



UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
ESCOLA DE INFORMÁTICA APLICADA

CATEGORIZAÇÃO SEMIAUTOMÁTICA DE TEXTOS APLICADA EM SISTEMA
WEB DE DIVULGAÇÃO DE NOTÍCIAS

GABRIEL RAMALHO DE ALBUQUERQUE

Orientadora

Dr^a. GEIZA MARIA HAMAZAKI DA SILVA

RIO DE JANEIRO, RJ – BRASIL

JULHO DE 2016

CATEGORIZAÇÃO SEMIAUTOMÁTICA DE TEXTOS APLICADA EM SISTEMA
WEB DE DIVULGAÇÃO DE NOTÍCIAS

GABRIEL RAMALHO DE ALBUQUERQUE

Projeto de Graduação apresentado à Escola de
Informática Aplicada da Universidade Federal do
Estado do Rio de Janeiro (UNIRIO) para obtenção do
título de Bacharel em Sistemas de Informação.

Aprovada por:

Dr^a. GEIZA MARIA HAMAZAKI DA SILVA (UNIRIO)

Dr^a. FERNANDA ARAUJO BAIÃO AMORIM (UNIRIO)

Dr^a. SIMONE BACELLAR LEAL FERREIRA (UNIRIO)

RIO DE JANEIRO, RJ – BRASIL

JULHO DE 2016

Agradecimentos

Primeiramente, agradeço a minha família, aos meus pais, pois sem eles não teria sido possível chegar até esse momento, obrigado por todo o esforço para que eu pudesse ter uma educação digna, pelos importantes valores que me passaram e pelo amor e carinho.

Agradeço aos meus amigos pelo carinho e pelo apoio. Aos amigos que fiz durante o curso e que levo para o restante da minha vida, em especial aos amigos Cecília, Jefferson, João Felipe, Pedro, Rodrigo e Davi, todos de alguma forma contribuíram nesse processo.

Agradeço especialmente a Bruna, que me trouxe paz e tranquilidade em momentos difíceis, pelos seus conselhos e também por toda a ajuda na revisão de partes deste trabalho.

Agradeço a todos os colegas que tive o prazer de trabalhar durante a minha carreira profissional, obrigado pelos conhecimentos compartilhados.

Agradeço a todos os professores da UNIRIO por todos os ensinamentos e aos servidores pelo excelente serviço prestado aos alunos.

Agradeço a professora Geiza por ter me orientado e também pela dedicação incrível e pela atenção que tem não só comigo, mas com todos os alunos. Ao professor Pimentel pelas ideias e sugestões dadas durante as aulas de Web Social. A professora Fernanda e a professora Simone por terem aceitado o convite para compor a banca examinadora.

A todos que participaram, direta ou indiretamente, do meu processo de formação, o meu muito obrigado.

RESUMO

O aprimoramento das Tecnologias de Informação e Comunicação, o surgimento da *internet* e das redes sociais, dos *blogs* e dos *microblogs* provocaram inúmeras transformações na sociedade, na forma como as pessoas compartilham informações. Todos os dias são postadas inúmeras notícias nestes sistemas, porém estes não foram projetados para funcionarem como *sites* jornalísticos. Neste contexto foi desenvolvido um protótipo de um sistema divulgação de notícias publicadas pelos próprios usuários, organizando os artigos em categorias, estados e cidades, tornando o compartilhamento de informações em algo mais público, democrático e organizado.

Para reduzir a tarefa do usuário na entrada de dados, foi aplicado um mecanismo de categorização semiautomática das notícias utilizando o algoritmo Naive Bayes Multinomial, demonstrando uma aplicação prática de técnicas de aprendizado de máquina e mineração de dados. Desta forma, o sistema proposto fornece aos usuários uma plataforma aberta de compartilhamento de informações além de auxiliá-los no processo de categorização das notícias.

Palavras-chave: KDD, mineração de dados, categorização de texto, Naive Bayes, web social.

ABSTRACT

The improvement of the Information and Communication Technology, the development of the internet, the social networks, blogs and microblogs caused many changes in the society and the way people share information. Every day are posted many news in these systems however, they do not have news websites features. In this context was developed a prototype of a news website which is open to the users and the news are organized in categories, states and cities, making the sharing of information into something more public, democratic and organized.

To minimize the data input task, a semi-automatic classification mechanism was applied, using the Multinomial Naive Bayes algorithm, demonstrating a practical application of machine learning and data mining techniques. Thus, the proposed system provides users an open sharing information system as well as assist them in the news classification process.

Keywords: KDD, data mining, text classification, Naive Bayes, web social.

Índice

| | | |
|---------|--|----|
| 1 | Introdução..... | 11 |
| 1.1 | Motivação | 11 |
| 1.2 | Objetivos | 12 |
| 1.3 | Organização do Texto | 12 |
| 2 | Fundamentação Teórica | 14 |
| 2.1 | Descoberta de Conhecimento em Banco de Dados | 14 |
| 2.1.1 | Seleção de Dados..... | 16 |
| 2.1.2 | Pré-processamento..... | 16 |
| 2.1.3 | Transformação dos Dados | 17 |
| 2.1.4 | Mineração de Dados | 17 |
| 2.1.4.1 | Categorizador Naive Bayes | 19 |
| 2.1.5 | Interpretação e Avaliação dos Resultados | 22 |
| 2.2 | Projetando Aplicações para a Web Social..... | 23 |
| 2.2.1 | Layout..... | 23 |
| 2.2.2 | Objeto Social | 23 |
| 2.2.3 | Funcionalidades | 24 |
| 2.2.3.1 | Cadastro | 24 |
| 2.2.3.2 | Login/Logout | 24 |
| 2.2.3.3 | Comentários | 24 |
| 2.2.3.4 | Classificação (“Rating”) de Publicações | 24 |
| 3 | Estado da Arte | 26 |
| 3.1 | Europe Media Monitor News Explorer | 26 |
| 3.2 | News Explorer | 27 |
| 3.3 | Cora..... | 27 |
| 3.4 | News Explorer for Web Streaming Data | 28 |
| 4 | Categorizador Semiautomático | 30 |

| | |
|---|----|
| 4.1 Weka..... | 30 |
| 4.2 Base de Dados e Seleção | 31 |
| 4.3 Pré-processamento | 33 |
| 4.4 Treinamento e Categorização | 34 |
| 4.5 Avaliando os Resultados..... | 34 |
| 4.6 News Share e Aplicação no Mundo Real | 37 |
| 4.6.1 Ferramentas Utilizadas | 37 |
| 4.6.1.1 HTML 5 | 37 |
| 4.6.1.2 Syntactically Awesome Style Sheets | 37 |
| 4.6.1.4 Bootstrap | 38 |
| 4.6.1.5 Ruby on Rails | 38 |
| 4.6.2 Layout..... | 39 |
| 4.6 Objeto Social | 42 |
| 4.6.3 Funcionalidades | 42 |
| 4.6.3.1 Publicação e Categorização de Notícias | 43 |
| 5 Conclusões e Trabalhos Futuros | 47 |
| Referências Bibliográficas | 50 |
| APÊNDICE A – Modelo Conceitual do Banco de Dados..... | 54 |
| APÊNDICE B – Modelo Lógico do Banco de Dados | 55 |
| APÊNDICE C – Diagrama de Casos de Uso..... | 56 |
| APÊNDICE D – Descrição dos Casos de Uso | 57 |

Índice de Tabelas

| | |
|--|----|
| Tabela 1 - Conjunto de documentos problemáticos para o Naive Bayes Multinomial. Fonte: Manning et al. (2009). | 20 |
| Tabela 2 - Exemplo de uma aplicação do Naive Bayes Multinomial..... | 21 |
| Tabela 3 - Matriz de confusão do J48..... | 35 |
| Tabela 4 - Matriz de confusão do PART. | 35 |
| Tabela 5 - Matriz de confusão do Naive Bayes. | 36 |
| Tabela 6 - Matriz de confusão do Naive Bayes Multinomial. | 36 |
| Tabela 7 - Descrição do caso de uso "Realizar Cadastro". | 57 |
| Tabela 8 - Descrição do caso de uso "Atualizar Dados". | 58 |
| Tabela 9 - Descrição do caso de uso "Realizar Login"..... | 59 |
| Tabela 10 - Descrição do caso de uso "Realizar Logout"..... | 60 |
| Tabela 11 - Descrição do caso de uso "Publicar Notícia". | 61 |
| Tabela 12 - Descrição do caso de uso "Comentar Notícia". | 62 |
| Tabela 13 - Descrição do caso de uso "Aprovar Notícia". | 63 |
| Tabela 14 - Descrição do caso de uso "Desaprovar Notícia". | 64 |

