UNIVERSIDADE FEDERAL DA PARAÍBA

Centro de Informática

PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

FRANCISCO PORFÍRIO RIBEIRO NETO

MELHORIA NA CLASSIFICAÇÃO DE TÓPICOS EM TEXTOS CURTOS USANDO *BACKGROUND KNOWLEDGE*

João Pessoa – Paraíba
Setembro / 2015

Francisco Porfírio Ribeiro Neto

MELHORIA NA CLASSIFICAÇÃO DE TÓPICOS EM TEXTOS CURTOS USANDO *BACKGROUND KNOWLEDGE*

Dissertação de Mestrado apresentada ao programa de Pós-Graduação em Informática da Universidade Federal da Paraíba para obtenção do título de Mestre.

Orientador: Andrei de Araújo Formiga

João Pessoa – Paraíba
Setembro/ 2015

Ata da Sessão Pública de Defesa de Dissertação de Mestrado de **FRANCISCO PORFIRIO RIBEIRO NETO**, candidato ao título de Mestre em Informática na Área de Sistemas de Computação, realizada em 31 de agosto de 2015.

Aos trinta e um dias do mês de agosto do ano de dois mil e quinze, às quatorze horas, no Centro de Informática - Universidade Federal da Paraíba (unidade Mangabeira), reuniram-se os membros da Banca Examinadora constituída para julgar o Trabalho Final do **Sr. Francisco Porfirio Ribeiro Neto** vinculado a esta Universidade sob a matrícula 2013102709, candidato ao grau de Mestre em Informática, na área de *"Sistemas de Computação"*, na linha de pesquisa *"Computação Distribuída"*, do Programa de Pós-Graduação em Informática, da Universidade Federal da Paraíba. A comissão examinadora foi composta pelos professores doutores: **Andrei de Araújo Formiga (PPGI-UFPB),** Orientador e Presidente da Banca, **Clauirton de Albuquerque Siebra (PPGI-UFPB),** Examinador Interno, **Thaís Gaudêncio do Rego (PPGI-UFPB)** Examinadora Interna e **Vinicius Ponte Machado** (UFPI), Examinador Externo à Instituição. Dando início aos trabalhos, o professor Presidente da Banca cumprimentou os presentes, comunicou aos mesmos a finalidade da reunião e passou a palavra ao candidato para que o mesmo fizesse, oralmente, a exposição do trabalho de dissertação intitulado *"Melhoria na Classificação de Tópicos em Textos Curtos Usando Background Knowledge"*. Concluída a exposição, o candidato foi arguido pela Banca Examinadora que emitiu o seguinte parecer: *"aprovado"*. Assim sendo, eu, Nadja Rayssa Soares de Almeida, Auxiliar em Administração, Secretária do Programa de Pós Graduação em Informática - PPGI, lavrei a presente ata que vai assinada por mim e pelos membros da Banca Examinadora. João Pessoa, 31 de agosto de 2015.

Nadja Rayssa Soares de Almeida

Profº Drº Andrei de Araújo Formiga
Orientador (PPGI-UFPB)

Profº Drº Clauirton de Albuquerque Siebra
Examinador Interno (PPGI-UFPB)

Profª Drª Thaís Gaudêncio do Rego
Examinadora Interna (PPGI-UFPB)

Profº Drº Vinicius Ponte Machado
Examinador Externo à Instituição (UFPI)

DEDICATÓRIA

A Deus por me proporcionar saúde, calma, paciência e discernimento, dimensões essenciais para realização deste projeto.

Ao meu pai Zizinho, minha mãe Ires e a minha esposa Ana Rita.

AGRADECIMENTOS

Muitos são os nomes que merecem aparecer nesta página, afinal, foram aproximadamente cinco anos de sonhos, angustias, dissabores, abdicações e felicidades. Tentarei segmentar o agradecimento por grupos de pessoas, mencionando as que mais tiveram relevância na construção desta dissertação.

Primeiramente devo agradecer a minha família, em especial aos meus pais, minha esposa e meu filho, que ainda está no ventre da mãe. Sendo os meus pais, responsáveis em permitir que eu iniciasse este sonho, e dando o empurrão inicial. Já a minha esposa, esteve presente em todas as etapas, desde o desejo da realização do mestrado, até o seu encerramento. E não poderia deixar de falar do meu filho Theo, que ao saber da sua existência, passei a ter um ânimo diferente, o que me estimulou na reta final deste projeto.

Agradeço também a família "Fortalecidos no Amor" (Grupo ECC), por toda corrente de oração destinada a conclusão do mestrado, podem ter certeza que vocês também fizeram toda a diferença.

Não poderia deixar de agradecer aos mestres que me ajudaram no ingresso do mestrado, e me deram palavras de incentivos ao longo de sua realização, por isso, fica aqui o meu muito obrigado ao Professor Doutor Luiz Maurício, e a Professora Mestre Vanessa Dantas.

Agradeço ainda ao Tribunal de Justiça da Paraíba na pessoa do Diretor de Tecnologia Ney Robson, por permitir a realização deste projeto, sem nenhuma objeção nos momentos em que precisei me ausentar para cumprir com as obrigações do mestrado.

E por último, mas que considero um dos mais importantes agradeço ao meu orientador, professor Doutor Andrei de Araújo Formiga. Pessoa extremamente humilde e acessível, e com uma capacidade tremenda de conduzir o aluno ao foco do trabalho, sem mesmo que o aluno perceba. Apesar de seu jeito tímido de ser, Andrei nunca deixou com que eu perdesse as esperanças, e com simples palavras de "Vamos Defender", me passava uma segurança que eu saia com um ânimo novo a cada reunião.

Para não pecar pela omissão, deixo registrado o meu muito obrigado a todos aqueles que contribuíram para conclusão deste projeto.

RESUMO

O poder da interação entre usuários na internet aumentou consideravelmente através do surgimento de ferramentas alinhadas com os conceitos da WEB 2.0, a exemplo dos blogs, fóruns de discussão, e redes sociais como o Facebook e Twitter. Estas aplicações são constituídas por uma troca contínua de mensagens entre os usuários, gerando com isso, uma expressiva massa de dados formada por várias pequenas mensagens. Estudos mostram que informações podem ser extraídas com base em tais dados através da classificação de textos. O desafio da classificação de texto nesse contexto é que as mensagens dos usuários nas redes sociais são curtas, por exemplo o *twitter*, que possui apenas cento e quarenta caracteres, fazendo com que ocorra o problema de escassez de dados e ausência de similaridade entre palavras relevantes. Nesta dissertação é apresentada uma técnica que propõe uma melhoria na classificação de tópicos em textos curtos usando *background knowledge*. A técnica proposta consiste em realizar a classificação de textos curtos em três etapas, usando um algoritmo de classificação de texto convencional, sendo o NaiveBayes escolhido para esta dissertação, realizando uma comparação simples de texto, aqui denominada de "Contador de Palavras" que verifica a existência ou não das palavras-chave da *background knowledge* nos textos e, finalmente, ao término das duas etapas, os resultados são combinados para que o texto seja efetivamente classificado. Para demonstrar a eficiência da melhoria proposta, foram extraídos mensagens do Twitter e construído um *Corpus* em português tendo como tema a "Violência". Esse *Corpus* foi utilizado em um experimento para determinar o desempenho de classificação da técnica proposta. Os resultados demonstram que a técnica de classificação proposta consegue melhorar o desempenho de classificação de tópicos em textos curtos.

ABSTRACT

The power of interaction between internet users has grown since the appearance of tools aligned with the principles of WEB 2.0, including blogs, forums and social networks like Twitter and Facebook. This kind of application is based on frequent message exchanges between users, generating large quantities of textual data comprised of small messages. Text classification techniques allow the extraction of relevant information from such messages. In this context, the challenges are related to the fact that the short messages common in social networks contain, individually, too little data for the traditional analyses. In this work a new technique for improving topic classification in short texts is proposed. This technique is based on the idea of combining a standard text classifier with a keywords-based simple classification scheme; the novelty here beyond the combination of two classification schemes is the use of a semi-automated, unsupervised technique for building the list of keywords reated to the desired topic; this technique is based on the use of topic modeling using the LDA algorithm. To demonstrate the validity of the proposed approach, a *Corpus* of twitter messages was built around the topic "violence". This *Corpus* was used in experiments to assess the performance of the proposed classification technique. Results show that topic classification for short texts is improved by the proposed technique.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

API – *Application Programming Interface*

RT – *Retweet*

IP – *Internet Protocol*

HTTP – *Hypertext Transfer Protocol*

IA - Inteligência artificial

AM - Aprendizado de Máquina

LDA – *Latent Dirichlet Allocation*

URL – *Uniform Resource Locator*

SUMÁRIO

1. Introdução

1.1 Contextualização e Motivação

Construir informações a partir de dados originados por diferentes pessoas e em diferentes momentos é um desafio constante da tecnologia da informação. Várias estratégias como a de *Business Intelligence* e Mineração de Dados continuam sendo amplamente estudadas, testadas e implantadas em organizações, com o objetivo único de buscar informações com base nas ações das pessoas.

Com surgimento da WEB 2.0, marcada pelo poder de interação entre internautas, tornou-se possível que novos dados pudessem contribuir ainda mais na construção de conteúdo (BOHN, 2010). As redes sociais são exemplos de ferramentas que fazem parte da WEB 2.0, que de forma resumida, pode ser definida como sendo uma rede de relacionamento estabelecidas pelos atores sociais no ambiente em que estão inseridos e que por meio delas, é possível fortalecer e consolidar as opiniões, permitindo que usuários tomem uma possível conclusão de fatos relevantes através de trocas de mensagens (NASCIMENTO; BEUREN, 2011)

Diante deste cenário, as redes sociais vêm sendo amplamente explorada por pesquisadores, como fonte de levantamento de informações, pelo fato de ferramentas como o *Twitter* permitirem que usuários passem *feedbacks* instantâneos dos fatos o que significa que qualquer ação no mundo real geralmente recebe uma reação instantânea na forma de *tweets* (RECUERO, 2009). No entanto, extrair informações de ferramentas como *Twitter* exige adoção de técnicas diferenciadas, afinal, trata-se de um grande volume de mensagens curtas de até 140 (cento e quarenta) caracteres.

Em virtude do curto comprimento das referidas mensagens, as mesmas não proporcionam co-ocorrência suficiente de palavras PHAN (2008), incorrendo no problema de esparsidade dos dados, portanto, os métodos normais de classificação de texto não conseguem atingir uma precisão aceitável devido à escassez de dados.

Após levantamento bibliográfico, percebeu-se que na literatura existem diversas comparações entre estratégias de classificação de textos longos e curtos, aplicadas a mensagens enviadas pelo Twitter. E foi comprovado em diversos casos, através de experimentos aplicados a textos curtos, que a classificação de mensagens do Twitter usando uma estratégia de classificação de textos curtos, possui melhores resultados se comparado a métodos tradicionais.

Ao ser analisado as estratégias de classificação de textos curtos apresentados nos trabalhos correlatos, foram identificadas que a grande maioria das técnicas apresentadas fazem uso de uma base secundária com a finalidade de complementar para fins de enriquecimento de dados denominada *Background Knowledge*, utilizada para sanar o problema de escassez dos dados. Diante deste cenário, questiona-se: existe um meio de usar a referida base como um elemento capaz de melhorar a classificação de tópico em textos curtos?

A problemática levantada no parágrafo anterior motivou a realização desta dissertação. E para responder ao questionamento,foi proposta uma melhoria na classificação de tópicos em textos curtos usando *background knowledge*.

Para construção da referida base, foi utilizado uma estratégia semi-automática, onde inicialmente foi realizada a modelagem de tópicos e, em um segundo momento, foram extraídos dos tópicos os termos mais relevantes para compor a *background knowledge*.

A ideia central da melhoria proposta, consiste em realizar a classificação de textos curtos em duas etapas, na primeira é usado um algoritmo de classificação de texto convencional, sendo o NaiveBayes escolhido para esta dissertação pelo fato deste já obter resultados comprovadamente satisfatórios na classificação de textos convencionais, e na segunda etapa, é realizado uma comparação simples de texto, aqui denominada de "Contador de Palavras", verificando a existência ou não das palavras-chave da *background knowledge* nos *tweets*. Ao término das duas etapas os resultados são combinados, para que assim o seja estabelecido um rótulo ao *tweet*.

Para validação da melhoria proposta ao processo de classificação, foi aplicado um experimento tendo como base de dados o *Corpus* "Violência". Esta escolha foi feita pelo fato da violência se caracterizar como um grave problema social, econômico e de saúde publica do Brasil. O contínuo e rápido incremento da violência cotidiana configura-se como aspecto representativo e problemático da atual organização da vida social (FREITAS; SILVA; MORAES, 2011). Diversas pesquisas voltadas para a violência no Brasil vêm sendo realizadas por WAISELFISZ (2008), dentre estas, pode-se destacar "O Mapa da Violência dos Municípios Brasileiros", que possui diversos estudos que possibilitam ponderar a situação e a evolução deste agravo em diversos locais do país, baseado em dados fornecidos pelo Subsistema de Informação Sobre Mortalidade (SIM) do Ministério da Saúde. A referida pesquisa tornou os Mapas da

Violência insumos essenciais para avaliação, e elaboração dos planos e estratégias de enfrentamento da violência no país.

Como conseqüência desta violência cotidiana instalada com tanta velocidade, se faz necessário que análises mais rápidas sejam realizadas em busca de mapear a violência no Brasil. RECUERO (2009), menciona que as reações dos usuários no *Twitter* são quase que instantâneas, identificou-se neste meio uma oportunidade de auxiliar a otimização da análise de violência no Brasil, através da melhoria de classificação de textos curtos aqui proposto (PRIER et. al, 2011)

Ao término deste trabalho serão apresentados os resultados obtidos do classificador convencional, na ocasião NaiveBayes, do contador de palavras e do classificador proposto, sendo utilizada a métrica F1 para ilustrar a eficiência da melhoria proposta aplicada ao *Corpus* "Violência".

