jornalismo Computacional e Data Journalism: estrutura, pensamento e prática profissional na Web de Dados**. Estudos em Comunicação, v. 12, p. 207-222, 2012.

MACHADO, Elias. **O ensino de jornalismo em tempos de ciberespaço**. In MACHADO, Elias e PALACIOS, Marcos. O Ensino de Jornalismo em Redes de Alta Velocidade. Metodologias e Softwares. Salvador: EDUFBA: 2007.

MAIGRET, E. **Sociologia da Comunicação e das Mídias**. São Paulo: Senac, 2010.

MANDEL, A., SIMON, I., & DELYRA, J. **Informação: computação e comunicação**. Revista da

USP, 35, 11-45, 1997.

MARCONDES FILHO, Ciro. **Pensar - pulsar: cultura comunicacional, tecnologias, velocidades.** São Paulo: Ed. NTC, 1995.

MARTINO, L.C. **História e Identidade: apontamentos epistemológicos sobre a fundação e fundamentação do campo comunicacional.** 2004, São Bernardo do Campo/SP: XIII COMPÓS, GT – Epistemologia da Comunicação. Disponível em: <http://www.compos.org.br/>. Acesso em: 05 de maio de 2013.

MONTILLA, Alfredo. Twitter y participación ciudadana en Venezuela. In: XXXIV CONGRESSO BRASILEIRO DE CIÊNCIAS DA COMUNICAÇÃO – GP COMUNICAÇÃO E DESENVOLVIMENTO REGIONAL E LOCAL. Recife, 2011. Recife. **Anais eletrônicos.** Universidade Católica de Pernambuco, 2011. Disponível em: <<http://www.intercom.org.br/papers/nacionais/2011/resumos/R6-2211-1.pdf>>. Acesso em out. 2014.

NORA, Pierre. **Entre memória e história: a problemática dos lugares.** Projeto História, São Paulo, n.10, dez. 1993, p.7-28.

PENTLAND, Alex. **A maior revolução em 300 anos.** [11 de março, 2015]. São Paulo: Revista VEJA. Entrevista concedida a Pieter Zalis.

RECUERO, Raquel. **Redes Sociais na Internet.** Porto Alegre: Sulina, 2009.

_____, Raquel. **RT, por favor: considerações sobre a difusão de informações no Twitter.** Revista Fronteiras, v. 12, p. 1-16, 2010.

RHEINGOLD, Howard. **The Heart of the WELL.** In HOLETON, Richard. Composing Cyberspace: Identity, Community and Knowledge in the Electronic Age. McGraw-Hill. USA, 1998.

SAID, G.; MAGALHÃES, T. M. S. **Redes Sociais, cidade e memória: lembranças compartilhadas de Teresina no Twitter.** In: Carlos Gerbase; Eduardo Campos Pellanda; Juliana Tonin. (Org.). Meios e Mensagens na Aldeia Virtual. 1ª ed. Porto Alegre: Editora Sulina, 2012, v. 1.

SANTAELLA, Lucia; LEMOS, Renata. **A cognição conectiva do Twitter.** São Paulo: Paullus, 2010.

SCHIESSL, José Marcelo. **Descoberta de conhecimento em texto aplicada a um sistema de atendimento ao consumidor.** 2007 Dissertação (mestrado em Ciência da Informação) – Universidade de Brasília, Brasília, 2007.

SIGNATES, L. **Da exogenia aos dispositivos: roteiro para uma teorização autônoma da comunicação.** 2012, Juiz de Fora/BH: XXI COMPÓS, GT – Epistemologia da Comunicação. Disponível em: <http://www.compos.org.br/>. Acesso em: 17 de maio de 2013.

SOARES, M. V. B. **Pretext ii: Descrição da reestruturação da ferramenta de pré-processamento de textos.** Relatório Técnico 333, Instituto de Ciências Matemáticas e de Computação, USP, São Carlos, 2008.

SODRÉ, Muniz. **Reinventado cultura.** 3ª ed. Pólis, RJ: Vozes, 1999.

SPINK, A., WOLFRAM, D., JANSEN, M. B., & SARACEVIC, T. **Searching the web: the public**

and their queries. Journal of the American Society for Information Science and Technology, Vol. 52, 226–234, 2001.

TURKLE, Sherry. **Life on the Screen: Identity in the Age of the Internet**. New York: Simon & Schuster, 1997.

WEINBERGER, D., **Why Open Spectrum Matters. The end of the broadcast nation**. In <http://www.evident.com>, 2003.

WEISS, S. M., INDURKHYA, N., ZHANG, T., e DAMERAU, F. (2005). **Text Mining: Predictive Methods for Analyzing Unstructured Information**. Springer Science+Business Media, Inc.

WELLMAN, B.; BOASE, J.; CHEN, W. **The Global Villagers: Comparing Internet Users and Uses Around the World**. In: WELLMAN, b.; HAYTHORNTHWAITE, C. **The Internet in Everyday Life**. (p.74-113). Oxford: Blackwell, 2002. _____. b **The Networked Nature of Community Online and Offline**. *IT & Society* n.1, vol 1, p.151- 165. Summer, 2002

WERTHEIN, Jorge- **Novas Tecnologias e a Comunicação Democratizando a Informação**, UNESCO, 2004 Disponível em: <http://www.qprocura.com.br/dp/60454/Novas-Tecnologias-e-a-Comunicacao-Democratizando-a-Informacao.html> Acesso em 08 de jun. 2004.

WING, J. M. **Computacional thinking. Communications of the ACM**, v. 39, n. 3, 2006. Disponível em: <http://www.cs.cmu.edu/afs/cs/usr/wing/www/publications/Wing06.pdf>. Acesso em: 19 de jun. 2014.

ZIKOPOULOS, P; DE ROOS, D; PARASURAMAN, K; DEUTSCH, T; GILES, J; CORRIGAN, D. **Harness the power of Big Data- The IBM Big Data Platform**. Emeryville: McGraw-Hill Osborne Media, 2012.