Índice de Figuras

| | |
|---|----|
| Figura 1 - Etapas do processo de KDD. Fonte: Steiner et al. (2006) apud Fayyad et al. (1996b)..... | 15 |
| Figura 2 - Procedimento geral de construção do modelo de categorização. Fonte: Raschka (2014). | 18 |
| Figura 3 - Tipos de <i>layouts</i> mais comuns. Fonte: Bell (2009), p. 244..... | 23 |
| Figura 4 - Tela inicial do <i>site</i> EMM News Explorer. Fonte: <i>print screen</i> do <i>site</i> EMM News Explorer. | 26 |
| Figura 5 - Tela de busca de artigos do News Explorer. Fonte: Desai et al. (2014). | 27 |
| Figura 6 - Página inicial do <i>site</i> Cora. Fonte: McCallum et al. (2000)..... | 28 |
| Figura 7 - Interface do NEWSD. Fonte: Mohiuddin et al. (2015)..... | 29 |
| Figura 8 - Tela inicial do Weka. | 31 |
| Figura 9 - Estrutura do arquivo <i>.arff</i> contendo a base de treinamento. | 32 |
| Figura 10 - Quantidade de instancias, atributos e a quantidade de instancias por atributo. | 32 |
| Figura 11 - Tipos de filtros disponíveis no Weka..... | 33 |
| Figura 12 - Funcionamento do “ <i>Cross-validation</i> ”. Fonte: Refaeilzadeh et al. (2008).. | 34 |
| Figura 13 - Página inicial do sistema com algumas notícias publicada, retiradas do <i>site</i> Globo.com. | 39 |
| Figura 14 – <i>Layout</i> versão <i>mobile</i> | 40 |
| Figura 15 - <i>Site</i> Globo.com. Fonte: <i>print screen</i> do <i>site</i> Globo.com | 41 |
| Figura 16 - Capa do jornal The New York Times. Fonte: Diário do Rio..... | 41 |
| Figura 17 - Exemplo de uma notícia publicada, retirada do <i>site</i> Globo.com. | 42 |
| Figura 18 - Exemplo de uma notícia publicada, retirada do <i>site</i> Globo.com. | 43 |
| Figura 19 - Diagrama de sequência do processo de categorização e publicação das notícias..... | 44 |
| Figura 20 - Formulário de cadastro de notícia com os campos preenchidos. | 45 |
| Figura 21 - Formulário com as opções de cidades e categorias. | 46 |
| Figura 22 - Modelo conceitual do banco de dados. | 54 |
| Figura 23 - Modelo lógico do banco de dados. | 55 |
| Figura 24 - Diagrama de casos de uso. | 56 |

Lista de Abreviaturas

TICs – Tecnologias de Informação e Comunicação

KDD – Knowledge Discovery in Databases

SGBD – Sistemas Gerenciadores de Bancos de Dados

SVM – Support Vector Machine

NBM – Naive Bayes Multinomial

MLP – Multi-Layer Perceptron

SOM – Self-Organizing Maps

NEWSD – News Explorer for Web Streaming Data

GUI – Graphical User Interface

API – Application Programming Interface

GNU – General Public License

HTML – HyperText Markup Language

Sass – Syntactically Awesome Style Sheets

CSS – Cascading Style Sheets

MVC – Model-View-Controller

1 Introdução

1.1 Motivação

Poucas inovações tecnológicas provocaram tantas mudanças em tão pouco tempo na sociedade como as novas Tecnologias de Informação e Comunicação (TICs) (Barbosa et al., 2004). Afirmar que as TICs modificam a vida da sociedade tornando-a mais dinâmica não é exagero; essas facilitam uma boa parte do trabalho cotidiano, contribuindo e agilizando processos que levariam semanas e até meses para serem solucionados (Carvalho, 2011).

O surgimento da *internet* e posteriormente as redes sociais fez com que a sociedade conhecesse os novos modelos e conceitos de interatividade entre usuários; onde os mesmos estão conectados por um ou vários tipos de relações, compartilhando valores e objetivos comuns, trocando informações a respeito de tudo e de todos (Carvalho, 2011).

Essa profunda transformação na sociedade contemporânea é denominada de Revolução da *Internet*, também conhecida por outros nomes como Revolução Digital ou Revolução Informacional. A iteração é contínua, ao longo de vários momentos do dia, seja pelo celular ou outro computador móvel; um evento interessante é motivo para *tuitar* e tornar a notícia compartilhada com todos (Pimentel e Fuks, 2011).

É possível perceber que existe o interesse da sociedade em compartilhar informações e as redes sociais, os *blogs*, os *microblogs* e os aplicativos de troca de mensagens acabaram recebendo essa função. Porém, esses sistemas não possuem funcionalidades típicas de *sites* de divulgação de notícias, como a categorização e exibição de notícias por regiões, etc. Já os *sites* jornalísticos não são abertos aos usuários, somente os jornalistas podem publicar os artigos.

Alguns sistemas de divulgação de notícias foram analisados (ver capítulo 3) e foi possível observar que estes não apresentam a funcionalidade de inserção de notícias pelos usuários, todos os dados exibidos são extraídos de fontes externas.

Dado este cenário é proposto um protótipo de um sistema de publicação e compartilhamento de informações, denominado *News Share*, que possui a funcionalidade de ser aberto aos usuários, como o que acontece nas redes sociais, mas ao mesmo tempo voltado especificamente para o compartilhamento de notícias, com funcionalidades típicas de *sites* jornalísticos.

1.2 Objetivos

O objetivo do trabalho é o desenvolvimento de um protótipo de um sistema, denominado *News Share*, para disponibilizar notícias fornecidas pelos usuários, no qual uma notícia publicada poderá ser acessada por todos os visitantes. Este sistema possuirá a funcionalidade de agrupamento de notícias por estados, cidades e categorias.

Com o intuito de demonstrar uma aplicação prática de técnicas de aprendizado de máquina e de mineração de dados, o sistema desenvolvido realiza parte da tarefa de categorização, o que antes seria feito pelo usuário.

1.3 Organização do Texto

O presente trabalho será desenvolvido da seguinte forma:

- Capítulo II: Fundamentação Teórica – São apresentados conceitos de descoberta de conhecimento em base de dados, técnicas de mineração de dados, aprendizado de máquina e os princípios, padrões e práticas utilizados para projetar aplicações *web* sociais. Esses conceitos são necessários para que o leitor possa compreender a solução que será proposta.
- Capítulo III: Estado da Arte – Nesse capítulo, são analisados alguns *sites* que utilizam a categorização automática de documentos como uma de suas funcionalidades.
- Capítulo IV: Categorizador Semiautomático – É apresentada a solução proposta para o emprego do categorizador semiautomático usado pelo sistema como um exemplo de aplicação real das técnicas abordadas no capítulo II, assim como as ferramentas utilizadas e funcionalidades do sistema.

- Capítulo V: Conclusões e Trabalhos Futuros – Reúne as considerações finais, assinalando as contribuições da pesquisa e sugerindo possibilidades de aprofundamento posterior.

2 Fundamentação Teórica

Para o desenvolvimento do categorizador semiautomático, foram utilizados conceitos em descoberta de conhecimento em banco de dados, técnicas de mineração de dados e aprendizado de máquina. Já para o desenvolvimento das interfaces e das funcionalidades do protótipo do sistema, foram seguidos alguns princípios e padrões. Neste capítulo serão apresentados os fundamentos para que o leitor possa compreender os passos da solução que será proposta.

2.1 Descoberta de Conhecimento em Banco de Dados

Com a popularização do computador e da internet, a capacidade da sociedade em gerar e coletar dados tem crescido rapidamente. Governos, agências, instituições científicas e empresas dedicam enormes recursos para coletar e armazenar dados (Kantardzic, 2011). Entretanto apenas uma pequena parcela desses dados será usada uma vez que, em muitos dos casos, o volume é muito grande para ser trabalhado ou a estrutura dos dados é complexa para ser eficientemente analisada. Em geral, não existia o desenvolvimento de um plano para que os dados fossem armazenados de forma eficiente e utilizados para análises futuras (Kantardzic, 2011).

Essa visão tem mudado uma vez que as instituições tem percebido que podem extrair conhecimento dessas bases e que este pode ser valioso para a tomada de decisão. Em paralelo, várias técnicas e ferramentas tem sido projetadas para dar suporte a extração de conhecimento a partir dos grandes e crescentes volumes de dados. Estas técnicas e ferramentas são os assuntos abordados pelo campo emergente da Descoberta de Conhecimento em Bancos de Dados (ou KDD do inglês *Knowledge Discovery in Databases*) (Fayyad et al., 1996a).

Existe uma diferença entre KDD e mineração dos dados (*data mining*). KDD refere-se a todo processo de descoberta de conhecimento útil nos dados, enquanto *data*

mining refere-se à aplicação de um ou vários algoritmos para extrair modelos (ou padrões¹) dos dados; até 1995, muitos autores consideravam os termos KDD e *data mining* como sinônimos (Oliveira, 2010). *Data mining* é uma etapa do processo de KDD (ver figura 1).

O KDD é um processo iterativo e iterativo, envolvendo várias etapas e decisões tomadas pelo usuário (Fayyad et al., 1996a; Fayyad et al., 1996b). Interativo, pois o usuário pode (e muitas vezes necessita) continuamente intervir e controlar o curso das atividades; iterativo, por ser uma sequência finita de operações em que o resultado de cada uma é dependente dos resultados das que a precedem (Prass, 2012).