1.2 Objetivos do Trabalho

O objetivo desta dissertação consiste em estabelecer uma melhoria na classificação de tópicos em textos curtos usando *background knowledge*. Para isso, foi construído um *Corpus* em português relacionado com "violência", que por sua vez foi utilizado nos experimentos.

1.3 Organização do Trabalho

Este trabalho está estruturado da seguinte forma: no Capítulo 2 foi feita uma breve introdução teórica dos conceitos básicos sobre redes sociais e serão abordados ainda conceitos acerca do *Twitter*; também foram apresentados conceitos de Inteligência Artificial e Aprendizado de Máquina, para finalmente adentrar no foco desta dissertação que é classificação de textos. O Capítulo 3 ilustra alguns trabalhos já realizados nesta área de classificação de textos curtos. No Capítulo 4 é apresentado o classificador proposto, bem como o detalhamento da etapa de construção do *Corpus*. Já no Capítulo 5 foi abordado o experimento realizado e os resultados obtidos. Por fim, serão apresentadas, no Capítulo 6 a conclusão e sugestões de trabalhos futuros.

2. Fundamentação Teórica de Classificação de Textos curtos

Com o aumento da interação entre usuários nas redes sociais, passou a existir a possibilidade de expressão e interação através das ferramentas de comunicação mediada pelo computador, permitindo assim, que estes indivíduos interajam nas redes sociais deixando evidências suficientes para construção de informação com base nas mensagens enviadas. Essas mensagens tornam possível a aplicação de diferentes estratégias de classificação de texto, estratégias estas que fazem parte de uma área da Inteligência Artificial conhecida como Aprendizado de Máquina.

Neste capítulo, serão descritas, de forma conceitual, as ferramentas e técnicas utilizadas para viabilizar o método de classificação proposto nesta dissertação.

2.1 Redes Sociais

De acordo com RECUERO (2009), uma rede social é definida como um conjunto de dois elementos: *atores* (pessoas, instituições ou grupos; os nós da rede) e suas *conexões* (interações ou laços sociais). Em pesquisa realizada pelo Ibope/NetRatings, foi detectado que desde o surgimento dos blogs e redes sociais (*Orkut, MySpace, Twitter, Facebook*, etc.), o número de atores brasileiros conectados aumentou. Praticamente metade dos brasileiros, 48%, faz uso da internet. O percentual de pessoas que a utilizam todos os dias creceu de 26%na PBM 2014 para 37% BRASIL (2015).

Já ROMANO (2013) Identificou-se que o Brasil é o segundo colocado em usuários do *Twitter* e do Facebook, atrás apenas dos Estados Unidos. Após pesquisa, foi encontrada uma entrevista com o Guilherme Ribenboim, diretor do *Twitter* no Brasil, onde foi perguntado “O que destaca o *Twitter* como rede social?”

> O *Twitter* é mais uma rede de interesse do que uma rede social. Não é sobre estar se relacionando é sobre estar próximo do conteúdo. O *Twitter*, muitas vezes, vem substituindo os jornais como primeira fonte de informação. É uma plataforma de consumo de conteúdo, mas que fica mais rica com a atmosfera social (ROMANO,2013).

Diante da resposta do Guilherme Ribenboim à entrevista, percebeu-se que os usuários de Brasil são seduzidos pelo *Twitter* especificamente, não apenas pelas relações, mas por terem acesso a acontecimentos de todo o mundo e de pessoas específicas, sem a necessidade de assistir a jornais, ou mesmo de encontrar com outras pessoas para serem atualizadas de um fato. Desta forma, conclui-se que o *Twitter* se enquadra muito bem neste estudo, uma vez que a forma como as pessoas usam tal ferramenta, segundo Guilherme Ribenboim, é favorável ao cumprimento do objetivo proposto: classificar mensagens curtas que giram em torno de um determinado fato.

2.1.1 Twitter

Segundo Santos (2013), o *Twitter* é uma ferramenta de micro mensagens, lançada em 2006, obtendo um rápido crescimento no Brasil e no Mundo por três motivos principais:

- Por possuir uma API robusta e versátil que permitiu aos desenvolvedores prover serviços entre diversos meios (celulares, navegadores, aplicativos isolados, etc.);
- Adoção por parte da mídia e celebridades como fonte de divulgação;
- Aprimoramento da ferramenta pelos usuários, como por exemplo adoção do uso de *hashtags* e r*e-tweet*.

Originalmente, os usuários do *Twitter* são convidados a responder à pergunta "O que você está fazendo?" em até 140 caracteres. Através do *Twitter* qualquer pessoa pode "seguir" e ser "seguido" por algum outro usuário da mesma ferramenta, estabelecendo com isso uma relação, sem a obrigação da recíproca ser verdadeira. Ou seja, se um usuário seguir outro, não necessariamente este deve ser seguido pelo mesmo usuário. Cada usuário possui autonomia para escrever e partilhar suas próprias mensagens com um grupo de seguidores, mensagens estas mais conhecidas como "*tweets"* (RECUERO; ZAGO, 2009). Os *tweets* podem variar bastante na temática, abordando desde tópicos mais rotineiros e pessoais, até notícias ou reações a eventos de importância social e internacional.

O *Twitter* pode ser definido também como um site de rede social que dispõe de um espaço na web capaz de permitir aos seus usuários construir perfis públicos, articular suas redes de contatos, tornarem visíveis essas conexões e interagir com outras

pessoas através de mensagens (BOYD; ELLISON, 2007 apud RECUERO; ZAGO, 2009).

Este instrumento já foi objeto de vários estudos, obviamente com finalidades distintas. Grande parte destes estudos são focados na apropriação dos dados de usuários, e na conversação entre eles, buscando com isso, extrair informações que sejam úteis a sociedade como um todo (RECUERO; ZAGO, 2009).

Em um dos primeiros trabalhos que exploram dados do *Twitter* JAVA et al. 2007 apud SANTOS (2013) analisam o uso do *Twitter*, estudando as propriedades topológicas e geográficas da rede. Através deste estudo, foi possível identificar três categorias de usuários do *Twitter*:

- Os que compartilham informações que são pessoas ou serviços automatizados que postam notícias.
- Os que mantêm relações de amizade, estes são a maioria dos usuários e que envolvem familiares, colegas de trabalho e algumas vezes até estranhos.
- Os que buscam informações e que raramente postam mensagens, porém, seguem outros perfis regularmente.

Ainda com relação ao mesmo estudo, os autores identificaram várias intenções para justificar o uso do *Twitter*, entre elas: conversas sobre o dia-a-dia, onde usuários discutem eventos de sua vida e seus pensamentos; divulgação de notícias; compartilhamento de mensagens de outros usuários, etc.

Outro estudo realizado avalia a topologia da rede e propagação de informação, ilustrando a relação entre os usuários, suas influências e os tópicos de tendência, revelando que há uma desigualdade na distribuição de seguidores entre os usuários da rede e uma baixa taxa de reciprocidade, mostrando que o *Twitter* mais se assemelha com uma rede de compartilhamento de informação do que uma rede social, Confirmando assim declaração feita por Guilherme Ribenboim (KAWAK et.al, 2010).

Sabendo que o *Corpus* "Violência" construído nesta dissertação, foi levantado com base em *tweets*, se faz necessário ter o conhecimento de quais são os dados que estão disponíveis para extração na referida rede social, bem como compreender os métodos de extração de dados do *Twitter*.

2.1.2 Dados provido pelo *Twitter*

Extrair dados do *Twitter* não é uma atividade trivial. Afinal, há uma dificuldade em torno da definição da metodologia que será aplicada para extração dos dados. Muitos pesquisadores buscaram em trabalhos relacionados maiores detalhes no método de extração adotado. Porém, tal detalhamento muitas vezes é precário, uma vez que os autores não mencionam características dos dados extraídos, por exemplo, idioma, geolocalização, período, etc

Manuel (2013) afirma que nos trabalhos relacionados sobre tal temática, em geral, o detalhe com que se apresenta o método de extração é demasiadamente vago, o que dificulta a compreensão de como o pesquisador extrai informações do *Twitter*. Muitos indicam apenas que os dados foram extraídos da *Timeline* dos usuários através da API do *Twitter*, não especificando a estratégia adotada para tal extração, dificultando a compreensão dos experimentos realizados, e até mesmo a possibilidade de sugestão de melhorias.

Diante da dificuldade relatada, foram abordados nesta seção os principais métodos de extração de dados do *Twitter*. Porém, para uma melhor compreensão dos métodos a serem elencados, antes serão apresentados os dados providos pelos usuários, e armazenados pelo *Twitter*. Manuel (2013) afirma que é possível identificar duas grandes fontes de dados na referida rede social, são elas:

- Perfis dos usuários
- *Tweets* emitidos por usuários

Um perfil de usuário do *Twitter* está associado a uma pessoa, ou entidade que criou uma conta em tal ferramenta existindo poucas informações associadas ao perfil pessoal (MANUEL, 2013). Na Tabela 1 estão as características vinculadas ao perfil do usuário.

Tabela 1 – Tabela Sumária dos Campos de um perfil de usuário

Campo	Descrição
Nome da Conta	Nome identificador da conta do usuário, único, mas passivo de modificação pelo usuário
Nome	Nome do usuário
Descrição	Descrição da conta do usuário em até 140 caracteres
Contagem de Favoritos	Número de *tweets* que o usuário definiu como favorito
Língua	Língua do perfil do utilizador
Localização	Campo Destinado a localização, a preencher pelo usuário
Privacidade	Informação sobre a privacidade dos dados da conta e *tweets*
Contagem de *Tweet*s	Número de *tweets* publicados
Fuso Horário	Fuso Horário definido pelo utilizador
URL	Campo reservado para uma hiperligação

Fonte: (MANUEL, 2013)

Tão importante quanto conhecer os dados disponíveis dos usuários, é compreender quais são os dados que estão disponíveis em um *tweet*. Sabe-se que a característica central do *Twitter* é a visualização dos *tweets* postados pelas pessoas que um determinado perfil segue, e que são mostrados ao seguidor em ordem cronológica inversa (BOYD et. al, 2010). Os *tweets* fazem parte da fonte mais dinâmica e diversa, no tipo e qualidade de informação que se encontram disponíveis para extração (MANUEL, 2013).

Conforme já mencionado anteriormente, os usuários possuem um limite de 140 caracteres para responder a pergunta "O que você está fazendo?". Com o uso massivo do *Twitter*, surgiu uma série de convenções que os usuários passaram a utilizar dentro do *tweet.* Por exemplo, os usuários desenvolveram formas de fazer referência a outros usuários, e também meios de mencionar e destacar um tópico (BOYD et. al, 2010).

Usuários do *Twitter* passaram a usar a sintaxe @user, para referenciar utilizadores específicos sendo muito útil quando se deseja direcionar ou mesmo mencionar um perfil. Já quando se deseja mencionar algum tópico que faz menção a algum conteúdo na web, os usuários fazem uso de uma combinação da hashtag (#) com uma palavra-chave (BOYD et. al, 2010). Por exemplo, a hashtag #ppgiufpb, remete ao usuário que o *tweet* possui alguma correlação com acontecimentos do Programa de Pós-Graduação em Informática da UFPB.

Além da marcação de usuários e uso de hashtags, outro elemento explorado por usuários de *Twitter* são os *retweets*, que podem ser definidos como um encaminhamento de uma mensagem já publicada, esta prática é bastante usada para disseminar alguma notícia, ou mesmo propaganda. Um *retweet* é identificado pelo uso da sigla "RT" no início da mensagem.

Além do texto do *tweet* e os elementos pré-definidos citados anteriormente, existem outros campos que podem ser utilizados para a extração de conhecimento do *Twitter*, já que um *tweet* contém também informação relativa ao seu autor, à data da sua criação e coordenadas geográficas. Na Tabela 2 é possível visualizar atributos inerentes a um *tweet*.

Tabela 2 – Tabela Sumária dos Campos de um *tweet*

Campo	Descrição
Autor	Autor do *tweet*
Texto	Corpo do *tweet* (140 Caracteres)
Data de Criação	Data de Publicação do *Tweet*
Coordenadas Geográficas	*Geocode* referente ao *Tweet*
Resposta a um *tweet*	Se for uma resposta a *tweet*, representado pelo id do *tweet*
Destinatário	Em sendo uma resposta, este campo representa o autor do *tweet* anterior
Re*tweet*	Se for um *retweet*
Entidades	Entidades presentes no *tweet*
Contagem de Favorito	Indicador aproximado do número de usuários que marcou o *tweet* como favorito
Língua	Língua detectada automaticamente pelo *Twitter*
Contagem de Re*tweet*	Número de vezes que o *tweet* foi republicado por outros usuários
Fonte	Fonte do *Tweet*

Fonte: MANUEL (2013)

2.1.3 Extração de Dados do *Twitter*

Tendo conhecimento dos dados providos pelos usuários e armazenados no *Twitter* é possível compreender os principais métodos de extração existentes.

2.1.3.1 Extração com Base nos Perfis do Usuário

O perfil do usuário é apenas um componente de uma estratégia de presença global no *Twitter*, porém é um elemento fundamental para a popularidade do perfil. E existem duas razões para tal, são elas (SHAINDLIN, 2010);

- Um perfil completo e cuidadosamente preenchido pode auxiliar as pessoas a encontrar uma organização no *Twitter* através de pesquisa;
- O conteúdo do perfil pode refletir a identidade da pessoa ou organização para com o seu público-alvo.

Para Manuel (2013), o uso dos atributos relacionados ao perfil do usuário citado na Seção 2.1.2.1 são bem explorados em diversos estudos sobre o *Twitter*, mas que estes não são os únicos dados possíveis de extração de um perfil de usuário. As próprias

relações estabelecidas entre o usuário e sua lista de seguidos e seguidores também são objetos de estudo no universo do *Twitter*.