As etapas definidas para este processo são: seleção, pré-processamento, transformação, *data mining* e interpretação dos resultados (ver figura 1).

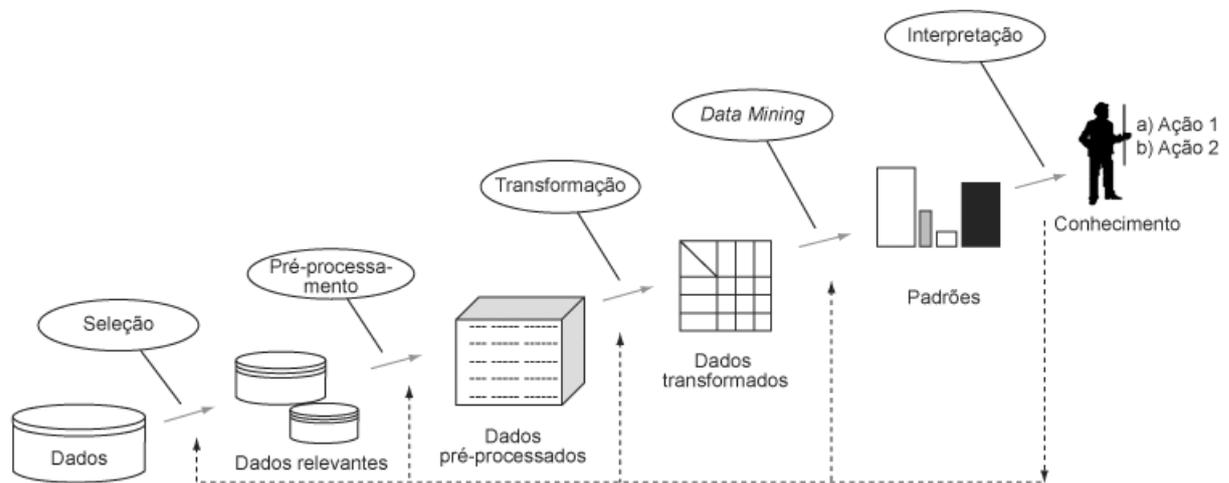


Figura 1 - Etapas do processo de KDD. Fonte: Steiner et al. (2006) apud Fayyad et al. (1996b).

O termo processo é utilizado para o KDD pois existem muitos passos envolvidos na preparação dos dados, procura por padrões, avaliação do conhecimento e refinamento — todos repetidos em múltiplas iterações (ver figura 1). Estes passos não são triviais, ou seja, envolvem a busca por estruturas, modelos, padrões ou parâmetros (Fayyad et al., 1996a). O padrão a ser encontrado não é um cálculo simples de quantidades pré-definidas, como calcular o valor médio de um conjunto de números.

¹ Um padrão é uma descrição de um subconjunto dos dados ou de um modelo aplicável ao subconjunto; assim, extrair um padrão também designa fazer qualquer descrição de alto nível de um conjunto de dados (Fayyad et al., 1996b).

Os padrões descobertos devem ser válidos e com algum grau de certeza; também devem ser novos (pelo menos para o sistema, e preferencialmente para o usuário) e potencialmente úteis para a tarefa ou o usuário. Finalmente, os padrões devem ser compreensíveis — se não imediatamente então após algum pós-processamento (Fayyad et al., 1996b).

Nas subseções seguintes são definidas as etapas do processo de KDD.

2.1.1 Seleção de Dados

A fase de seleção de dados é a primeira no processo de descobrimento de informação e possui impacto significativo sobre a qualidade do resultado final, uma vez que nesta fase é escolhido o conjunto de dados contendo todas as possíveis variáveis (também chamadas de características ou atributos) e registros (também chamados de casos ou observações) que farão parte da análise (Prass, 2012). Geralmente essa escolha é feita por um especialista do domínio.

Os dados podem estar disponíveis em diferentes formatos e estruturas e podem ser obtidos através de planilhas de texto, de um sistema legado², de um armazém de dados³ (*Data Warehouse*), etc.

2.1.2 Pré-processamento

Neste ponto, os dados redundantes e/ou inconsistente, que possuem valores atípicos, chamados de *outliers*, deverão ser avaliados para que possam ser excluídos ou não.

Na categorização de documentos, deve ser realizada a análise léxica, na qual o documento será adaptado. Essencialmente, eliminam-se os dígitos e os sinais de pontuação e isolam-se os termos e efetua-se a conversão de letras maiúsculas para minúsculas (Camargo, 2007). Palavras que não possuem valor semântico, denominadas “*stopwords*” (conjunções, preposições, artigos, conectores textuais e léxicos, etc.), também são eliminadas.

² Sistema legado é um sistema de software que, embora seja antigo, permanece vital para uma organização.

³ Armazém de dados é um repositório de informações coletadas de várias fontes, armazenadas sob um esquema unificado e, geralmente, residente em um único local; são construídos através de um processo de limpeza, integração, transformação e carregamento de dados e de forma periódica (Han et al., 2011).

Numerais e sinais gráficos como o apóstrofo, parênteses, colchetes, etc., também devem ser removidos uma vez que não acrescentam valor para o conjunto de treinamento no caso da categorização de documentos.

Por exemplo, a frase “Esse equilíbrio era tido como condição para o sucesso do plano econômico.” após o pré-processamento ficará “equilibrio era tido como condicao sucesso plano economico”.

Uma representação utilizada na categorização de textos é a formatação do documento em um vetor de palavras ou *strings*. Esse modelo é chamado de “*bag of words*”, que é um modelo simples onde o vocabulário é formado pelas diferentes palavras que ocorrem na base de teste. A ordem de ocorrência dessas palavras, a semântica e a sintaxe não são consideradas.

2.1.3 Transformação dos Dados

Nesta fase os dados são transformados em estruturas que possam ser interpretadas pelas ferramentas e algoritmos de mineração de dados.

Em grandes corporações é comum encontrar computadores executando diferentes sistemas operacionais e diferentes Sistemas Gerenciadores de Bancos de Dados (SGDB); estes dados que estão dispersos devem ser agrupados em um repositório único (Prass, 2012).

Além disto nesta fase, se necessário, é possível obter dados faltantes através da transformação ou combinação de outros, são os chamados “dados derivados”. Um exemplo de um dado que pode ser calculado a partir de outro é a idade de um indivíduo, que pode ser encontrada a partir de sua data de nascimento. Outro exemplo é o valor total de um financiamento que pode ser calculado a partir da multiplicação do número de parcelas pelo valor da parcela (Prass, 2012).

2.1.4 Mineração de Dados

Mineração de dados consiste na exploração e análise de grandes bases de dados com objetivo de descobrir padrões e regras (Linoff e Berry, 2011) anteriormente desconhecidos utilizando um algoritmo de aprendizado de máquina.

Os dois principais objetivos da mineração de dados tendem a ser a predição e a descrição. Predição envolve o uso de variáveis presentes nos dados para predizer valores futuros ou desconhecidos; e a descrição, por sua vez, tem como objetivo

encontrar padrões que descrevam os dados e que possam ser interpretados por humanos (Fayyad et al., 1996a; Fayyad et al., 1996b; Kantardzic, 2011).

Os objetivos da predição e da descrição são alcançados utilizando as seguintes técnicas: regressão, clusterização, sumarização, modelo de dependência, detecção de desvio e categorização (Fayyad et al., 1996a; Fayyad et al., 1996b; Kantardzic, 2011).

Regressão é a descoberta de uma função que mapeia um item a uma variável de previsão. Clusterização é uma tarefa descritiva em que se procura identificar um conjunto finito de categorias ou grupos para descrever um conjunto de dados. Sumarização é uma tarefa descritiva que envolve métodos para encontrar uma descrição compacta para um conjunto (ou subconjunto) de dados. Modelo de dependência consiste em encontrar um modelo que descreve dependências significativas entre as variáveis em um conjunto de dados ou em uma parte de um conjunto de dados. Detecção de desvio se concentra em descobrir as mudanças mais significativas em um conjunto de dados de valores previamente medidos. A categorização é a descoberta de uma função que categoriza um item em uma das várias categorias predefinidas (Fayyad et al., 1996a; Fayyad et al., 1996b; Kantardzic, 2011).

Formalmente, um categorizador é um modelo ou função M que prediz a categoria y de uma dada entrada x , sendo que $y = M(x)$, onde $y \in \{c1, c2, \dots, ck\}$ e cada ci é uma categoria (Zaki e Wagner, 2014). A figura 2 representa o diagrama simplificado do procedimento geral de construção do modelo de categorização.



Figura 2 - Procedimento geral de construção do modelo de categorização. Fonte: Raschka (2014).

Para criar o modelo, são necessários alguns dados corretamente categorizados, o que é chamado de conjunto de treinamento. Após treinar o modelo M , é possível prever a categoria de uma nova entrada. Têm sido propostos muitos tipos diferentes de modelos de classificação como Árvores de Decisão, categorizadores probabilísticos, Máquina de Vetores de Suporte (ou SVM do inglês *Support Vector Machine*), etc. (Zaki e Wagner, 2014).