O estudo realizado por Kawak et. al (2010) busca responder alguns questionamentos do tipo, como as pessoas estão conectadas no *Twitter*? Quais os perfis mais influentes? Tais questionamentos motivaram o autor a estudar as características topológicas do *Twitter* e seu poder como um novo meio de compartilhamento de informações.

No referido trabalho, o autor supracitado busca identificar usuários influentes na rede, com base no número de seguidores. É válido destacar que, para isso, apenas dados existentes nos atributos dos perfis dos usuários foram usados. Para que a verificação da influência fosse realizada, se fez necessário escolher um perfil a ser usado como perfil raiz, ou seja, ponto de partida para varredura de perfis, e em um segundo momento, realizar uma busca em largura na árvore, para extração de novos perfis. Dessa forma, escolheram o perfil de Perez Hilton, que possuía na época em que o trabalho foi escrito, mais de um milhão de seguidores.

Em virtude da limitação da API do *Twitter* de 20.000 requisições por hora de um mesmo IP, foram usadas 20 máquinas com IP's distintos, com uma extração de 10.000 requisições por hora. E então, foi possível extrair no período de Julho de 2009 um total de 41.700 milhões de perfis distintos de usuários do *Twitter*. Após tal extração, dentre as várias descobertas, Kawak et. al (2010) identificou que abaixo de Perez Hilton, há uma série de pessoas com grande influência na rede, e que tais pessoas, também podem servir de raiz para outras análises dependendo do contexto em que se deseja estudar.

Para Manuel (2013), outro atributo do perfil do usuário bastante explorado em pesquisas, é a geolocalização do mesmo. O grande problema desta abordagem é destacado por Perez (2011), onde ele afirma que, de um universo de 5.282.657 usuários do *Twitter*, 34% não possui uma localização válida, e os outros 66%, apesar de ter uma localização válida, não há um padrão de uso, uma vez que as pessoas em determinados momentos colocam apenas o estado onde residem, em outros, apenas as cidades, e em muitos casos colocam as cidades de forma abreviada ou mesmo apenas os bairros, dificultando o agrupamento de informações.

Já para os autores Cheng; Caverlle; Lee (2010), que buscam identificar a localização do usuário com base no conteúdo postado, afirmam que em um universo de um milhão de usuários, apenas 26% possui uma localização precisa, os demais, se enquadram no mesmo problema destacado por Perez (2010).

2.1.3.2 Extração com base nos *tweets*

A extração de dados com base em *tweets* é um dos métodos mais explorados nas pesquisas envolvendo tal rede social.

De acordo com dados estatísticos publicados pelo próprio *Twitter,* mais de 500 milhões de *tweets* são enviados por dia *Twitter* (2015). Tais números justificam o fato da extração de dados com base nos *tweets* serem cobiçadas por pesquisadores da área, tendo em sua grande maioria como finalidade entender a dinâmica e o comportamento humano dos usuários desta rede.

Apesar de muito explorada, a obtenção de dados de mídias sociais como um todo, não é uma tarefa trivial, uma vez que não existe uma padronização dos dados publicados, e ainda há restrições de acesso aos dados por parte das ferramentas.

O *Twitter*, mídia social explorada neste trabalho, dispõe de um conjunto de instruções para os desenvolvedores construírem aplicações que dependem de dados do próprio *Twitter*. Este conjunto de instrução é comumente chamado de Interface de Programação de Aplicativos, que vem do Inglês *Application Programing Interface* (API). Atualmente existem basicamente duas API's disponibilizadas gratuitamente pelo próprio *Twitter* para extração de *tweets*, e existe uma outra que permite extrair a massa total de dados, porém não é gratuita. Segue abaixo:

- *Twitter*'s Search API: Esta API foi projetada, para que usuários possam consultar conteúdo no *Twitter* publicado recentemente. Isso pode incluir um conjunto de *tweet*s com palavras-chave específicas, encontrar *tweet*s que fazem referência a um usuário específico, ou encontrar *tweet* de um usuário específico, tudo isso, sem a necessidade de realizar autenticação com algum usuário. (TWITTER, 2013)
- *Twitter*'s Streaming API: Esta API é a amostra em tempo real do *Twitter Firehose*, descrito a seguir. Através dela, é possível capturar uma amostra de 1% de todos os *tweets* públicos (MORSTATTER et. al, 2013). Destina-se a desenvolvedores que estão construindo um produto de mineração de dados, ou que estão interessados em análise de pesquisa. Permite grandes quantidades de palavras-chave a serem especificadas e controladas, recuperam os *tweet*s de determinados usuários que possuem o status público. Para isso, se faz necessário uma conexão HTTP de longa duração (TWITTER, 2013).

- *Twitter*'s *Firehose*: É o nome dado ao massivo fluxo de *tweet*s que fluem em tempo real no *Twitter*, através deste, é possível ultrapassar a barreira imposta pelo Streaming API, podendo acessar até 100% dos *tweets* públicos. Possuem basicamente as mesmas funcionalidades do Streaming API. O que diverge, é o volume de dados capturado. Morstatter et. al (2013) afirma que ao coletar *tweet*s na região em torno da Síria no período entre 14 de Dezembro de 2011 e 10 de Janeiro de 2012, foram coletados 1.288.344 *tweet*s pelo *firehose*, já pelo Streaming API, foram coletados 528.592 *tweet*s. É importante destacar que o conjunto de palavras-chave usadas para a busca foi exatamente o mesmo.

As API's gratuitas que foram mencionadas possuem algumas limitações.

Sousa (2012) afirma que a principal limitação do Search API, por exemplo, é que esta retorna apenas dados postados a um período entre 6 a 9 dias anteriores a data da busca, não permitindo que consultas complexas com mais de 1000 caracteres sejam realizadas. Outra limitação mencionada pelo autor que deve ser levada em consideração, é o fato de ter um limite de quantidade de requisições. A documentação não menciona qual o limite, entretanto, em algum momento da extração, o erro 420, informando que a taxa limite de busca foi atingido, é retornado ao desenvolvedor.

Já *Stream* API tem como uma de suas principais restrições, a necessidade de autenticação com um usuário válido do *Twitter* para ter acesso aos métodos de busca, além disto, esta API não permite que requisições simultâneas vindas de um mesmo usuário sejam feitas.

Para cumprir os objetivos deste trabalho, foi preciso extrair aleatoriamente dados que estão sendo "twitados" pelos usuários, sem necessariamente seguir ou está sendo seguido por outro usuário, para que em um segundo momento, seja possível realizar a classificação de textos.

2.2 Inteligência Artificial

A inteligência artificial consiste em desenvolver mecanismos e dispositivos tecnológicos que possam simular o raciocínio humano, ou seja, a inteligência que é característica dos seres humanos. Para tal, deve-se considerar a construção de máquinas capazes de aprender a partir de experiências passadas, demonstrando que estas podem apresentar significante nível de aprendizado, tornando-as inteligentes (CRISTIANINI; SHAWE-TAYLOR, 2000).

Diversos são os problemas que exigem grande complexidade de solução se comparado as técnicas de programação convencional. Por exemplo, não se sabe como escrever um programa convencional de computador que classifique um texto na categoria a qual ele pertence. Para que isto seja possível, o humano precisa apresentar previamente os elementos (caracteres e textos) individuais para posterior reconhecimento (CRISTIANINI; SHAWE-TAYLOR, 2000).

As redes sociais são alvos de inúmeras pesquisas relacionadas com classificação de textos. Classificar textos neste ambiente é uma tarefa desafiante, uma vez que a estrutura da rede é muito heterogênea, os textos possuem poucas palavras, e pode variar de acordo com a necessidade do que está se buscando (BREVE, 2010). Tal heterogeneidade dificulta o agrupamento de dados, e por isso, necessita muitas vezes de uma amostragem devidamente reconhecida pelo humano para que, posteriormente, o problema venha a ser solucionado por máquinas.

2.2.1 Aprendizado de Máquina

O grande volume de dados sob a forma de texto, e a possibilidade de fácil acesso através da rede mundial de computadores, deu origem a uma demanda de organização de tais conteúdos, difícil de ser suprida através de programação convencional. Para suprir estas demandas, surgiu o aprendizado de máquina (*Machine Learning*).

O Aprendizado de Máquina (AM) é uma subárea de pesquisa da Inteligência Artificial (IA). Breve (2010) afirma que a referida subárea explora projetos e desenvolvimentos de algoritmos que melhoram automaticamente com a experiência, imitando o comportamento do aprendizado dos humanos.

Já os autores Conduta; Magrin (2010) afirmam que o AM é uma subárea da Inteligência Artificial onde a máquina aprende a melhor maneira de resolver um problema à medida que tarefas relacionadas são executadas, permitindo assim alcançar a conclusão do aprendizado.

O principal foco do AM é aprender automaticamente a reconhecer padrões, permitindo, com isso, a tomada de decisão inteligente com base em dados, mesmo em sistemas onde não é possível ou viável encontrar soluções exatas.

De acordo com Breve (2010), muitas aplicações de AM utilizando diferentes técnicas já foram desenvolvidas, variando de programas de mineração de dados que detectam transações fraudulentas em cartões de crédito a sistema que aprendem os hábitos de usuários em redes sociais. Batista (2003) complementa informando que

existem diversos métodos de AM, dentre eles cabe citar o aprendizado supervisionado e não supervisionado.

Os algoritmos da categoria de aprendizado não supervisionado buscam identificar como os dados estão organizados. Os dados de testes consistem apenas de exemplos de entradas, sem rótulos ou valores de saída. O objetivo é encontrar padrões na respectiva amostra, para descobrir como os dados estão organizados, formando agrupamento de exemplos que tenham características similares ou dissimilares. Ou seja, os exemplos são descritos em agrupamentos, ou clusters, de acordo com alguma medida de similaridade (MARTINS, 2003). Por exemplo, um agente aprendendo a dirigir pode gradualmente desenvolver um conceito de dias de bom tráfego e dias de tráfego congestionado, mesmo sem nunca ter recebido rótulos de um "supervisor".

Um algoritmo é tido como supervisionado, quando é possível aprender a partir de um conjunto de exemplos. Um requisito básico para tal aprendizado é que o conceito aprendido deve estar relacionado com fatos observados (BATISTA, 2003). O processo de aprendizado supervisionado é caracterizado pela apresentação do conjunto de dados de treinamento a um algoritmo de aprendizado denominado indutor. Cada elemento pertencente ao conjunto de treinamento possui uma classe específica associada (MARTINS, 2003).

Existem diversas tarefas de aprendizado de máquina associados a métodos de aprendizado supervisionado e não supervisionado. Alguns exemplos de tarefas são a Modelagem de Tópicos que geralmente é realizada com técnicas de aprendizado não-supervisionado e a classificação de textos que geralmente é realizada com técnicas de aprendizado supervisionado. Ambas as tarefas são utilizadas nesta dissertação.

2.2.1.1 Modelagem de Tópicos

Conforme já mencionado, o conhecimento coletivo continua a ser digitalizado e armazenado na WEB em forma de notícias, blogs, artigos científicos, imagens, som, vídeo e redes sociais, tornando-se mais difícil encontrar algo que esteja sendo pesquisado e que seja relevante. Afinal, o humano não consegue ler e posteriormente correlacionar o grande volume de conteúdo eletrônico disponível.

Para suprir esta demanda, pesquisadores da AM desenvolveram a modelagem de tópico. Um modelo de tópico é um modelo generativo, que é baseado na ideia de que documentos são formados por uma distribuição probabilística de tópicos, e que cada tópico é determinado por uma distribuição probabilística de palavras. Sendo assim, é

possível através da modelagem de tópicos obter uma representação vetorial onde no lugar de palavras como atributo tem-se os tópicos (ALENCAR, 2013)

Os algoritmos que implementam a modelagem de tópicos analisam as palavras dos textos originais para descobrir os tópicos através deles. Tais algoritmos não requerem rotulagem dos documentos, os temas surgem a partir da análise dos textos originais, caracterizando um aprendizado não supervisionado (BLEI, 2012)

Considerado como sendo um dos algoritmos mais simples de modelagem de tópicos, o algoritmo LDA (*Latent Dirichlet Allocation*) auxilia a descoberta automatizada destes (BLEI; ANDREW; JORDAN, 2003). Os autores afirmam ainda que a ideia básica do LDA é representar os documentos como uma mistura de tópicos latentes, onde cada tópico é caracterizado por uma mistura de palavras conforme a Figura 1.

Topic 247

word	prob.
DRUGS	.069
DRUG	.060
MEDICINE	.027
EFFECTS	.026
BODY	.023
MEDICINES	.019
PAIN	.016
PERSON	.016
MARIJUANA	.014
LABEL	.012
ALCOHOL	.012
DANGEROUS	.011
ABUSE	.009
EFFECT	.009
KNOWN	.008
PILLS	.008

Topic 5

word	prob.
RED	.202
BLUE	.099
GREEN	.096
YELLOW	.073
WHITE	.048
COLOR	.048
BRIGHT	.030
COLORS	.029
ORANGE	.027
BROWN	.027
PINK	.017
LOOK	.017
BLACK	.016
PURPLE	.015
CROSS	.011
COLORED	.009

Topic 43

word	prob.
MIND	.081
THOUGHT	.066
REMEMBER	.064
MEMORY	.037
THINKING	.030
PROFESSOR	.028
FELT	.025
REMEMBERED	.022
THOUGHTS	.020
FORGOTTEN	.020
MOMENT	.020
THINK	.019
THING	.016
WONDER	.014
FORGET	.012
RECALL	.012

Topic 56

word	prob.
DOCTOR	.074
DR.	.063
PATIENT	.061
HOSPITAL	.049
CARE	.046
MEDICAL	.042
NURSE	.031
PATIENTS	.029
DOCTORS	.028
HEALTH	.025
MEDICINE	.017
NURSING	.017
DENTAL	.015
NURSES	.013
PHYSICIAN	.012
HOSPITALS	.011

Figura 1 - Exemplo de quatro (de 300) tópicos extraídos do *Corpus* TASA
Fonte–(STEYVERS et al., 2006)

Cada tópico descoberto pelo LDA é representado por uma distribuição de probabilidade que determina a correlação entre palavras e um tópico específico. Este processo define uma distribuição de probabilidade das variáveis aleatórias observadas e as ocultas. As variáveis aleatórias observadas são as palavras nos documentos, e as ocultas são as estruturas de tópicos.