Neste trabalho foi utilizado o categorizador probabilístico Naive Bayes, apresentado na próxima subseção.

2.1.4.1 Categorizador Naive Bayes

O Naive Bayes é um tipo de categorizador Bayesiano, que utiliza o teorema de Bayes para prever a categoria de uma dada entrada. Está entre uma das abordagens mais práticas para certos tipos de problemas de aprendizagem (Mitchell, 1997), na medida em que pressupõe que todos os atributos dos exemplos são independentes uns dos outros, dado o contexto da categoria (McCallum e Nigam, 1998). Frequentemente usado em problemas que envolvem a categorização de textos; pode-se afirmar que é computacionalmente eficiente (o treinamento e a categorização são realizados com apenas uma passagem sobre os dados) e de fácil implementação; são dois os tipos de modelos comumente utilizados para esta tarefa: o multivariado Bernoulli e o multinomial (McCallum e Peng, 2004).

O modelo multivariado Bernoulli baseia-se em dados binários: cada termo no vetor característico de um documento está associado com o valor 1 ou 0 (Raschka, 2014). O valor 1 significa que a palavra ocorre no documento e 0 significa que a palavra não ocorre no documento (Raschka, 2014).

O modelo multinomial é uma abordagem alternativa para categorização de documentos – em vez de valores binários – é utilizada a frequência de um termo; que é definida como o número de vezes que um determinado termo t (isto é, uma palavra ou *token*) aparece em um documento d (esta abordagem é, por vezes, também chamada de frequência bruta) (Raschka, 2014).

Os estudos de McCallum e Nigam (1998) indicam que o modelo multinomial obtém resultados melhores do que o modelo multivariado Bernoulli na tarefa de categorização de documentos.

Na prática o Naive Bayes compete bem se comparado com classificadores mais sofisticados (Chhajed et al., 2015). O estudo realizado por Michie et al. (1994) que

compara o Naive Bayes com outros algoritmos como Árvore de Decisão e Redes Neurais em diferentes tipos de tarefas comprovaram que o Naive Bayes obtém bons resultados quando os termos são independentes. Para a tarefa específica de categorização de textos, os estudos de Camargo (2007), Melo (2007) e Corrêa (2002) mostraram que o Naive Bayes Multinomial (NBM) obteve bons resultados quando comparado com o algoritmo SVM, redes neurais artificiais do tipo Multi-Layer Perceptron (MLP) e do tipo Self-Organizing Maps (SOM) e técnicas de aprendizado tradicionais como árvores de decisão (C4.5) e regras de decisão (PART).

Para que o NBM obtenha esses resultados, é importante que na etapa de pré-processamento sejam selecionados termos diferentes (independentes) e relevantes, uma vez que este categorizador utiliza a frequência com que as palavras aparecem nos documentos. O conjunto de textos representados na tabela 1 é um exemplo que afetaria o desempenho pois os termos são os mesmos, o que diferencia é a ordem com que estes aparecem, ou seja, não são independentes.

| | |
|---|--|
| 1 | Ele saiu de Londres, Ontario para Londres, Inglaterra. |
| 2 | Ele saiu de Londres, Inglaterra para Londres, Ontario. |
| 3 | Ele saiu da Inglaterra para Londres, Ontario. |

Tabela 1 - Conjunto de documentos problemáticos para o Naive Bayes Multinomial. Fonte: Manning et al. (2009).

O teorema de Bayes é definido por:

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

Onde $P(c)$ é a probabilidade *a priori* de que a hipótese c seja verdadeira na ausência de qualquer evidência específica, ou seja, é a probabilidade da categoria sem considerar qualquer entrada. $P(x)$ é a probabilidade *a priori* de que a hipótese x seja verdadeira na ausência de qualquer evidência específica, ou seja, é a probabilidade da entrada sem considerar as categorias.

$P(x/c)$ é a probabilidade de que a hipótese x seja verdadeira dada a evidência de c . De modo mais geral, escreve-se $P(x/y)$ para denotar a probabilidade de x dado y

(Mitchell, 1997). Nesse caso é a probabilidade de uma palavra x dada a evidência de uma categoria c .

Em problemas de aprendizado de máquina, estamos interessados na probabilidade $P(c/x)$ onde a hipótese c é verdadeira dada a evidência de x . $P(c/x)$ é chamada de probabilidade a *posteriori* de c , pois reflete a probabilidade de que c seja verdadeira dada a existência de x . É possível notar que a probabilidade a *posteriori* $P(c/x)$ reflete a influência de x , em contraste com a probabilidade a *priori* $P(c)$ que é independente de x (Mitchell, 1997). Nesse caso é a probabilidade de uma categoria c dada a evidência de uma palavra x .

A seguir é apresentado um exemplo⁴ de uma aplicação do NBM na categorização de documentos. Considere o seguinte conjunto de treinamento representado na tabela 2.

| | Documento | Palavras | Categoria |
|-------------------------|-----------|--------------------------------|-----------|
| Conjunto de Treinamento | 1 | china pequim china | c |
| | 2 | china china xangai | c |
| | 3 | china macao | c |
| | 4 | toquio japao china | j |
| Teste | 5 | china china china toquio japao | ? |

Tabela 2 - Exemplo de uma aplicação do Naive Bayes Multinomial.

Supondo que é preciso descobrir a categoria mais provável para o documento 5 entre c (China) e j (Japão):

1) Calcula-se as probabilidades prévias:

$$P(\text{categoria}) = (\text{número de casos da categoria})/(\text{número total de casos})$$

$$P(c) = 3/4$$

⁴ Manning, C. D. et al. (2009) An Introduction to Information Retrieval. Inglaterra, Cambridge University Press, p. 253 – 288.

$$P(j) = 1/4$$

2) Calcula-se as probabilidades condicionais, ou seja, a probabilidade de cada uma das palavras em relação a cada categoria. Para evitar possíveis zeros (caso em que a palavra não existe em nenhuma instância da categoria), é adicionado 1 (um) ao número de ocorrências da palavra x na categoria. O termo $|V|$ é o número de palavras existentes no vocabulário, que no caso são 6 (seis) diferentes palavras:

$$P(x|categoria) = (\text{número de ocorrências da palavra } x \text{ na categoria} + 1) / (\text{número total de palavras da categoria} + |V|)$$

$$P(\text{china}|c) = (5+1)/(8+6) = 3/7$$

$$P(\text{toquio}|c) = (0+1)/(8+6) = 1/14$$

$$P(\text{japao}|c) = (0+1)/(8+6) = 1/14$$

$$P(\text{china}|j) = (1+1)/(3+6) = 2/9$$

$$P(\text{toquio}|j) = (1+1)/(3+6) = 2/9$$

$$P(\text{japao}|j) = (1+1)/(3+6) = 2/9$$

3) Com as probabilidades prévias e condicionais calculadas é possível calcular a probabilidade do documento pertencer a cada uma das categorias. Como a probabilidade $P(x)$ possui valor constante, esta não precisa ser considerada:

$$\text{argmax } P(\text{categoria}|x_1 \dots x_i) = \text{argmax } \prod_i P(x_i|\text{categoria}) \times P(\text{categoria})$$

$$\text{argmax } P(c|d5) = \text{argmax } P(\text{china}|c)^3 * P(\text{toquio}|c) * P(\text{japao}|c) * P(c) = (3/7)^3 * 1/14 * 1/14 * 3/4 = 0.0003$$

$$\text{argmax } P(j|d5) = \text{argmax } P(\text{china}|j)^3 * P(\text{toquio}|j) * P(\text{japao}|j) * P(j) = (2/9)^3 * 2/9 * 2/9 * 1/4 = 0.0001$$

É possível verificar que a probabilidade do documento 5 pertencer a categoria c (China) é maior do que a probabilidade de ele pertencer a categoria j (Japão).

2.1.5 Interpretação e Avaliação dos Resultados

Como o objetivo é a categorização de uma notícia, nesta etapa o resultado obtido deve ser avaliado para verificar se a notícia foi corretamente categorizada ou não. Pode-se ainda utilizar um ou mais especialistas nessa fase para ajudar nesta avaliação.

Caso o resultado não seja o esperado, é possível voltar o processo de KDD para uma das etapas (ver figura 1) e fazer os devidos ajustes até que o resultado final seja satisfatório.

2.2 Projetando Aplicações para a Web Social

Para o desenvolvimento da interface e das funcionalidades do protótipo do sistema voltadas para *web social*, foram seguidos alguns princípios, práticas e padrões propostos por Bell (2009) e por Crumlish e Malone (2009), que possuem o intuito de tornar uma aplicação mais atrativa e social para os seus usuários. Esses princípios, práticas e padrões são apresentados nesta seção.

2.2.1 Layout

Diferentemente dos *sites* tradicionais e dos *blogs*, *sites* voltados para a web social possuem o menu de navegação no topo junto com a marca. Na parte inferior da página está uma navegação mais completa com *links* para outras seções do site (Bell, 2009). Na figura 3 é possível visualizar as disposições dos elementos nos diferentes tipos de *layouts*.

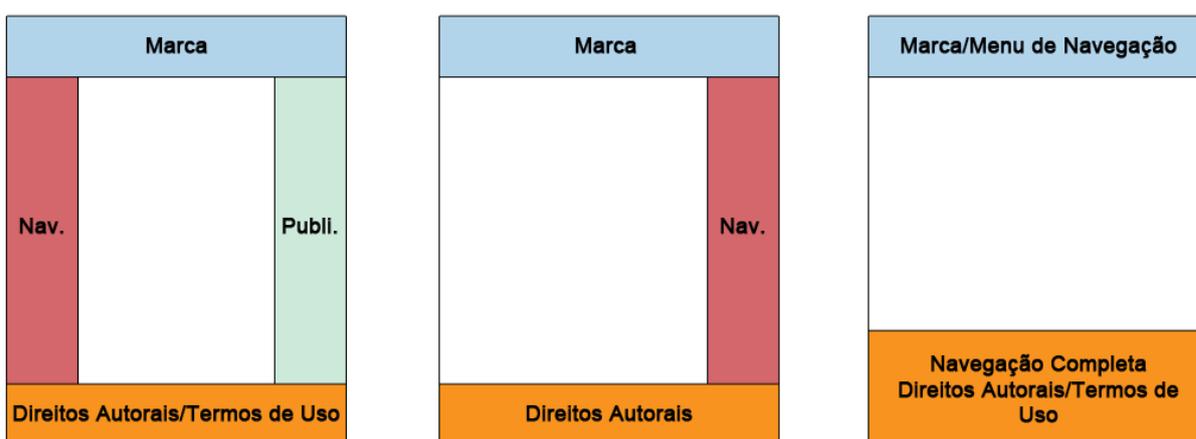


Figura 3 - Tipos de *layouts* mais comuns. Fonte: Bell (2009), p. 244.

2.2.2 Objeto Social

Segundo Bell (2009), ao desenvolver um *site* voltado para *web social*, é necessário identificar o objeto que irá estimular a interação entre os usuários. O Twitter,

por exemplo, tem como objeto social as mensagens de texto curtas, as *hashtags* e os *retweets*, já o Youtube os vídeos, os votos positivos e negativos e os comentários.

2.2.3 Funcionalidades

Nesta seção são apresentadas as funcionalidades voltadas para *web* social que foram aplicadas no protótipo.

2.2.3.1 Cadastro

Segundo Crumlish e Malone (2009), o cadastro é uma funcionalidade essencial para aumentar o engajamento dos usuários. Com o cadastro, o usuário poderá acessar parte do *site* ou aplicação onde necessita criar ou salvar informações pessoais.

2.2.3.2 Login/Logout

Ainda segundo Crumlish e Malone (2009), o sistema de *login*, assim como o cadastro, é importante para aumentar o engajamento dos usuários com o *site*. Através do *login*, o usuário poderá acessar sua informação personalizada que está armazenada no *site*. Com o *logout*, o usuário termina a sua sessão. O usuário poderá fazer o *login* novamente após o *logout*.

2.2.3.3 Comentários

O comentário é uma ferramenta que possibilita ao usuário adicionar mais informações em algo criado por outra pessoa (Bell, 2009). Permitir que as pessoas comentem ou marquem o conteúdo de outras é a essência de muitas atividades de mídia social; existe um ciclo: O usuário compartilha algum conteúdo e obtém um feedback da comunidade, o que provavelmente irá incentivá-lo a continuar a sua participação (Bell, 2009).

Jornais permitem que comentários sejam feitos em suas histórias, revistas científicas permitem comentários em seus artigos, e em muitos *sites* sociais, os usuários podem comentar em conteúdos gerados por outros. Comentários estão se tornando parte dos *sites* jornalísticos e é onde a interação dos leitores ocorre (Bell, 2009).

2.2.3.4 Classificação (“Rating”) de Publicações

Classificação de conteúdo se enquadra na categoria de atividades com baixa sobrecarga cognitiva. Avaliação em uma escala de cinco pontos (conhecido como a

escala de *Likert*), é bastante comum, ao lado da classificação simples do tipo sim/não. Essa é uma técnica bastante popular e presente em alguns *sites* que permite uma pessoa classificar o conteúdo de outra e indiretamente classificar a pessoa que criou esse conteúdo (Bell, 2009).

3 Estado da Arte

Nesta seção são analisados alguns *sites* de publicação e divulgação de notícias que utilizam a categorização automática de documentos como uma de suas funcionalidades.

3.1 Europe Media Monitor News Explorer

O Europe Media Monitor News Explorer (EMM News Explorer) é um *site* desenvolvido pela *Joint Research Centre* para a Comissão Europeia denominada *Europe Media Monitor* (ver figura 4). É um sistema de monitoramento de mídia em tempo real que detecta notícias publicadas em *sites* de notícias na internet em 30 idiomas. Segundo Fogelman-Soulié et al. (2007), a União Europeia depende do sistema para receber alertas de notícias mais mencionadas. Também é possível comparar como um mesmo evento foi relatado em diferentes localidades.

The screenshot shows the EMM News Explorer interface. At the top, there are tabs for 'EMM NewsBrief' and 'EMM NewsExplorer', along with search fields for 'Name Search' and 'Text Search'. The main header reads 'News Analysis' and 'Daily News Analysis, across languages and over time'. Below this, there's a 'Main Menu' section with 'News Summary' and 'About EMM NewsExplorer'. A 'News language and date' section allows filtering by language (en-English) and date (Nov 2015). The central part of the page is titled 'Clustered news for Tuesday, November 3, 2015' and features a world map with location markers. Below the map, there's a list of countries with their respective news counts: United States (287), China (54), India (62), Russian Federation (62), United Kingdom (61), Germany (60), Pakistan (41), South Africa (41), Iraq (31), Spain (31), Australia (20), Viet Nam (29), Japan (27), New Zealand (24), Turkey (23), Korea, Republic Of (19), Canada (17), Switzerland (17), Myanmar (16), Yemen (15), Cuba (15), Syrian Arab Republic (14), Israel (13), France (12), and Egypt (11). To the right, there are sections for 'This Week's New Stories' and 'This Month's New Stories', each listing recent news items with dates and brief descriptions. At the bottom, there's a 'Timeline' section for 11/2015 and a 'Biggest Story this Month' section highlighting the Russian plane crash in Egypt.

Figura 4 - Tela inicial do *site* EMM News Explorer. Fonte: *print screen* do *site* EMM News Explorer.

O sistema possui várias funcionalidades como a sugestão de artigos com assuntos parecidos com os visualizados pelo usuário e o agrupamento automático de notícias de acordo com o seu conteúdo. O *site* não possui informações sobre qual foi o algoritmo utilizado para essa tarefa.

3.2 News Explorer

O *News Explorer* é um sistema que categoriza notícias e permite visualizá-las com base na sua localização, organização e autores; permite que artigos sejam categorizados em criminais, políticos, etc. (Desai et al., 2014). É um projeto que teve como inspiração sistemas como o *EMM News Explorer*, porém com o intuito de desenvolver um melhor mecanismo de categorização e também uma interface gráfica mais intuitiva (ver figura 5).

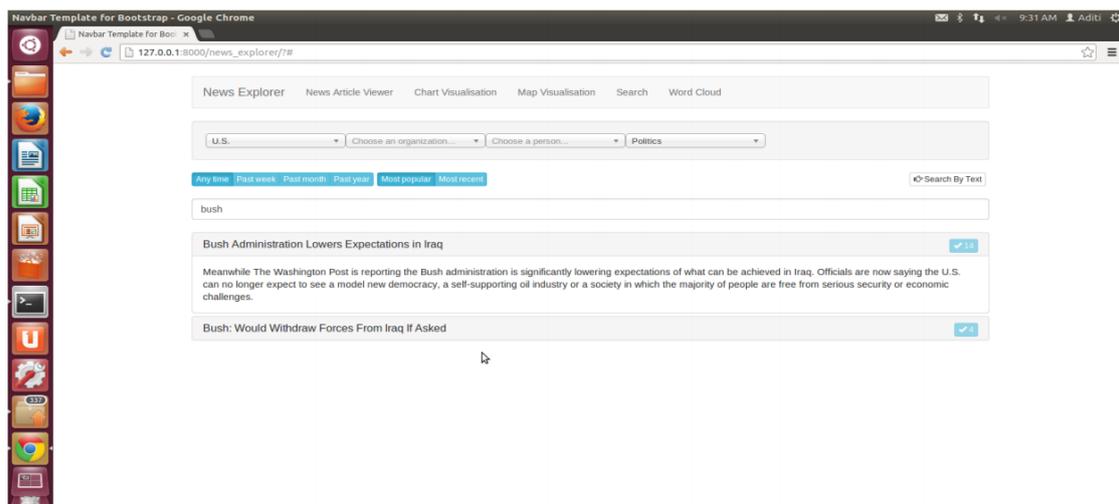


Figura 5 - Tela de busca de artigos do News Explorer. Fonte: Desai et al. (2014).

Para a categorização das notícias foram utilizados algoritmos como o “k-vizinhos mais próximos”, e o Naive Bayes. Segundo Desai et al. (2014), esses categorizadores obtiveram boa acurácia.