Inferir uma estrutura de tópico tendo como base apenas os documentos, é uma tarefa computacionalmente complexa, e amplamente estudada por pesquisadores na área de AM, mais conhecido como o problema de computar a distribuição posterior, ou seja, a distribuição condicional das variáveis ocultas dados os documentos. Como exemplo, um tópico sobre assassinato será aquele que contém palavras relacionadas a assassinato

com maior probabilidade de ocorrência. É importante ressaltar que um tópico que se relacione com qualquer outro assunto distinto conterá palavras sobre assassinato, porém, com probabilidade muito baixa

É importante destacar que o algoritmo LDA é não-determinístico, o que significa que os resultados podem variar mesmo que as entradas (documentos e parâmetros de modelagem) sejam as mesmas. Entretanto, a expectativa é que a variação observada seja pequena e pouco relevante na prática.

A característica mais marcante do LDA é o mínimo de intervenção humana requerida para sua aplicação. Além desta, o fato de os documentos não serem rotulados com tópicos ou palavras-chaves, caracterizam este modelo como não supervisionado (BLEI; ANDREW; JORDAN, 2003).

2.2.1.2 Classificação de Textos

Classificação de textos é uma área onde algoritmos de classificação são aplicados em documentos de texto. A tarefa consiste em atribuir uma classe ou várias classes a um determinado documento, com base em seu conteúdo. Por exemplo, classificar artigos de notícias por assunto, ou simplesmente rotular um artigo como bom ou ruim (SRIRAM, 2010).

De acordo com o número de rótulos a ser empregado na classificação, é possível diferenciar a classificação em dois tipos:

- Classificação binária – Classifica o objeto de entrada em uma de duas classes
- Classificação Multi-classe – Classifica o objeto de entrada em uma de múltiplas classes

A classificação binária é mais estudada e demonstra bons resultados, como mostrado na literatura de aprendizado de máquina. Além disso, a classificação multi-classe pode ser encarada como uma série de classificações binárias que tentam determinar a que classe pertence um item. Nesta dissertação é feita uma classificação binária para determinar se um texto curto pertence ou não a um tópico relacionado com violência.

A tarefa de classificação é normalmente realizada através de técnicas de aprendizado supervisionado. Para a criação de um classificador de texto eficiente, é necessário ter uma amostra conhecida e cuidadosamente rotulada, que é conhecida como o conjunto de treinamento. É comum também o uso de um outro conjunto de

dados, o chamado conjunto de teste, podendo este ser rotulado ou não. O conjunto de treinamento é usado como um conjunto de exemplos para o aprendizado do classificador, enquanto que o conjunto de teste é usado para estimar o poder de generalização e o desempenho do classificador em dados desconhecidos (SRIRAM, 2010).

Vários fatores, além de uma escolha cuidadosa dos dados de treinamento, contribuem para elevar a acurácia de uma dada classificação. A escolha do tamanho do conjunto de dados de teste é muito importante. Se o classificador é alimentado com um pequeno número de documentos para treinar, não é possível adquirir conhecimento substancial para classificar os dados de teste corretamente (SRIRAM, 2010).

Um outro problema mencionado por Sá (2008) é que, caso o volume de textos usados para classificação seja muito grande e ruidosa, é possível recair no problema de alta dimensionalidade de atributos, causando uma grande quantidade de termos redundantes e ou irrelevantes.

Nesta seção foram apresentados conceitos de classificação de textos de forma geral. Porém, o autor Sriram (2010) afirma que, as estratégias consagradas para classificação de textos longos, não necessariamente possuem o mesmo sucesso quando aplicadas aos textos curtos, uma vez que a frequência de ocorrência das palavras é bem inferior.

2.2.1.3 Classificação de Textos Curtos

De acordo com Sriram (2010), considera-se texto os documentos que abrigam o texto propriamente dito. Tais documentos geralmente são ricos em conteúdo e quantidade de palavras, permitindo que técnicas tradicionais de representação como *Bag-of-Words* sejam suficientes para uma boa classificação; entretanto, o mesmo não se aplica para textos curtos.

Uma característica marcante das redes sociais como o *Twitter*, é a possibilidade que os usuários possuem de propagar uma notícia com número restrito de caracteres. Esta característica tem dificultado consideravelmente a classificação de *tweets*, exatamente pela limitação de caracteres que é bem pequena se comparado a um artigo de notícias por exemplo. Outro fator que dificulta bastante a classificação é que, devido à popularização do uso do *Twitter*, erros de ortografia, abreviações de palavras e linguagem gramaticalmente incorreta é bastante frequente, sendo este um complicador a mais na tarefa de classificação (SANKARANARAYANAN et. al, 2009).

Em virtude das limitações apresentadas no parágrafo anterior, pesquisadores têm proposto diferentes métodos de classificar de forma mais eficiente os textos curtos. Alguns métodos serão apresentados no Capítulo 3, a seguir.

3. Trabalhos Relacionados

A partir do levantamento acerca das pesquisas na área de classificação de textos curtos no *Twitter*, foram encontrados alguns trabalhos relacionados. Os mais relevantes para esta dissertação estão listados ao longo deste capítulo.

Um problema muito comum enfrentado nas pesquisas que envolvem classificação de textos curtos é a esparsidade dos dados. Em uma representação *bag-of-words* (comum na classificação de textos), a dimensionalidade dos dados pode chegar às dezenas de milhares (isso depende do tamanho do vocabulário), enquanto que um texto curto publicado em rede social vai ter por volta de 100 palavras. Desta forma, o vetor que representa o texto curto no espaço dos dados será bastante esparso. Diversos estudos tentaram usar outras fontes de dados para superar a esparsidade dos dados e obter uma melhor classificação.

Inspirado pela ideia de usar fontes de dados externas mencionadas acima, os Phan; Nguven; Horiguchi (2008) propõe um quadro geral de classificadores, construídos com base em temas ocultos descobertos a partir de coletas de dados em grande escala. A ideia do quadro é que, para cada tarefa de classificação, seja realizada uma grande coleta de dados na web, chamado "conjunto de dados universal", em seguida, é construído um modelo de classificação, sendo este formado por um pequeno conjunto de dados de treinamento rotulados, e devidamente associado a tópicos obtidos com base no conjunto de dados universal.

De forma análoga aos autores citados no parágrafo anterior, Guo et. al (2013) acreditam que uma boa estratégia para classificação de textos curtos, também é enriquecendo os mesmos com conteúdos externos. Para entender melhor as mensagens do *Twitter*, os autores propõem a tarefa de ligar um *tweet* com uma notícia que é relevante. Para isto, foi criado um conjunto de dados denominado “padrão-ouro” com mensagens que possuem URL para links da CNN e NYTimes, bem como todas as notícias publicadas por ambos em seu perfil. O objetivo é prever a url da reportagem com base no texto de cada mensagem, e com isso indicar que aquele *tweet* é uma notícia.

Percebendo a importância da proliferação de uma notícia de última hora, Sankaranarayanan et. al (2009) investigaram o uso do *Twitter* para construir um sistema de processamento de notícias de última hora a partir de *tweet*s, chamado TwitterStand.

O que o difere do trabalho realizado pelos autores Phan, Nguyen; Horiguchi (2008), é que para eles se faz necessário antes relacionar uma possível notícia com algo que já foi publicado pela mídia, enquanto que Sankaranarayanan et al. (2009) teve a idéia de capturar *tweet*s que correspondem a notícias de última hora sem necessariamente estabelecer o link com uma URL do CNN ou NYTimes, por exemplo. O resultado final é análogo a um serviço de distribuição de notícias, a exemplo do Google News, Bing News e Yahoo. A diferença é que as identidades dos "repórteres" não são conhecidas com antecedência. Além disso, os *tweet*s não são enviados de acordo com um cronograma, afinal eles são publicados conforme a notícia está acontecendo, e tendem a possuir dados bastante ruidosos, uma vez que geralmente os *tweet*s chegam em uma alta taxa de transmissão. O autor em questão meciona que encontrou basicamente três dificuldades:

- Limitação da quantidade de caracteres
- Agrupamento de diferentes *tweet*s falando de uma mesma notícia
- Identificar que o *tweet* é uma notícia

Em linhas gerais, pode-se afirmar que o autor encontrou dificuldades em identificar *tweet*s que são considerados notícias atuais e de agrupa-los com base em sua equivalência.

Inicialmente, Sankaranarayanan et. al. (2009) classificaram os *tweet*s de entrada como sendo lixo ou notícias, onde os *tweet*s "lixo" têm uma boa chance de não estarem relacionados com a notícia, portanto, descartados, enquanto os *tweet*s de notícias têm uma boa chance de serem relacionados efetivamente a uma notícia. O objetivo não é se livrar completamente de ruídos, mas sim, jogar fora tantos *tweet*s possíveis, sem perder muitos relacionados a notícias. Para isso, foi usado um classificador NaiveBayes que é treinado em um *Corpus* de treinamento de *tweet*s que já foram marcadas como notícia.

A fim de assegurar que *tweet*s relacionados a uma notícia não foram classificados como lixo, foi usado um *Corpus* dinâmico e um outro estático. O estático é composto por uma grande coleção de *tweet*s notícias rotulados como tal, bem como uma outra grande coleção de *tweet*s que são lixo. Além do *Corpus* estático, também foi utilizado um dinâmico, com mensagens rotuladas como notícias que são periodicamente obtidas a partir do módulo de agrupamento. Este *Corpus* contém os *tweet*s de notícias recentes pertencentes aos tópicos de notícias no momento. A idéia é que o estático

auxilie na identificação das mensagens notícias sobre temas que não foram encontrados anteriormente, enquanto o dinâmico auxilia na identificação dos *tweet*s sobre eventos atuais.

Um outro método é descrito por Zelikovitz; Hirsh (2002), e combina dados de treinamento rotulado, a um *Corpus* secundário não rotulado que possui relação com documentos mais longos. A este *Corpus* secundário da-se o nome de *background knowledge*. A utilização de um *Corpus* secundário, mesmo que sem rótulo, auxilia consideravelmente a redução da taxa de erros na classificação de textos curtos, uma vez que o problema de esparsidade dos dados é reduzido. Esta abordagem é bastante útil quando a mão de obra para se rotular texto é muito grande, e quando há bastante texto sobre o assunto na internet.

Para atingir o objetivo de seu estudo, os autores citados no parágrafo anterior fazem uso do WHIRL, que é um sistema com alguns operadores especiais para comparação de texto. Para que o uso do Whirl seja útil, é necessário considerar os exemplos de treinamento como uma tabela, o *background knowledge* como outra tabela e finalmente, um exemplo de teste como uma terceira tabela. Desta forma o WHIRL permite fazer uma classificação por similaridade usando as três tabelas, de forma análoga a uma junção, muito utilizada na linguagem SQL. A principal diferença, desta abordagem é que além de fazer uso de um conjunto de dados de treinamento para classificar um novo conjunto de testes, os autores também fazem uso de uma base de conhecimento não rotulada.

Ainda no mesmo trabalho, são apresentados resultados de experimentos aplicados a diferentes *Corpus*, demonstrando o êxito para todos eles, é válido destacar que alguns experimentos não construíram a *backgorund knowledge* a partir da mesma fonte de dados do conjunto de treinamento, porém os resultados não foram afetados.

Já Sriram (2010) descreve que nos serviços de microblogging como o *Twitter*, os usuários podem ficar sobrecarregados com uma quantidade excessiva de dados sem a existência de algum tipo de agrupamento, para isto, o referido autor afirma que a solução para este problema é realizar a classificação dos *tweet*s. Porém, assim como os demais autores, Siriram (2010), também percebeu que textos curtos como os *tweet*s não fornecem ocorrência de palavras suficiente para os métodos de classificação que utilizam abordagens tradicionais, como "Bag Of-Words". Para resolver este problema, o referido autor propõe a utilização de um pequeno conjunto de características específicas de domínio, extraídos do perfil e texto do autor.

A abordagem proposta classifica o texto de forma eficaz a um conjunto pré-definido de classes genéricas, tais como notícias, eventos, opiniões, ofertas e mensagens privadas.

Para selecionar as características, o autor Sriram (2010) leva em consideração oito características (8F), que consistem em uma nominal (autor) e sete características binárias (presença de encurtamento de palavras e gírias, frases de tempo de eventos, palavras opinativas, ênfase em palavras, moeda e sinais de porcentagem, "@ username" no o início do *tweet*, e "@ username" dentro o *tweet*). Na fase de classificação, esses recursos são levados em consideração, a seguir serão descritos como os recursos representam as classes pré-determinadas.

- Notícias – Todos os *tweet*s que não possuem emoções e gírias
- Eventos – Considerando como evento "algo que acontece em um determinado lugar e tempo", a presença do participante, local e informações de tempo pode determinar a existência de um eventono texto. Com isso, foram coletadas informações atualizadas de horário e tempo de duração do evento, que são coletados a partir de um conjunto de *tweet*s com base na observação geral de usuários. Informações dos participantes também são capturadas através da presença do caráter"@"seguido do nome do usuário dentro de *tweet*s.
- Presença de Opiniões – É determinada por uma pesquisa na lista de palavras que consiste de cerca de 3000 palavras opiniosas obtidas a partir da Web.
- Negócio – Foram considerados os *tweet*s com caracteres especiais no texto, tais como moedas e sinais de porcentagem.
- Mensagem Privada – Foram considerados privativos, aqueles *tweet*s que possuem a referência a outro autor no início do *tweet*.