3.3 Cora

Cora é um sistema que foi desenvolvido para apresentar a aplicação real de várias técnicas de aprendizado de máquina. Segundo McCallum et al. (2000), o sistema obtém artigos científicos voltados para área da computação disponíveis na *internet* e

oferece aos usuários funcionalidades como a busca por artigos através de palavras-chave, agrupamento de forma hierárquica de acordo com o seu tópico, mapeamento das relações entre os artigos de acordo com os *links* de citações e o fornecimento das informações bibliográficas (ver figura 6).



Figura 6 - Página inicial do site Cora. Fonte: McCallum et al. (2000).

Para a categorização dos artigos, o sistema também utiliza o categorizador Naive Bayes Multinomial. Segundo McCallum et al. (2000), este categorizador obteve resultados satisfatórios.

3.4 News Explorer for Web Streaming Data

O News Explorer for Web Streaming Data (NEWSD) é uma ferramenta de mineração de texto que possui uma Interface Gráfica do Usuário (ou GUI do inglês *Graphical User Interface*) desenvolvido para a categorização de dados existentes na *web* (Mohiuddin et al., 2015) (ver figura 7).



Figura 7 - Interface do NEWSD. Fonte: Mohiuddin et al. (2015).

O sistema extrai dados das páginas de *sites* de notícias no Facebook para então realizar as etapas de pré-processamento, categorização e visualização. Para extrair os dados do Facebook, foi usada a Interface de Programação de Aplicação (ou API do inglês *Application Programming Interface*) do Facebook Graph; o sistema estabelece uma conexão com o Facebook e então obtém acesso aos dados disponíveis nas páginas de divulgação de notícias em tempo real. Para realizar as etapas de pré-processamento e categorização, foi usado a API da ferramenta de mineração de dados Weka. Os algoritmos de categorização usados foram o Naive Bayes e o J48 (implementação do algoritmo C4.5 na linguagem de programação Java).

Segundo Mohiuddin et al. (2015), o categorizador Naive Bayes obteve uma acurácia de 61.2857% enquanto o J48 obteve 52.1428%.

4 Categorizador Semiautomático

Para o desenvolvimento do categorizador semiautomático, foi realizado os seguintes passos: seleção, pré-processamento, transformação, mineração dos dados, interpretação e avaliação dos resultados. Foram realizados testes iniciais utilizando a ferramenta de *data mining* Weka. Para alcançar o objetivo deste trabalho, foi usada a categorização, que é a tarefa usada para prever a categoria de uma entrada e o algoritmo aplicado foi o Naive Bayes Multinomial.

Para o desenvolvimento das funcionalidades do protótipo voltadas para a web social foram adotados princípios e padrões que também serão apresentados neste capítulo.

4.1 Weka

O Weka é um software para mineração de dados bastante popular. Foi desenvolvido pela Universidade de Waikato (Nova Zelândia) e possui a licença *General Public License* (GNU), portanto é um software livre.

Algumas das funcionalidades do Weka envolvem o pré-processamento de dados, algoritmos para categorização, regressão, clusterização, regras de associação e visualização. Como ele possui a licença GNU, outras funcionalidades podem ser desenvolvidas por qualquer pessoa.

Foi implementado na linguagem Java e possui uma GUI, mas também pode ser empregado a partir de linhas de comandos. Também pode ser utilizado como uma biblioteca Java para aplicativos desenvolvidos nesta linguagem.



Figura 8 - Tela inicial do Weka.

Possui vários tipos de aplicações disponíveis como o “*Explorer*” e o “*Experimenter*” (ver figura 8). Para realizar as tarefas descritas nas próximas seções foi utilizado a opção “*Explorer*”.

O Weka foi aplicado nos testes iniciais para verificar o desempenho dos categorizadores disponíveis (ver subseção 4.5).

4.2 Base de Dados e Seleção

A base de dados utilizada para formar o conjunto de treinamento é composta por notícias extraídas do Corpus de Extratos de Textos Electrónicos NILC/Folha de S. Paulo (CETENFolha) que é um corpus criado pelo projeto Processamento Computacional do Português formado por artigos do jornal Folha de São Paulo do ano de 1994 (mil novecentos e noventa e quatro) e que estão divididos em: esporte, imóveis, informática, política e turismo. Neste trabalho foram usados somente os artigos pertencentes as categorias esporte e política. Também foram coletadas notícias sobre economia que foram retiradas do *site* Globo.com. No total, a base de treinamento possui 105 (cento e cinco) artigos, 35 (trinta e cinco) para cada categoria.

As notícias foram salvas em um arquivo do tipo *.arff*, que é uma das extensões utilizadas pelo Weka, o qual contém os atributos e a lista dos dados. Foram definidos dois atributos, um denominado “*text*”, que se refere ao conteúdo, e outro “*category*”, que define a categoria de cada artigo, podendo ser “Esporte”, “Política” ou “Economia” (ver figura 9). Para abrir o arquivo no Weka, é preciso selecionar a opção “*Open file*” e

então é possível visualizar a quantidade de instancias, atributos e a quantidade de instancias por atributo (ver figura 10).

```

1 @relation noticias
2 @attribute text String
3 @attribute category {'Esporte','Política','Economia'}
4 @data
5 'parlamento grego aprova referendo sobre oferta credores deputados votaram favor referendo proposto alexis tsipras referendo realizado o
6 'dilma chega nova york lado itiva ministros presidente tentara atrair investimentos tera encontro obama saida brasilia atrasou devido reunia
7 'ministro aloizio mercadante aponta enfase ataque pt supostamente citados delacao premiada empreiteiro ricardo pessoa doou campanha
8 'ajuste ministros trocam gabinete popo deputados orientacao negociar pessoalmente propostas levy eliseu padilha gabas circulam diariam
9 'ministro vazamento seletivo fez doacoes dilma edinho silva tesoureir campanha defendeu entrevista chamou mentiras supostas declarac
10 'delator aponta teriam recebido dinheiro esquema revista veja listou nomes supostamente delatados empreiteiro dono utc ricardo pessoa t
11 'cpi petrobras remarca acareacoes investigados lava jato pedro barusco renato duque serao confrontados cpi dia seguinte barusco passa
12 'mercosul estuda negociacao sindical coletiva trabalhadores bloco declaracao assinada ministros submetida presidentes bloco tambem a
13 'china disposta importar brasil ministro wang yang viajou brasil reuniu dilma rousseff michel temer vice brasileiro quer autorizacao frigorific
14 'camara empresas entrega projetos shopping prazo terminava nesta sexta estendido agosto primeirosecretario contesta interpretacao obr
15 'camara barra pl reduz incentivo empresas bebidas parlamentares estado afirmam proposta impacta mil empregos proposta poderia eleva
16 'thiago silva nega penalti lembro tocar mao bola zagueiro surpreende infracao garantiu empate paraguai partida tirou brasil copa america ir
17 'neymar desfalca brasil jogo eliminatorias camisa cumprira suspensao copa america outubro neymar desfalcara brasil jogos selecao elimi
18 'robinho evita apontar viloes apos eliminacao boeira nossa atacante autor gol empate paraguai selecao conseguiu matar jogo polemiza te
19 'leo moura rescinde strikers pede ajuda jogar itiba rafael bertani presidente conselho administracao saida jogador amigavel garante enviara
20 'roger nega rendimento elogia jogo feito fora casa tricolor marcou gols minutos sofreu pressao segundo triunfo sobre avai ressacada neste
21 'marreta acerta chutaco cabeça nocauteia bosse segundos exparticipante tuf brasil consegue vitoria espetacular despachar steve bosse
22 'jogo festivo cruzeiro tem gol feio alex penalti samuel rosa acostumado marcar golacos exmeia homenageado partida mineirao vitoria time
23 'bahia marca vence luverdense entra serie garoto base joao leonardo marca gol vitoria tricolor ajuda time aproximar primeiras posicoes lu
24 'gol chape empata sport precisara secar rivais seguir lider pernambucanos marcam primeiro gol partida etapa inicial tentam suportar press
25 'exercicios dicas podem amenizar sindrome joelho redor sindrome femoropatelar gera dor parte joelho problematica quem pratica rida con
26 'federer critica regra vestimenta wimbledon ridiculamente rigida heptacampeao torneio suico reclama tradicao obriga tenistas usarem aper
27 'abc náutico ficam empate jogo movimentado cheio gols pernambucanos ficam tres vezes vantagem placar frasqueirao potiguares buscar
28 'rivaldo campo crb bate lanterna mogi fora casa gol contra zagueiro renato camilo define vitoria regatiana interior paulista banco rivaldo ass

```

Figura 9 - Estrutura do arquivo *.arff* contendo a base de treinamento.