Os resultados experimentais mostram que a abordagem *Bag-of-Words* realiza uma classificação razoável, porém a abordagem proposta “8F” demonstrou melhor desempenho com o conjunto de classes genérica pré-definida. Com o uso de um pequeno conjunto de características discriminativas, a abordagem 8F fornece uma linha de base para classificar novos *tweet*s em linha com uma precisão melhor. No entanto, os dados mais ruidosos podem degradar o desempenho da abordagem proposta; logo, técnicas de remoção de ruído são essenciais.

Conforme apresentado ao longo deste capítulo, pode-se observar que diversos foram os estudos realizados acerca da classificação de textos curtos nos últimos dez

anos (PHAN; NGUYEN; HORIGUCHI, 2008). Entretanto, dentre as pesquisas já mencionadas, foi identificado que algumas possuem uma certa similaridade com a melhoria proposta nesta dissertação por fazerem uso de uma base de dados secundária para "enriquecer" os dados do *Corpus* principal. São elas:

- Método que combina dados de treinamento rotulado, mais um *Corpus* secundário não rotulado que possuem relação com documentos mais longos (ZELIKOVITZ e HIRSH, 2002);
- Método de classificação formado por um pequeno conjunto de dados de treinamento rotulados e um rico conjunto de rótulos descobertos a partir da coleta de dados no *Twitter* (PHAN; NGUYEN; HORIGUCHI, 2008);
- Método que propõe criar um conjunto de dados denominado "padrão-ouro" com *tweet*s que possuem URL para notícias. E em seguida, confrontar com outro conjunto de *tweet*s, buscando prever a url da reportagem com base no texto de cada *tweet* (GUO et. al, 2013).
- Os autores Sankaranarayanan et. al (2009) fazem uso de um *Corpus* dinâmico e outro estático. O *Corpus* estático é composto de uma grande coleção de *tweet*s notícias rotulados como notícias, bem como uma outra grande coleção de *tweet*s que são lixo. Já o *Corpus* dinâmico, contempla *tweet*s classificados como notícias recentes em fases de treinamentos anteriores.
- Utilização de metadados do próprio *tweet* base complementar, para que esta sirva de complemento das mensagens emitidas pelos usuários do *Twitter* (SRIRAM, 2010).

O que chama mais atenção, é o fato de que apesar de todos os autores mencionarem a criação de uma base secundária para enrriquecer os *tweet*s, apenas Zelikovitz; Hirsh (2002), destacam que fazem uso de *background knowledge* para classificar de forma mais eficiente textos curtos, porém, entende-se que as demais estratégias por fazerem uso de uma base secundária também adotam a estratégia de utilização de uma *background knowledge*. Desta forma, por identificar um número elevado de trabalhos fazendo usando tal base, e demonstrando êxito em seus experimentos, optou-se em fazer uso da *background knowledge* no intuito não de enriquecer os dados a serem classificados, mas sim de realizar uma comparação simples entre o texto a ser classificado, e as palavras-chave que compõem a base secundária.

Outra observação a ser feita sobre os trabalhos correlatos é o fato de apenas o Sriram (2010) fazer uso de dados do próprio *tweet* para compor a base secundária, tal observação, aliado ao bom desempenho demonstrado nos experimentos e ainda a comprovação da possibilidade de aplicação deste método para outros domínios, fez com que a estratégia de classificação aqui proposta também utilizasse uma *background knowledge* formada por dados do próprio *Twitter*.

4. Classificador de Texto Curto Proposto

Com base nas estratégias exploradas por outros pesquisadores, identificou-se a possibilidade de propor uma melhoria na estratégia de classificação de tópicos em textos curtos usando uma forma de *background knowledge* diferente das que já foram apresentadas nesta dissertação.

A principal diferença entre o método de classificação proposto e os demais mencionados, é que no método proposto, a *background knowledge* é utilizada como artefato para estabelecer uma classificação do *tweet*, e nos demais métodos, a referida base é utilizada unicamente para "enriquecer" os dados a serem treinados por um dado classificador. Na Figura 2 é possível visualizar, de forma mais clara, o processo de classificação de texto curto proposto como um todo.

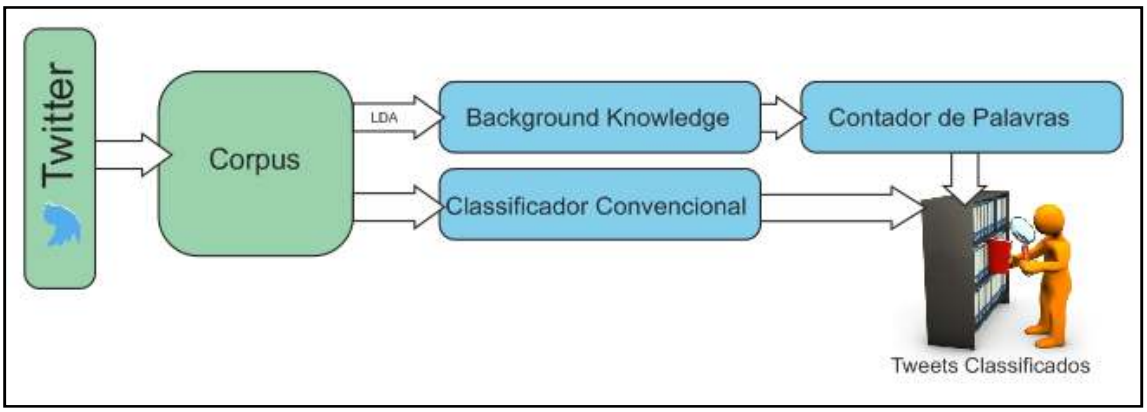

Figura 2 – Método de Classificação de Texto Curto Proposto

Fonte – Próprio Autor

É possível observar ainda que na Figura 2 há elementos do processo de classificação em verde e azul. Em verde são os elementos que servem como entrada para o classificador proposto, tais elementos serão detalhados na Seção 4.1, já os elementos em azul são aqueles que efetivamente formam o classificador, sendo estes detalhados na Seção 4.2.

4.1 Construção do *Corpus*

Os elementos em verde da Figura 2 representam os dados de entrada que serão submetidos ao classificador. Nesta seção será detalhado como se deu a construção do *Corpus* utilizado nesta dissertação. Desta forma, considerando que foi dado um *zoom* no

elemento "*Corpus*" da Figura 2, foram identificados os elementos apresentados na Figura 3.

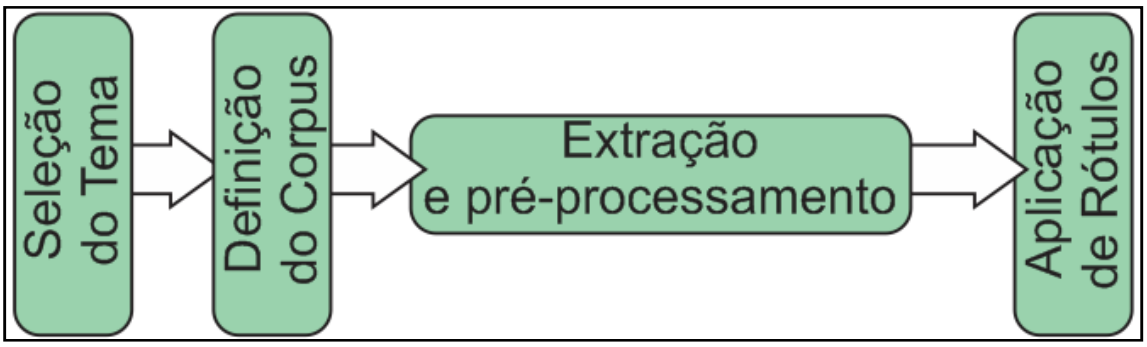

Figura 3 – Etapas para construção do *Corpus*

Fonte – Próprio Autor

- Seleção do tema – Consiste em identificar qual será o assunto que se deseja extrair para construir um determinado *Corpus*.
- Definição do *Corpus* – Apresenta definições sobre o *Corpus* usado neste trabalho, bem como a metodologia empregada para criação do mesmo;
- Método de extração e pré-processamento de dados do *Twitter* – Extração de *tweets*, e remoção de possíveis ruídos;
- Aplicação de Rótulos – Classificação prévia de parte dos dados extraídos

4.1.1 Seleção do Tema

No ano de 2012, segundo Waiselfisz (2014), 112.709 pessoas morreram em situações de violência no país. O número equivale a 58,1 habitantes a cada grupo de 100 mil, e é o maior da série histórica do estudo, divulgado a cada dois anos.

Segundo o responsável pela pesquisa citada acima, ainda não é possível concluir se o que ocorreu no ano de 2012. Trata-se de um surto de rápida passagem ou se de fato trata-se de um novo ciclo ou uma nova tendência. Greves de agentes das forças de segurança ou ataques de grupos criminosos organizados são fatores que contribuíram consideravelmente para o aumento da violência. Pode-se afirmar que atualmente há uma disseminação da violência nas diferentes regiões e cidades do País.

Entre 2002 e 2012, os quantitativos só não cresceram no Sudoeste do Brasil. As regiões Norte e Nordeste experimentaram rápido aumento na violência. No Norte, por exemplo, foram registrados 6.098 homicídios só em 2012, mais que o dobro dos 2.937 verificados em 2002. Já no Nordeste, estados como Maranhão, Bahia e Rio Grande do Norte mais que triplicaram os homicídios (WAISELFISZ, 2014).

As estatísticas apresentadas através dos mapas de violência já publicados possuem significante relevância para área de Segurança Pública do Brasil, isso se confirma pelo prêmio de Direitos Humanos entregues pela Presidenta Dilma Rousself ao autor desta série de estudos (Waiselfisz, 2014).

Mesmo percebendo tamanha significância na série publicada, observa-se a necessidade de uma estatística atualizada com uma freqüência maior, afinal, o Mapa da Violência sempre faz referência a fatos ocorridos há dois anos antes de uma nova publicação da série. O amplo intervalo de tempo para divulgação de novas análises estimulou a escolha do tema "Violência" para ser o assunto do *Corpus* construído e utilizado nesta dissertação.

4.1.2 Definição do *Corpus*

De acordo com Aluísio; Almeida (2006) ocorreu uma mudança no conceito de *Corpus*. Eles entendem que *Corpus* seja um "conjunto, tão variado quanto possível, de enunciados efetivamente emitidos por usuários da referida língua em determinada época" Ducrot; Todorov (2001) apud Aluísio; Almeida (2006).

Para Trask (2004) apud Aluísio; Almeida (2006), *Corpus* é "um conjunto de textos escritos ou falados em uma língua, disponível para análise".

Segundo Sinclair (2005) apud Aluísio; Almeida (2006), responsável pelo trabalho pioneiro na área de léxico com o dicionário COBUILD, o primeiro a ser compilado a partir de um *Corpus c*omputadorizado, um *Corpus* é uma coleção de pedaços de texto em formato eletrônico, selecionados de acordo com critérios externos para representar uma fonte de dados para a pesquisa lingüística.

Ao observar essas definições, percebe-se uma diferença sutil entre as mesmas, ou seja, os dados devem estar em formato eletrônico. O que significa dizer que uma grande quantidade de livros, ou de revistas, ou mesmo de textos impressos não é considerada *Corpus* por Sinclair (2005) apud Aluísio; Almeida (2006), já que tais dados não estão em um formato que possam ser processados pelo computador.

Para construção do *Corpus* computadorizado, é necessário observar um conjunto de requisitos que impactarão na validade e confiabilidade da pesquisa baseada no mesmo (ALUÍSIO; ALMEIDA, 2006). Para o *Corpus* elaborado nesta dissertação, foi considerado que os seguintes requisitos impactam na sua construção.

- Autenticidade – Os textos devem ser escritos em linguagem natural. Além disso, os textos devem ser escritos por falantes nativos, neste caso, Português do Brasil (SARDINHA, 2000 apud ALUÍSIO; ALMEIDA, 2006);
- Balanceamento – Deve existir um equilíbrio de gêneros discursivos (informativo, científico, religioso, etc.), ou de autores ou de todos esses itens juntos, demonstrando que os textos foram escolhidos criteriosamente, isto se faz necessário pelo fato de não conhecer com precisão quem fala sobre violência no país;
- Diversidade – O *Corpus* deve considerar a diversidade de dialetos, uma vez que diferentes termos podem ser usados para um mesmo ato de violência;

Após identificar os requisitos considerados relevantes para a elaboração do *Corpus* relacionado à violência, foi possível iniciar a sua construção propriamente dita. Sendo assim, para compilação do referido *Corpus* foram realizados os seguintes passos:

- Projeto de *Corpus* – Abrange a indicação de que tipo de texto deve ser extraído, levando em consideração os requisitos discutidos;
- Captura/Compilação e Manipulação do *Corpus* – Consiste no armazenamento em arquivos predeterminados de todos os textos selecionados e no pré-processamento dos dados, que consiste em fazer uma remoção de ruído nos textos extraídos;
- Anotação – Apesar de existirem as anotações estruturais, e lingüísticas, neste trabalho foi explorada apenas a estrutural, uma vez que esta basicamente tem a finalidade de separar um texto de seus metadados, em se tratando de *tweet*, os seus metadados podem ser o horário, localização, autor do *tweet*, etc.