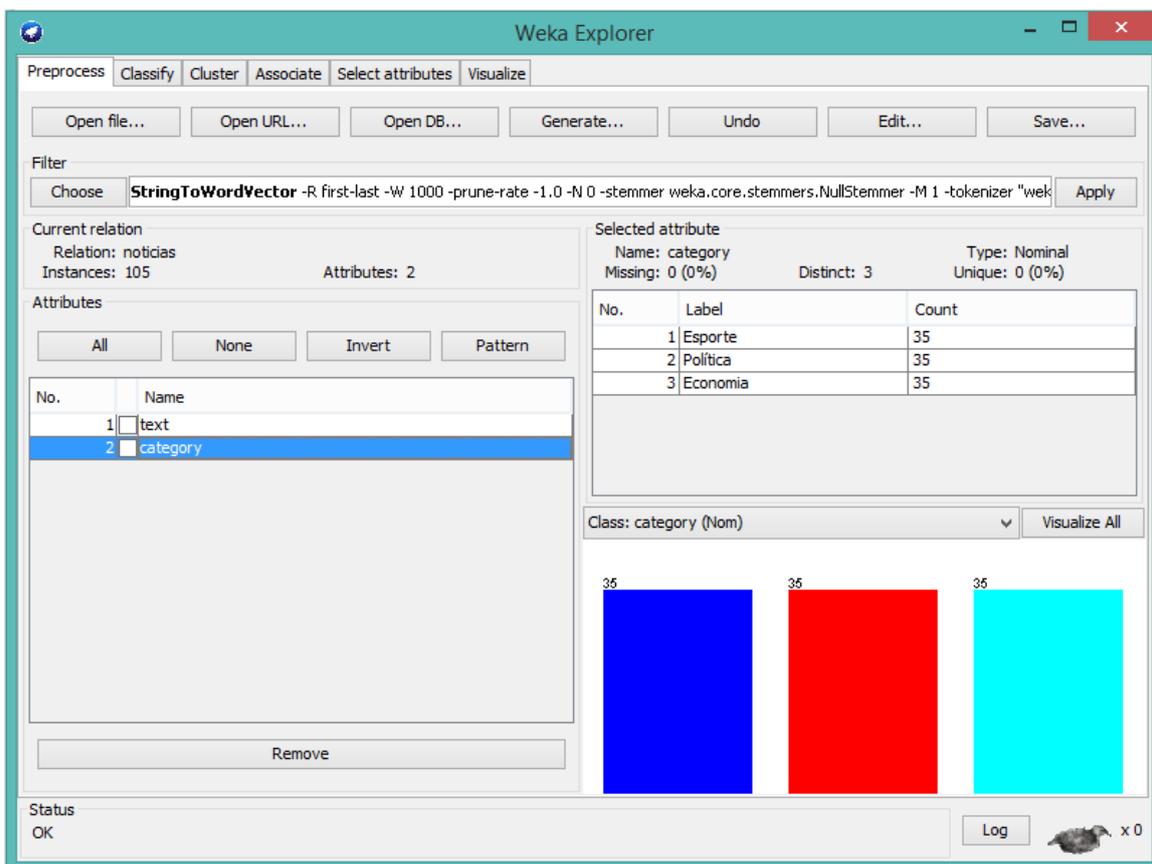


Figura 10 - Quantidade de instancias, atributos e a quantidade de instancias por atributo.

4.3 Pré-processamento

Como as notícias não estavam pré-processadas, foi necessário realizar o procedimento de remoção dos sinais de pontuação, conjunções, preposições, artigos e as “*stopwords*”, além da troca das letras maiúscula pelas minúsculas e de outros caracteres como descrito na subseção 2.1.3.

Os sinais diacríticos (acentos gráficos, cedilha, til, trema) também foram retirados, pois, se o usuário inserir uma palavra sem o sinal, esta será considerada diferente da palavra escrita corretamente. Numerais e sinais gráficos como o apóstrofo, parênteses, colchetes, etc., também foram removidos pois não acrescentam valor para o conjunto de treinamento no caso de categorização de notícias.

Foi necessário modificar as notícias, que se encontravam em forma de uma única *string*, para o modelo “*bag of words*”. O Weka possui vários tipos de filtros e um deles realiza esse tipo de tarefa. Para selecionar esse filtro deve-se escolher a opção “*Choose*”, abrir as opções em “*filters*” e em “*unsupervised*” e então selecionar o filtro “*StringToWordVector*” (ver figura 11). É necessário marcar o atributo “*category*” antes de aplicar o filtro.

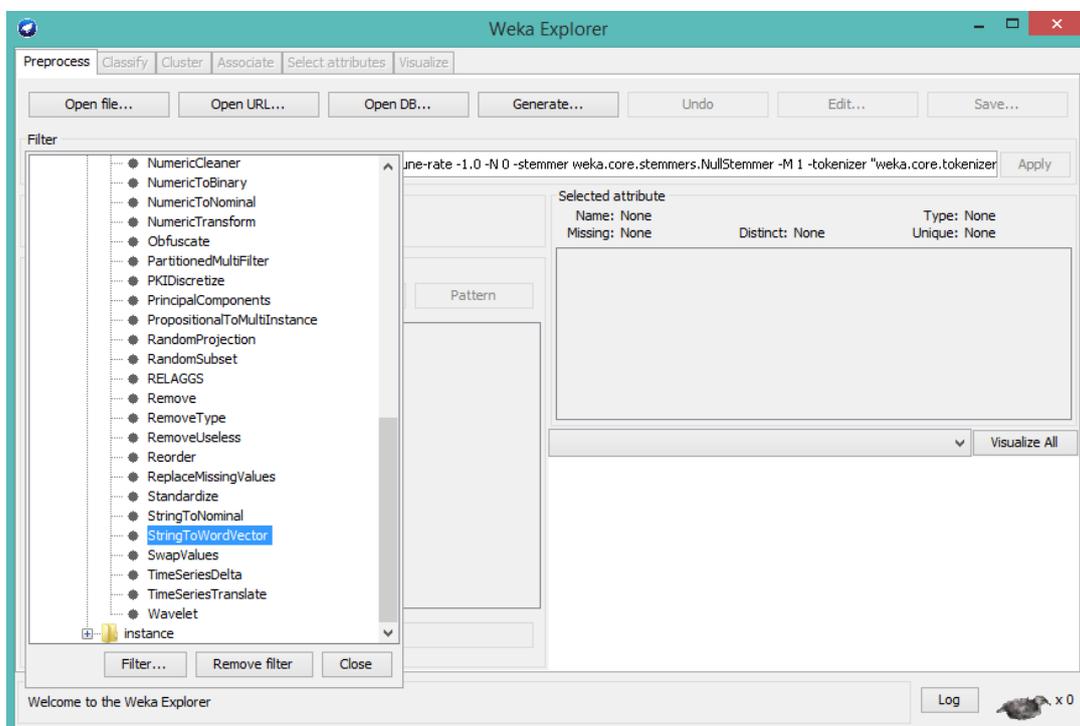


Figura 11 - Tipos de filtros disponíveis no Weka.

4.4 Treinamento e Categorização

Para utilizar um categorizador, é preciso treiná-lo com a base que possui os dados que foram selecionados e pré-processados e então realizar os testes. O Weka possui vários tipos de categorizadores. Antes de escolher um deles, é preciso marcar o atributo “category”. Para selecionar um deles, deve-se clicar na aba “Classify” e depois na opção “Choose”.

É preciso selecionar uma das opções de teste, neste trabalho foi selecionada a opção “Cross-validation” com dez conjuntos de teste. Ele irá produzir dez conjuntos, cada um dividido em dois grupos: noventa instâncias utilizadas para o treinamento e dez instâncias utilizadas para o teste. O algoritmo é executado para esse conjunto e então outro conjunto é criado e esse processo é repetido até que todos os dez conjuntos tenham sido criados e testados e então é feita uma média do desempenho do algoritmo (ver figura 12).

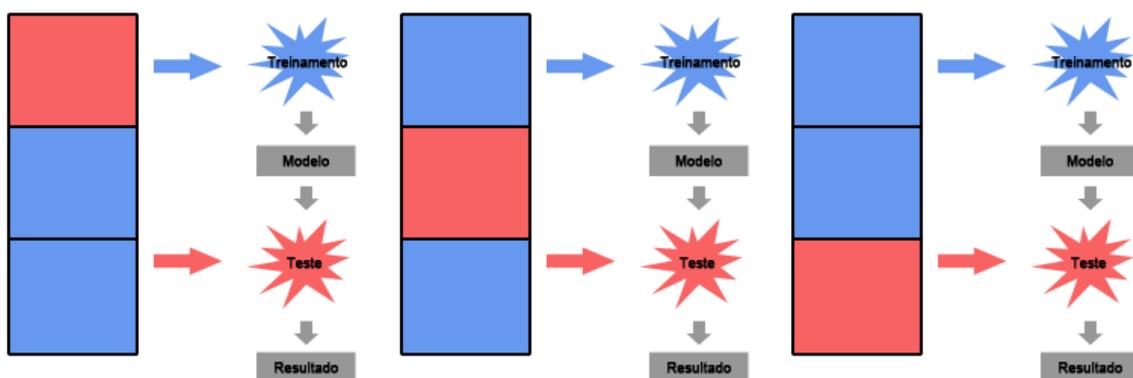


Figura 12 - Funcionamento do “Cross-validation”. Fonte: Refaeilzadeh et al. (2008).