Na primeira etapa de Projeto do *Corpus*, os seguintes requisitos foram estabelecidos para a seleção dos textos:

- Os textos devem necessariamente ser escritos originalmente em português;
- Os textos selecionados devem abranger diferentes domínios, com a condição de atender o requisito de balanceamento e diversidade;
- Seleção de uma quantidade de textos suficientes para a elaboração de um dicionário que contemple uma variedade de palavras que esteja relacionada com o termo violência, por exemplo, roubo, estupro e latrocínio.

Uma vez definido o *Corpus*, e indicado os textos que seriam extraídos, deu-se início a etapa de extração dos dados do *Twitter* e pré-processamento dos dados.

4.1.3 Métodos para Extração e Pré-processamento de Dados

Nesta dissertação, foram verificados diversos métodos de extração de dados do *Twitter*. Entretanto, é importante destacar que, na maioria dos trabalhos analisados percebeu-se que o nível de detalhes no qual é descrito o método de extração de *tweet*s é demasiadamente vago, se restringindo apenas a informar que os *tweet*s foram extraídos da *Public Timeline* dos usuários (MANUEL, 2013). Sendo assim, como uma das contribuições desta dissertação é o *Corpus* construído, então será apresentado com mais detalhes a estratégia de extração adotada.

Para extração dos dados foi construído um programa que faz uso da API do *Twitter* versão 1.1 (TWITTER, 2013), em conjunto com o *Twitter4J* na versão 3.0.5 (TWITTER4J, 2013), uma biblioteca *Java* que permite integrar aplicações com os serviços do *Twitter*. O extrator consiste basicamente em executar a rotina de consultas e persistir os dados em uma tabela do banco de dados *MySql* conforme Tabela 3.

Descrição da Coluna	**Nome da Coluna**
Identificador do Usuário que postou o *tweet*	User_id
Cidade do usuário que postou o *tweet*	Ds_cidade
Palavra chave mencionada	Ds_keyword
Latitude do *tweet* postado	Nu_latitude
Longitude do *tweet* postado	Nu_longitude
Nome do usuário no *tweet*	Ds_nome
Mensagem Original, (*tweet*) propriamente dita	Ds_*tweet*
Mensagem pré-processada	Ds_*tweet*_modificado
Indicação se o *tweet* é relacionado ao crime ou não	Fl_crime

Tabela 3 – Estrutura da tab. "tb_key_word_mention" que armazena os dados dos *tweet*s

Fonte: Próprio Autor

Como já mencionado, o extrator construído para ser usado nesta dissertação fez uso do Twitter4j e *Twitter* API, usando o método *Search* para extrair os *tweet*s. Uma limitação desta API é que só podem ocorrer, no máximo, 180 requisições em cada período de 15 minutos, sendo extraídos no máximo 100 *tweet*s a cada nova requisição.

Após isso, é necessário monitorar os limites impostos pelo *Twitter* para que a extração não pare.

Tendo em vista que para realização dos experimentos e avaliação dos resultados é necessário uma quantidade elevada de *tweet*s, foi razoável implementar no extrator uma pequena lógica capaz de extrair uma quantidade ampla de *tweet*s, sem a necessidade de intervenção humana no momento em que o máximo de *tweet*s predeterminado pela API fosse extrapolado. Tal programa realiza a extração dos 100 primeiros *tweet*s, posteriormente captura o menor TwitterID, que são classificados em ordem cronológica aproximada, em seguida, faz uma nova requisição para o *Twitter*, de forma análoga à primeira, entretanto, é necessário informar que o TwitterID máximo que se deseja recuperar é um menor do que o menor já extraído, desta forma serão retornados os *tweet*s de um dado momento do tempo para trás. Este processo se repete até atingir os limites impostos pelo *Twitter*, após isso, o programa monitora o momento em que pode ser feita uma nova requisição. Uma vez liberada a nova requisição, o extrator passa a extrair *tweets* novamente e armazenar no banco de dados (MCGUINNESS, 2013).

Além do programa implementado para extrair um número maior de *tweets*, foi necessário aplicar dois filtros, sendo um para capturar apenas os *Tweets* escritos em Português, e outro para não considerar os Re-*Tweets*, e ainda remover os *Emojions*[1] antes mesmo de inseri-los na base de dados. Para isto foi construída uma expressão regular com a finalidade de desprezar todos os caracteres que não compreendem os caracteres com códigos ASCII que vão de "x20" ao "x7e" e do "x80" ao "xff". A Tabela 4 apresenta um exemplo de *tweet* antes e depois da remoção dos *Emojions*

Tweet com *Emojions*	**Tweet sem *Emojions***
Se você não der este bjo eu roubo ☺☺☺☺!!!	Se você não der este bjo eu roubo!!!

Tabela 4 – Exemplo de *tweet* antes e depois da remoção dos *Emojions*

Fonte: Próprio Autor

[1] Biblioteca de figuras prontas (Smilles, Coração, etc), amplamente utilizada em redes sociais.

Após implementação do programa para extração de *tweets*, deu-se início ao processo de extração, que por usa vez foi realizada de duas formas, sendo a primeira com palavra-chave e a segunda sem palavra-chave. Na Tabela 5 são mostradas as quantidades de *tweets* extraídos por palavra-chave.

Palavra-Chave	Quantidade
Latrocínio	1075
Homicídio	5390
Assalto	13409
Roubo	10101
Estupro	3114

Tabela 5 - Quantitativo de *Tweets* Extraídos por palavra-chave

Fonte: Próprio Autor

Após extração por palavra-chave, foram extraídos *tweets* de forma aleatória, sem considerar as palavras chaves. Três extrações foram realizadas utilizando este critério: a primeira considerou *tweets* com geolocalizacao, a segunda sem geolocalização, e uma última sem nenhum tipo de filtro. A Tabela 6 demonstra o quantitativo extraído.

Método de Extração	Primeira Extração
Com Geolocalizacao	227760
Sem Geolocalizacao	3250820
Aleatório	479088

Tabela 6 - Quantitativo de *Tweets* extraídos sem palavra-chave

Fonte: Próprio Autor

Tendo os *tweets* armazenados no banco de dados MySQL, iniciou-se a etapa de pré-processamento que consiste na execução dos seguintes passos:

- Remoção de URL's
- Remoção de acentuação
- Atualização de todas as mensagens para UPPERCASE
- Remoção de registros em duplicidade

É importante destacar que os passos apresentados foram executados na mesma ordem, afinal, para que a etapa de remoção de duplicidade seja eficaz, é necessário que os dados estejam devidamente normalizados. A Tabela 7 ilustra exemplos de *tweets* antes e depois da fase de pré-processamento.

	Tweet
Antes	Faculdade de Medicina da USP tem dois novos casos de estupro http://t.co/R18x8sLyWg
Depois	FACULDADE DE MEDICINA DA USP TEM DOIS NOVOS CASOS DE ESTUPRO
Antes	Porto Nacional: Jovem investigado por tráfico, roubo e receptação é preso.
Depois	PORTO NACIONAL: JOVEM INVESTIGADO POR TRAFICO, ROUBO E RECEPTACAO E PRESO.
Antes	@mayormillls e um casal gay não pode mas traição roubo mentiras mortes pode, a gente se acostumou com a violência mas não com o amor
Depois	MAYORMILLLS E UM CASAL GAY NAO PODE MAS TRAICAO ROUBO MENTIRAS MORTES PODE, A GENTE SE ACOSTUMOU COM A VIOLENCIA MAS NAO COM O AMOR
Antes	Tentativa de Latrocínio é registrada em Maravilha.
Depois	TENTATIVA DE LATROCINIO E REGISTRADA EM MARAVILHA.
Antes	@coefigueiredo não consigo dormir aí tem q fazer algo e é perturbar vocês no snap hahaha
Depois	COEFIGUEIREDO NAO CONSIGO DORMIR AI TEM Q FAZER ALGO E E PERTURBAR VOCES NO SNAP HAHAHA

Tabela 7 – *Tweets* antes e depois da fase de pré-processamento

Fonte: Próprio Autor

Ao término da etapa de pré-processamento, obtiveram-se os quantitativos apresentados nas Tabelas 8 e 9.

Palavra-Chave	Qtd Sem Duplicidade
Latrocínio	274
Homicídio	2187
Assalto	4768
Roubo	5932
Estupro	1609

Tabela 8 - Qtd. de *Tweets* extraídos por palavra chave, após remoção de duplicidade

Fonte: Próprio Autor

Método de Extração	Quantidade
Com Geolocalizacao	210100
Sem Geolocalizacao	2205433
Aleatório	479088

Tabela 9 - Qtd. de *Tweets* extraídos sem palavra-chave, após remoção de duplicidade

Fonte: Próprio Autor

4.1.4 Aplicação de Rótulos

A fase de construção do *Corpus* é finalizada nesta etapa, que consiste basicamente em aplicar rótulos aos *tweets* extraídos. Esta fase é de fundamental

importância para o estabelecimento do *Corpus*, afinal, através destes rótulos foi possível treinar um classificador e medir o seu desempenho de classificação.

Os *tweets* foram rotulados manualmente por um único humano, indicando se um *tweet* tem relação com a violência ou não. Para isso, foi alimentada a coluna fl_crime igual "S" para sim, caso o *tweet* tivesse correlação com violência e, "N" para não. A Tabela 10 consolida o quantitativo de dados rotulados ou não, que compõem o *Corpus* proposto.

	Qtd. Extraída	Qtd. Remoção de Duplicidade	Qtd. Rotulado
Latrocínio	1075	274	274
Homicídio	5390	2187	1068
Assalto	13409	4768	1406
Roubo	10101	5932	999
Estupro	3114	1609	994
Com Geolocalizacao	227760	210100	1406
Sem Geolocalizacao	2806820	2205433	2825
Aleatório	730010	479088	9995

Tabela 10 - Quantitativo de *tweets* Classificados Manualmente

Fonte: **Próprio Autor**

4.2 Classificador Proposto

Nesta seção serão detalhados os elementos que compõem o classificador proposto, que são os elementos em azul da Figura 2.

- *Background Knowledge*
- Contador de Palavras
- Classificador convencional

Assim como em alguns trabalhos correlatos apresentados no Capítulo 3, a finalidade da *background knowledge* é fazer parte do processo de classificação de textos curtos. Porém, nesta dissertação, a *background knowledge* é usada no classificador "Contador de Palavras", para obter um *score* parcial de classificação, em seguida, o seu resultado será combinado com o resultado do "Classificador Convencional", diferente das abordagens já apresentadas nos trabalhos correlatos, onde os respectivos autores fazem uso da *background knowledge* para enriquecer os "textos curtos", e posteriormente, submeter a um único classificador.

Para construir a *background knowledge* utilizada no classificador proposto nesta dissertação, é necessário realizar uma modelagem de tópicos nos dados, imediatamente

após a etapa de pré-processamento da construção do *Corpus* (ou seja, utilizam-se os dados sem rótulos). Tal modelagem é realizada utilizando o algoritmo LDA implementado no Mallet[2]. Após realização da modelagem, é obtido como saída um arquivo contendo N tópicos e M palavras por tópico, onde N e M são parâmetros ajustáveis do Mallet. Um resultado típico é mostrado na Figura 4 (para N = 20 e M = 9). Note que a saída do algoritmo LDA é um conjunto de tópicos, no qual cada tópico é representado por um conjunto de palavras co-relacionadas. Como é um algoritmo totalmente não-supervisionado, o LDA não detecta que tópicos estão presentes no *Corpus* de entrada, apenas agrupa palavras por tópico.

```
0   --> causa noticia crime seguem historia mina laden lindo ajuda
1   --> dilma pais namorada sul delegacia registra uol diretor frente
2   --> assalto homicidio estupro tentativa pm casos anos homem via
3   --> rio latrocinio tiros preso tentativa mesmo suspeitos brasil ir
4   --> morte durante puta assaltos dele lei serio leia ve
5   --> quase betalab direito dinheiro prova atriz mata fuga pensar
6   --> rua terca militar estadao miliciano fase alta aconteceu cama
7   --> roubo sai santana culposo suicidio criancas alto trabalhar gato
8   --> registrado indiara dela sapatos joao buri pena legal queimada
9   --> amo tiro dormir cometer nega amanha passa preciso caralho
10  --> dupla campinas usp imagens tentavam reage justica indios preto
11  --> gente operacaobetalab bandidos vereador entao volta ficar policiais civil
12  --> queria formato civil amigos mortos feliz festa pa testemunhas
13  --> cidade arma qualquer seja pai cade disse pois real
14  --> policia medicina vida troca filho pessoas vitimas pessoa suspeito
15  --> morre policial olha termina feridos evitem mesma manaus muita
16  --> sair sinto adolescente haha bebe janeiro ligar uberl recebe
17  --> roubo sp bem cara pessoas amor deus julgado coletivo
18  --> paulo armada emabiggestfans cabelo certo esclarecer doloso acordar garoto
19  --> bom durante carro flerte tipo comigo falando supremo claro
```

Figura 4 - Saída da Modelagem de Tópicos feita com o Mallet
Fonte: Próprio Autor

Após realização da modelagem de tópicos, se faz necessário indicar as palavras que são mais relevantes para o tópico que está sendo classificado; para o caso do presente trabalho, foi escolhido "violência". Os termos contidos nos tópicos selecionados serão os elementos que irão compor a base aqui denominada de *background knowledge*. Uma vez concluída a etapa de construção da *background knowledge*, pode-se usar as palavras-chave selecionadas no classificador "Contador de Palavras".

O "Contador de Palavras" realiza uma comparação simples de texto, verificando a existência ou não das palavras-chave da *background knowledge* nos *tweets*. Após o

[2]MALLET é um pacote baseado em Java para modelagem de tópicos usando LDA.

processamento, tem-se como saída desta etapa o quantitativo de palavras encontradas para cada *tweet*, saída esta utilizada para calcular o *Score* Final de cada *tweet* para o tópico em questão.

Na etapa denominada "Classificador Convencional" foi utilizado a API do Weka[3] para treinar um classificador NaiveBayes. Esse tipo de classificador foi escolhido por ter normalmente um bom desempenho na classificação de texto, mas o esquema proposto de combinação com uma contagem de palavras pode ser usado com qualquer classificador de texto. Após aplicação do "Classificador Convencional", cada *tweet* recebe duas probabilidades previstas pelo classificador, sendo uma indicando a probabilidade deste pertencer à classe chamada de positiva (que representa, no caso deste trabalho, o fato do *tweet* estar relacionado ao tópico "violência"), e a outra indica a probabilidade de pertencer à classe negativa (ou seja, não estar relacionado ao tópico "violência").

Tendo os valores referentes ao "Contador de Palavras" e as probabilidades referente ao "Classificador Convencional", foi estabelecida uma fórmula para calcular o *Score* Final da Classificação, como uma média ponderada da saída dos dois módulos de classificação conforme Figura 5.

$$K = \frac{w_1 \times P_{nb} + w_2 \times C}{w_1 + 4w_2}$$

Figura 5 - Cálculo do Score Final para classificação do Tweet

Fonte: Próprio Autor

Onde P_{nb} é a probabilidade (para a classe positiva) de saída do classificador NaiveBayes, e C é o contador de palavras do tópico presente no *tweet*. O peso 4 multiplicando w_2 no denominador é baseado na idéia que a ocorrência de 4 palavras em um *tweet* é um número alto. W_1 e W_2 dizem respeito à atribuição de peso a probabilidade de saída do classificador NaiveBayes e do contador de palavras respectivamente.

Por exemplo, na Figura 6 se $w_1 = 1$ e $w_2 = 0{,}25$ (o peso 0,25 é pensado para fazer com que C = 4 contribua com um valor 1 na média), teríamos:

[3]Weka é um pacote de Software cujo objetivo é agregar algoritmos provenientes de diferentes abordagens/paradigmas na sub-área da inteligência artificial dedicada ao aprendizado de máquina.

$$K = \frac{P_{nb} + 0.25 \times C}{2}$$

Figura 6 - Exemplo de utilização da fórmula

Fonte: Próprio Autor

É preciso definir um limiar para o valor de K, de forma que se K for maior que esse limiar, o *tweet* é classificado como um *tweet* relacionado com o tópico "violência"; caso contrário, se K for menor que o limiar, o *tweet* não está relacionado com o tópico. Cabe-se destacar, que o valor de K, pode variar de *Corpus* para *Corpus*, não tendo este um valor único que seja aplicado a toda e qualquer classificação de texto curto.

5 Experimentos e Análises

O esquema de classificação proposto consiste em duas partes principais: primeiro, o uso de um modelo de tópicos sobre um *Corpus* representativo de *tweets* para determinar um conjunto de palavras-chave importantes para um tópico; segundo, o uso das palavras-chave descobertas para melhorar o desempenho de classificação de tópico em textos curtos, principalmente quando este ocorre raramente em textos reais.

Neste capítulo são descritos experimentos que foram realizados para validar a estratégia proposta. Os experimentos aplicados foram norteados pelos objetivos propostos desta dissertação, e consiste na aplicação da classificação de textos curtos apresentado no Capítulo 4, e aplicado no *Corpus* "Violência" detalhada no mesmo capítulo.

Inicialmente, aplicou-se a modelagem de tópicos para obter uma lista de palavras-chave agrupadas por tópico. Por inspeção direta da lista resultante, pretendeu-se verificar se o modelo de tópicos é capaz de isolar um ou mais grupos de palavras relacionados à "violência"; e escolher um ou mais grupos de palavras da lista para compor a *background knowledge* e usar no classificador proposto.

Com a *background knowledge* criada, o classificador proposto pôde ser testado em comparação com um classificador NaiveBayes simples, a fim de verificar se ocorre melhoria no desempenho de classificação. O classificador proposto também foi comparado com um classificador de palavras-chave simples, para verificar se o melhor desempenho resulta da combinação dos dois módulos de classificação.

5.1 Background knowledge

A *background knowledge* foi construída através da criação de um modelo de tópicos baseado no *Corpus* criado (descrito no Capítulo 4).

Vários experimentos poderiam ser realizados modificando apenas os parâmetros da modelagem de tópicos e, conseqüentemente, os dados que compõem a *background knowledge*. Entretanto, para esta dissertação, foram utilizados os mesmos parâmetros de modelagem para todos os experimentos, parâmetros estes mostrados na Tabela 11.

Nº de Tópicos	Nº Palavras Por Tópico	Nº de Iterações
20	10	1000

Tabela 11 - Parâmetros utilizados no Mallet para modelagem de tópicos

Fonte: Próprio Autor

Uma vez estabelecido os parâmetros para modelagem, foi realizado a primeira execução do Mallet (ferramenta para modelagem de tópicos), e conforme destacado em vermelho na Figura 7 identificou-se a presença de palavras e também tópicos que não contribuíam semanticamente para o tema "violência".

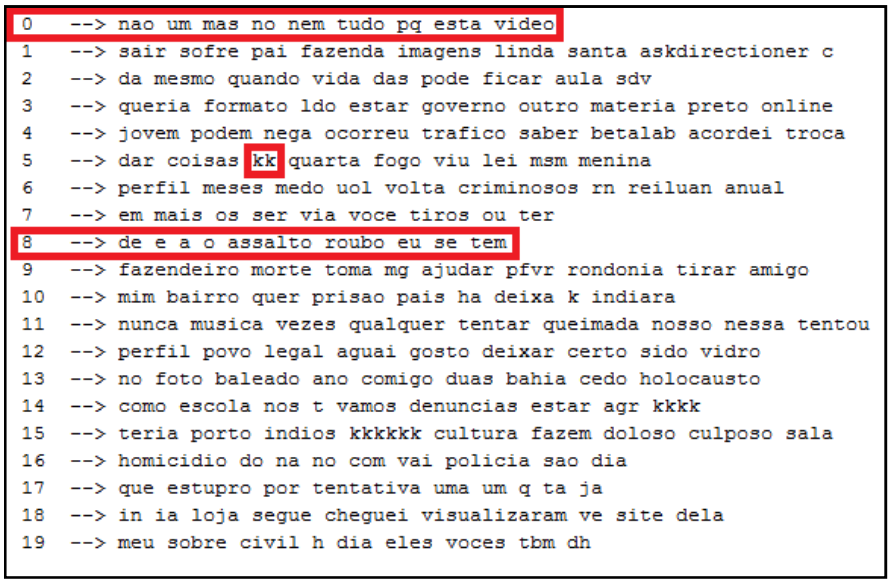

Figura 7 - Resultado da execução do Mallet sem Stop Words
Fonte: Próprio Autor

Visando diminuir a ocorrência de palavras não compatíveis semanticamente com o tema violência, foi necessário fazer uso de uma lista de *stop words*[4] em português do Brasil, para que o modelo de tópicos se concentrasse em palavras com maior conteúdo semântico relacionado a algum tópico.

Desta forma, foi construída uma lista contendo 458 termos, considerando desde artigos, pronomes e afins, até dados ruidosos dos *tweets*, a exemplo "kkkk" ou "rsrsrsrs". O conhecimento do *Corpus* foi crucial neste aspecto, para que os resultados da modelagem não influenciassem negativamente o classificador proposto como um todo. Ao término da construção da *stop words*, e nova execução do Mallet, obteve-se como saída os tópicos apresentados na Figura 8.

[4] Palavras que podem ser consideradas irrelevantes para a modelagem de tópico

```
0   --> causa noticia crime seguem historia mina laden lindo ajuda
1   --> dilma pais namorada sul delegacia registra uol diretor frente
2   --> assalto homicidio estupro tentativa pm casos anos homem via
3   --> rio latrocinio tiros preso tentativa mesmo suspeitos brasil ir
4   --> morte durante puta assaltos dele lei serio leia ve
5   --> quase betalab direito dinheiro prova atriz mata fuga pensar
6   --> rua terca militar estadao miliciano fase alta aconteceu cama
7   --> roubo sai santana culposo suicidio criancas alto trabalhar gato
8   --> registrado indiara dela sapatos joao buri pena legal queimada
9   --> amo tiro dormir cometer nega amanha passa preciso caralho
10  --> dupla campinas usp imagens tentavam reage justica indios preto
11  --> gente operacaobetalab bandidos vereador entao volta ficar policiais civil
12  --> queria formato civil amigos mortos feliz festa pa testemunhas
13  --> cidade arma qualquer seja pai cade disse pois real
14  --> policia medicina vida troca filho pessoas vitimas pessoa suspeito
15  --> morre policial olha termina feridos evitem mesma manaus muita
16  --> sair sinto adolescente haha bebe janeiro ligar uberl recebe
17  --> roubo sp bem cara pessoas amor deus julgado coletivo
18  --> paulo armada emabiggestfans cabelo certo esclarecer doloso acordar garoto
19  --> bom durante carro flerte tipo comigo falando supremo claro
```

Figura 8 - Tópicos Modelados e Tópicos Selecionados utilizando *Corpus* "Violência"

Fonte: Próprio Autor

Ainda sobre a Figura 8, os tópicos 2 e 3 estão destacados em vermelho, pois representam os tópicos selecionados para compor a *background knowledge* utilizada neste experimento.

5.2 Classificador NaiveBayes

O esquema de classificação de tópicos proposto (Figura 2, Capítulo 4) tem como um de seus componentes um classificador que serve como base do processo. Nesta dissertação foram utilizados classificadores NaiveBayes, que costumam ter bom desempenho em tarefas de classificação de textos. Para treinar o classificador, que segue um processo típico de aprendizado supervisionado, deve ser usado um *Corpus* rotulado.

O *Corpus* descrito no Capítulo 4 foi utilizado para treinar classificadores NaiveBayes para classificação de textos em relação ao tópico "violência", foram utilizados os segmentos do *Corpus* separados por palavra-chave e localização, sendo um classificador treinado para cada segmento, bem como um classificador para o segmento de *tweets* aleatórios e um classificador treinado com todos os segmentos do *Corpus*. Em cada segmento, separou-se 2/3 das instâncias rotuladas para servir como conjunto de treinamento, o restante ficando designado como conjunto de testes. No total, foram construídos nove classificadores NaiveBayes treinados a partir dos segmentos do *Corpus*.

5.3 Classificador Proposto

Tendo o classificador NaiveBayes usado como base no esquema de classificação proposto, e a *background knowledge* com a lista de palavras-chave, a construção do classificador proposto depende apenas do módulo contador de palavras e dos parâmetros para combinação dos dois módulos (conforme visto no Capítulo 4). Após alguns testes empíricos, foram utilizados os valores para os parâmetros segundo a Tabela 12.

Peso NaiveBayes	Peso Contador	Média de palavras encontradas	Limiar de Classificação
1	0.25	4	0.8

Tabela 12 - Parâmetros utilizados no classificador

Fonte: Próprio Autor

5.4 Testes dos classificadores

Para cada classificador treinado, os testes foram realizados com o conjunto de testes obtido da separação de 1/3 das instâncias rotuladas de cada segmento do *Corpus*. Esta separação se fez necessária pela necessidade de verificar a eficiência do modelo proposto, observando os valores que foram rotulados manualmente, sendo estes a referência. Os testes estimam a capacidade de classificação em novas instâncias e demonstram se o esquema de classificação proposto tem melhor desempenho que o classificador usado como base.

Para fins de comparação, os testes foram realizados com os três esquemas de classificação: apenas o classificador básico (Naive Bayes), apenas o contador de palavras-chave (usando as palavras-chave da *background knowledge*, obtidas através do modelo de tópicos), e o classificador proposto combinando ambos os mecanismos. Os resultados dos testes (F1-*score* de cada classificador para cada segmento do *Corpus*) são mostrados na Figura 9.

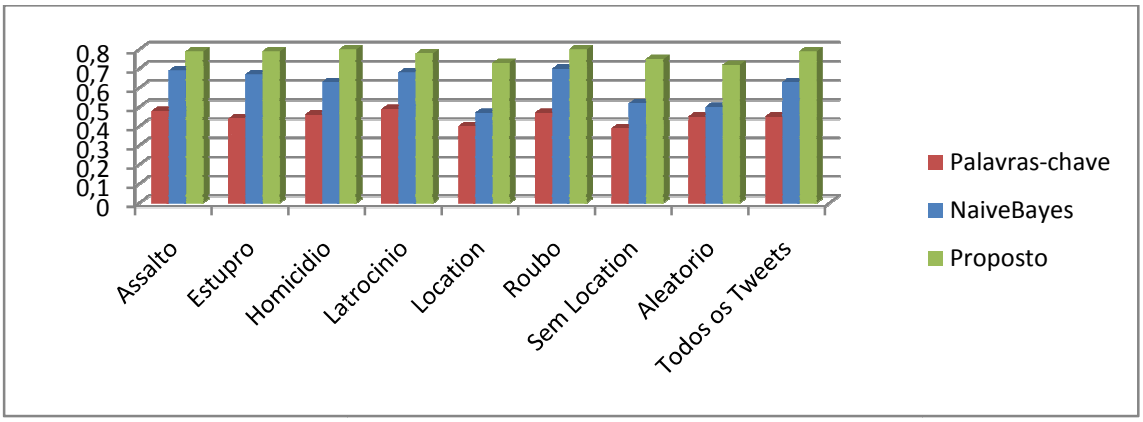

Figura 9 - Resultado dos Experimentos realizados (F1 Score)
Fonte: Próprio Autor

Em todos os segmentos, é possível verificar uma melhoria no desempenho de classificação do classificador proposto com relação ao classificador de base e a um contador de palavras-chave simples. Isso indica que o uso de palavras-chave obtidas através de um modelo de tópicos para auxiliar na classificação de tópicos contribui positivamente para o desempenho de classificação, ao mesmo tempo em que o uso apenas das palavras-chave obtém o pior desempenho. Como visto em esquemas de classificação usando *ensembles* (conjuntos de classificadores combinados), a combinação entre o classificador NaiveBayes e o contador de palavras aproveita os pontos fortes dos dois módulos isolados, compensando as falhas de cada um.

6 Conclusão

O trabalho descrito no presente documento propõe a melhoria na classificação de tópicos em textos curtos usando *background knowledge,* com objetivo de identificar *tweets* escritos em português do Brasil relacionados com a violência.

Para realização do trabalho, foram realizadas pesquisas acerca de classificação de textos curtos, visando aquisição de bom embasamento teórico, fundamental para implementação da melhoria proposta.

Em um segundo momento foi pesquisado os trabalhos correlatos, e identificado que boa parte das estratégias fazem uso da *background knolwedge*, justificando assim o uso da referida base neste trabalho. A base utilizada nesta dissertação foi construída tendo como referência uma modelagem de tópicos utilizando a ferramenta Mallet.

Concluída a etapa de definição de como a *background knowledge* seria composta, iniciou-se a implementação de um programa em Java capaz de verificar a quantidade de termos da *background knowledge* presente em cada *tweet*, bem como fazer uso de um algoritmo de classificação de textos convencionais implementado pelo Weka, na ocasião NaiveBayes. Além destas duas tarefas, o referido programa combinou os dois resultados com a finalidade de estabelecer o score final de classificação para cada *tweet*, indicando se está relacionado com a violência ou não.

Uma vez finalizado a implementação do programa responsável pela classificação de textos curtos, foi necessário realizar um estudo sobre construção de *Corpus* antes da fase de realização de testes, afinal, não é possível realizar os testes e experimentos sem o a existência do *Corpus*.

Após estudo, foi construído um *Corpus* com um total de 2.909.391 (Dois milhões novecentos e nove mil trezentos e noventa e um *tweets*) extraídos por um programa feito em Java, respeitando todas as etapas de construção de um *Corpus* mencionadas no Capítulo 4, deste total 18967 (Dezoito Mil Novecentos e sessenta e sete) *tweets* foram rotulados manualmente. A extração destes *tweets* foi segmentada em nove partes, sendo cinco segmentos relacionados com palavra-chave (Assalto, Estupro, Homicídio, Latrocínio e Roubo), um considerando *tweets* com e outro sem geolocalização, outro segmento com *tweets* totalmente aleatórios, e por ultimo um segmento contendo todos os *tweets* por palavra-chave.

Finalmente, tendo o *Corpus* construído e o programa com a melhoria proposta devidamente implementada, os testes foram iniciados com a construção da *background*

knowledge, sendo esta composta por dezoito termos obtidos a partir da execução do Mallet. Em seguida, 1/3 de cada segmento do *Corpus* foi submetido ao programa, com a finalidade de verificar o desempenho da melhoria proposta, após execução pode-se observar que para todos os nove segmentos a melhoria proposta se demonstrou mais eficiente, se comparado ao NaiveBayes puro, e ao contador de palavras.

Sendo assim, é possível concluir que esses resultados validam a melhora no esquema de classificação de texto curto proposto neste trabalho, incluindo o uso de modelagem de tópicos para criar uma *background knowledge* e a associação de um contador de palavras simples, baseado na *background knowledge*, com um classificador de base para obter uma melhoria na classificação de tópicos quando o tópico desejado acontece raramente em textos reais.

6.1 Sugestões para Trabalhos Futuros

Nesta dissertação foram demonstrados os resultados alcançados na aplicação da melhoria proposta a classificação de textos curtos, na ocasião, *tweets* enviados por usuários. Utilizou-se para construção da *background knowledge* um *Corpus* criado relacionado a violência, em seguida, foi feita a classificação "Contador de Palavras" com base na *background* modelada.

Por fim, foi utilizado como classificador convencional, Naive Bayes. A combinação das duas estratégias de classificação aliadas a parametrização do classificador proposto, culminou em uma nova classificação para o *Tweet*.

Entretanto, perceberam-se alguns aspectos que podem ser trabalhos em pesquisas futuras:

- Realização da validação do *Corpus* – Conforme mencionado ao longo do texto, o *Corpus* criado foi classificado manualmente por uma só pessoa, podendo esta ficar viciada e não classificar adequadamente os dados, impactando assim no modelo proposto como um todo;
- Automatização da Construção da *Background Knowledge* – Nesta dissertação a *background knowledge* é construída de forma semi-automática. Sendo assim passiva de falhas humanas, ou mesmo, escolha de tópicos tendenciosos;
- Realização de novos experimentos – O algoritmo "NaiveBayes" utilizado na classificação final pode ser substituído por outro algoritmo de classificação. Os

parâmetros determinados no ato da combinação dos resultados podem ser ajustados conforme a necessidade.

Referências Bibliográficas

ALUÍSIO, S.; ALMEIDA, G. **O que é e como se constrói um *Corpus*? Lições aprendidas na compilação de vários *corpora* para pesquisa lingüística.** Calidoscópio Vol. 4, n. 3, Unisinos, Rio Grande do Sul, Brasil, 2006.

ALENCAR, A; **Visualização da Evolução Temporal de Coleções de Artigos Científicos,** Tese (Doutorado em Ciências da Computação e Matemática Computacional) – – Instituto de Ciências Matemáticas e Computação USP. Brasil, 2013.

BATISTA, G. **Pré-processamento de Dados em Aprendizado de Máquina Supervisionado,** Tese (Doutorado em Ciências da Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e Computação USP. Brasil, 2003.

BLEI, D.; ANDREW, N.; JORDAN M. **Latent Dirichlet allocation**. Journal of Machine Learning Research, January 2003.

BLEI, D. **Probabilistic topic models.** *Communications of the ACM*, 55(4):77–84, 2012

BOYD, D.; SCOTT, G.; GILAD, L. **Tweet, Tweet, Re*tweet*: Conversational Aspects of Re*tweet*ing on Twitter**. HICSS-43. IEEE: Kauai, HI, January 6. 2010.

BOHN, VANESSA. "**Comunidades de Prática na Formação Docente: Aprendendo a usar ferramentas na Web 2.0**" Dissertação (Mestrado em estudos Linguísticos) – Universidade Federal de Minas Gerais, 2010, 158 p.

BREVE, F. **Aprendizado de Máquina em Redes Complexas.** Tese (Doutorado em Ciências da Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e Computação da USP, BRAZIL, 2010.

CHENG, Z.; CAVERLEE, J.; LEE, K. **You Are Where You Tweet: A Content-Based Approach toGeo-locating Twitter Users.** CIKM'10, October 26–30, 2010, Toronto, Ontario, Canada.

CONDUTA, B.; MAGRIN, D. **Aprendizagem de Maquina**, Dissertação de Mestrado da Faculdade de Tecnologia da Universidade Estadual de Campinas,Campinas, BRASIL, 2010.

COSTA, A. "E-Branding – **A internet como instrumento de fixação de marca: Tecnisa no Twitter**" Trabalho de conclusão do Curso de Pós-Graduação – Estácio UniRadial (MBA em Gestão Empresarial). São Paulo, 2011, 74p.

CRISTIANINI, N.; SHAWE-TAYLOR, J.**An Introduction to Support Vector Machinesand Other Kernel-based Learning Methods**.Cambridge University Press, 2000.

FREITAS, W.W.L.; SILVA, V.F.; MORAES, R. M. **Análise Espacial de uma década de mortes violentas por homicídios na cidade de João Pessoa**. XI Safety, Health and Environment World Congress. Santos, BRAZIL, 2011.

GUO, W.; LI, H.; JI, H.; DIAB, M. **Linking *Tweets* to News: A Framework to Enrich Short Text Data in Social Media**. The 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013) Sofia, Bulgaria, 2013.

KAWAK, H.; LEE, C.; PARK, H.; MOON, S. **What is Twitter, a Social Network or a News Media?.**WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA**.**

MANUEL, T. **Amostragem e Caracterização de Coleções de Dados do Twitter**. Dissertação (Mestrado Integrado em Engenharia Informática e Computação) – Curso de Pós-Graduação em Engenharia Informática e Computação, Faculdade de Engenharia da Universidade do Porto. Portugal, 2013, 58p.

MARTINS, C. **Uma Abordagem Para Pré-processamento de Dados Textuais em Algoritmos de Aprendizado,** Tese (Doutorado em Ciências da Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e Computação USP. Brasil, 2003.

MCGUINNESS, C.. **How to retrieve more than 100 *tweets* with the Twitter API and Twitter4J**. Novembro de 2013. Disponível em <http://www.socialseer.com/twitter-programming-in-java-with-twitter4j/how-to-retrieve-more-than-100-*tweets*-with-the-twitter-api-and-twitter4j/> Acesso em 12 de Junho de 2014

MORSTATTER, F.; PFEFFER, J.; LIU, H.; CARLEY, K. .**Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's *Firehose***, Seventh International AAAI Conference on Weblogs and Social Media, 2013, Cambridge, Massachusetts, 400-408.

NASCIMENTO, S.; BEUREN, I. **Redes Sociais na Produção Científica dos Programas de Pós-Graduação de Ciências Contábeis no Brasil.** V.15, art.3 RAC, Curitiba, Brasil, 2011.

PHAN, X.; NGUYEN, L.; HORIGUCHI, S. **Learning to Classify Short and Sparse Text & Web with Hidden Topics from Large-scale Data Collections.**WWW 2008 / Refereed Track: Data Mining – Learning, Beijing. China, 2008.

PEREZ, S. **Only 66% Use Twitter Profile Location Field as Intended, Says PARC Research Study, 2011**. Disponível em <http://readwrite.com/2011/01/18/66_percent_use_twitter_profile_location_field_says_parc_research_study#awesm=~oA8pD4STy7rsjq>. Acesso em 01/04/2014.

PRIER, K.; SMITH, M.; CARRIER, C.; HANSON, C. **Identifying Health-Related Topics on Twitter**. Social Computing, Behavioral-Cultural Modelingand Prediction Lecture Notes in Computer Science Volume 6589, 2011.

RECUERO, R. **"Redes Sociais na Internet".** PortoAlegre**:** Meridional LTDA**,** 2009**,**192p.

RECUERO, R.; ZAGO, G. **Em busca das "redes que importam": redes sociais e capital social no Twitter.** XVIII Encontro do Compós. Líbero – São Paulo – v. 12, n. 24, p. 81-94, dez. de 2009**.**

ROMANO, O. **Brasil vira 'potência' das redes sociais em 2013.** Disponível em <http://blogs.estadao.com.br/link/em-2013-brasil-vira-potencia-das-redes-sociais/> Acesso em 01/04/2014.

SANKARANARAYANAN, J.; SAMET, H.; TEITLER, B. E.; LIEBERMAN, and M. D.; SPERLING, J. **TwitterStand: news in *tweets***. In Proc. ACM GIS'09, Seattle, Washington, Nov. 2009.

SANTOS, A. **Descobrindo Eventos Locais Utilizando Análise de Séries Temporais nos dados do Twitter.** Dissertação (Mestrado em Ciência da Computação) – Programa de Pós-Graduação em Computação, Universidade Federal do Rio Grande do Sul. Brasil, 2013, 72p.

SÁ, Hially. **Seleção de Características para Classificação de Textos**. Trabalho de Conclusão de Graduação – Centro de Informática UFPE. Brasil, 2008.

SECOM, **Hábitos de Consumo de Mídia Pela População Brasileira**. Dezembro de 2015 Disponível em < http://www.secom.gov.br/atuacao/pesquisa/lista-de-pesquisas-quantitativas-e-qualitativas-de-contratos-atuais/pesquisa-brasileira-de-midia-pbm-2015.pdf >. Acesso em 02 de Outubro de 2015

SHAINDLIN, A. **Alumni Organization Twitter Profiles: A QuickGuide to 10 Best Pratices.** Dezembro de 2010. Disponível em http://www.alumnifutures.com/whitepapers.html. Acesso em 30 de Março de 2014.

SOUSA, G. **Tweetmining: Análise De Opinião Contida Em Textos Extraídos Do Twitter** (Graduação em Sistemas da Informação). Universidade Federal de Lavras. Minas gerais, Brasil, 2012.

SRIRAM, B. **Short Text Classification In Twitter To Improve Information Filtering.** Dissertação (Mestrado na escola de graduação de Ciências) **-** Ohio StateUniversity. USA ,2010.

STEYVERS, M.; GRIFFITHS, T.**Probabilistic topic models**. Latent semantic analysis: A road to meaning. Mahwah, NJ: Erlbaum, 2006.

TWITTER4J. **Integracao de aplicacoes Java com o Twitteratraves do Twitter4j. Disponível em: <**http://twitter4j.org>. Acessado: em dezembro de 2013.

TWITTER. **Portal do Twitter para Desenvolvedor**, Disponível em <https://dev.twitter.com/start>, Acessado em dezembro de 2013.

Twitter, **Portal de informações sobre o Twitter**. Disponível em: < http:// about.twitter.com/pt/company>. Acesso em 07 de Maio de 2015.

WAISELFISZ, J. J. **Mapa da Violência dos Municípios Brasileiros 2008.** Ministério da Saúde 1ª Edição**.** Brasília, BRASIL, 2008.

ZELIKOVITZ, S.; HIRSH, H. **Improving Short-TextClassificationUsingUnlabeled Background KnowledgetoAssessDocumentSimilarity.**Computer Science Department, Rutgers University, 110 Frelinghuysen Road, Piscataway, NJ 08854-8019 USA, 2002.

WAISELFISZ, J. J.. **Mapa da Violência dos Municípios Brasileiros 2008.**Ministério da Saúde 1ª Edição**.** Brasília, BRASIL, 2008.

WAISELFISZ, J.J. **Mapa da Violência do Brasil 2014.** Secretaria Geral da Presidência da República, 1a Edicao, Brasilia, BRASIL, 2014.