4.5 Avaliando os Resultados

Foram realizados testes no Weka com os algoritmos J48 (árvore de decisão), PART (regra de decisão), Naive Bayes (modelo Bernoulli) e o Naive Bayes Multinomial.

O J48 categorizou 67 (sessenta e sete) instâncias corretamente (63.8095%). O Weka também produz uma matriz de confusão que mostra como as instancias foram categorizadas em cada categoria. A tabela 3 representa a matriz gerada.

| Categorizado como -> | Esporte | Política | Economia |
|----------------------|---------|----------|----------|
| Esporte | 20 | 3 | 12 |
| Política | 4 | 15 | 16 |
| Economia | 1 | 2 | 32 |

Tabela 3 - Matriz de confusão do J48.

É possível verificar que para as categorias “Esporte”, “Política” e “Economia”, 20 (vinte), 15 (quinze) e 32 (trinta e duas) instancias foram categorizadas corretamente respectivamente. Para a categoria “Esporte” 3 (três) foram categorizadas erroneamente como “Política” e 12 (doze) como “Economia”, já para “Política”, 4 (quatro) foram categorizadas erroneamente como “Esporte” e dezesseis como “Economia” e para “Economia”, 1 (uma) foi categorizada erroneamente como “Esporte” e 2 (duas) como “Política”.

O PART categorizou sessenta e quatro instancias corretamente (60.9524%). A tabela 4 representa a matriz de confusão gerada.

| Categorizado como -> | Esporte | Política | Economia |
|----------------------|---------|----------|----------|
| Esporte | 21 | 3 | 11 |
| Política | 3 | 17 | 15 |
| Economia | 3 | 6 | 26 |

Tabela 4 - Matriz de confusão do PART.

É possível verificar que para a categoria “Esporte”, “Política” e “Economia”, 21 (vinte e uma), 17 (dezesete) e 26 (vinte e seis) instancias foram categorizadas corretamente respectivamente. Para a categoria “Esporte” 3 (três) foram categorizadas erroneamente como “Política” e 11 (onze) como “Economia”, já para “Política”, 3 (três) foram categorizadas erroneamente como “Esporte” e 15 (quinze) como “Economia” e para “Economia”, 3 (três) foram categorizadas erroneamente como “Esporte” e 6 (seis) como “Política”.

O Naive Bayes categorizou 78 (setenta e oito) instancias corretamente (74.2857%). A tabela 5 representa a matriz de confusão gerada.

| Categorizado como -> | Esporte | Política | Economia |
|----------------------|---------|----------|----------|
| Esporte | 25 | 1 | 9 |
| Política | 2 | 23 | 10 |
| Economia | 1 | 4 | 30 |

Tabela 5 - Matriz de confusão do Naive Bayes.

É possível verificar que para a categoria “Esporte”, “Política” e “Economia”, 25 (vinte e cinco), 23 (vinte e três) e 30 (trinta) instancias foram categorizadas corretamente respectivamente. Para a categoria “Esporte” 1 (uma) foi categorizada erroneamente como “Política” e 9 (nove) como “Economia”, já para “Política”, 2 (duas) foram categorizadas erroneamente como “Esporte” e 10 (dez) como “Economia” e para “Economia”, 1 (uma) foi categorizada erroneamente como “Esporte” e 4 (quatro) como “Política”.

Já o Naive Bayes Multinomial categorizou 83 (oitenta e três) instancias corretamente (79.0476%) e levou um tempo muito baixo para gerar o modelo. A tabela 6 representa a matriz de confusão gerada.

| Categorizado como -> | Esporte | Política | Economia |
|----------------------|---------|----------|----------|
| Esporte | 28 | 3 | 4 |
| Política | 1 | 26 | 8 |
| Economia | 1 | 5 | 29 |

Tabela 6 - Matriz de confusão do Naive Bayes Multinomial.

É possível verificar que para a categoria “Esporte”, “Política” e “Economia”, 28 (vinte e oito), 26 (vinte e seis) e 29 (vinte e nove) instancias foram categorizadas corretamente respectivamente. Para a categoria “Esporte” 3 (três) foram categorizadas

erroneamente como “Política” e 4 (quatro) como “Economia”, já para “Política”, 1 (uma) foi categorizada erroneamente como “Esporte” e 8 (oito) como “Economia” e para “Economia”, 1 (uma) foi categorizada erroneamente como “Esporte” e 5 (cinco) como “Política”.

Por possuir um bom desempenho comparado com os outros algoritmos e baixo tempo de execução, além das qualidades discutidas na subseção 2.1.4.1, foi escolhido o categorizador Naive Bayes Multinomial para realizar a tarefa de categorização das notícias no protótipo do sistema.

Também foi possível verificar, através das matrizes de confusão, que os algoritmos costumam errar mais na categorização entre notícias do tipo política e economia e vice-versa, o que é considerado normal uma vez que pode ser complicado mesmo para os especialistas pois podem possuir conteúdos semelhantes.

4.6 News Share e Aplicação no Mundo Real

O protótipo do sistema de divulgação de notícias foi desenvolvido com o foco em *web social*, portanto a interface e as funcionalidades foram elaboradas seguindo princípios e padrões apresentados nesta subseção.

A documentação do sistema (modelos conceitual e lógico do banco de dados, diagrama de casos de uso e descrição dos casos de uso) encontra-se nos apêndices.

4.6.1 Ferramentas Utilizadas

Nesta subseção são apresentadas as ferramentas utilizadas para a implementação do protótipo do sistema.

4.6.1.1 HTML 5

O sistema foi desenvolvido para a Web para que pudesse ser acessado tanto de computadores quanto de *smartphones* e *tablets*. Foi utilizada a linguagem HTML 5 (*Hypertext Markup Language*, versão 5), a versão mais atual e com mais recursos disponíveis, aplicada na estruturação e apresentação de conteúdos para a Web. Como quase todos os navegadores já possuem suporte para esta linguagem, o sistema possui menos risco de ficar defasado no futuro próximo.

4.6.1.2 Syntactically Awesome Style Sheets

Syntactically Awesome Style Sheets (Sass) é um pré-processador de *Cascading Style Sheets* (CSS), ele permite escrever pseudo-códigos que serão convertidos em códigos CSS, que são folhas de estilos baseadas em *tags* HTML. A vantagem em utilizar um pré-processador como o Sass, é que neste é possível declarar variáveis, criar funções e cálculos matemáticos, o que não é possível com o CSS.

4.6.1.4 Bootstrap

Como o intuito é que o sistema possa ser acessado de qualquer lugar e em qualquer momento, sua interface foi desenvolvida de forma adaptativa para diferentes tamanhos de telas. Para que isso fosse possível foi utilizado o *framework* CSS Bootstrap, muito popular para esse tipo de tarefa, pois possui uma *grid* adaptativa e vários componentes prontos para serem utilizados.

O sistema utiliza a *grid* do Bootstrap e alguns de seus componentes, porém o visual padrão do Bootstrap não foi adotado e quase todos os componentes tiveram seu visual modificados, como novas cores, imagens, bordas, etc.

4.6.1.5 Ruby on Rails

A parte *Back-End* do sistema foi desenvolvida utilizando a linguagem Ruby e seu *framework* Rails. Esta foi escolhida por possuir uma pequena curva de aprendizado além de incluir vários tipos de *plugins* e extensões o que tornou o desenvolvimento mais ágil.

O Rails utiliza o padrão *Model-View-Controller* (MVC), que determina a separação de responsabilidades de uma aplicação em componentes de modelo, visão e controle. O modelo é formado por entidades de negócio que representam os dados do sistema. Componentes de visão (telas gráficas ou páginas HTML, por exemplo) têm o objetivo de apresentar as informações pertencentes ao modelo e capturar eventos de usuário. Componentes de controle tratam os eventos capturados, notificam e coordenam componentes de modelo. Os componentes de controle fazem a ponte entre componentes de visão e do modelo (Souza et al., 2003).

Como a linguagem escolhida foi Ruby, o sistema não utilizou a biblioteca do Weka por ser desenvolvida em Java. Foi então utilizado um algoritmo em Ruby que implementa o categorizador Naive Bayes Multinomial. O algoritmo utiliza uma função de *pré-cache*, armazenado alguns parâmetros, como as palavras. Esses dados são consultados antes da categorização (Idris, 2012).

4.6.2 Layout

O *layout* foi desenvolvido seguindo o padrão proposto por Bell (2009) para aplicações voltadas para a *web* social e que foi apresentado na subseção 2.2.1. Na parte superior está o menu de navegação, no centro da página estão as notícias publicadas pelos usuários; qualquer visitante pode visualizá-las mesmo que não tenha realizado o *login* ou que não esteja cadastrado.

As notícias são exibidas pela ordem com que foram publicadas, sendo que as quatro mais recentes possuem destaque, são maiores do que as outras e ficam em um “carrossel animado” (ver figura 13). Na versão *mobile* do sistema, o tamanho das notícias é reduzido e é utilizado somente uma coluna ao invés de duas, para poder melhorar a navegação devido ao tamanho da tela (ver figura 14).

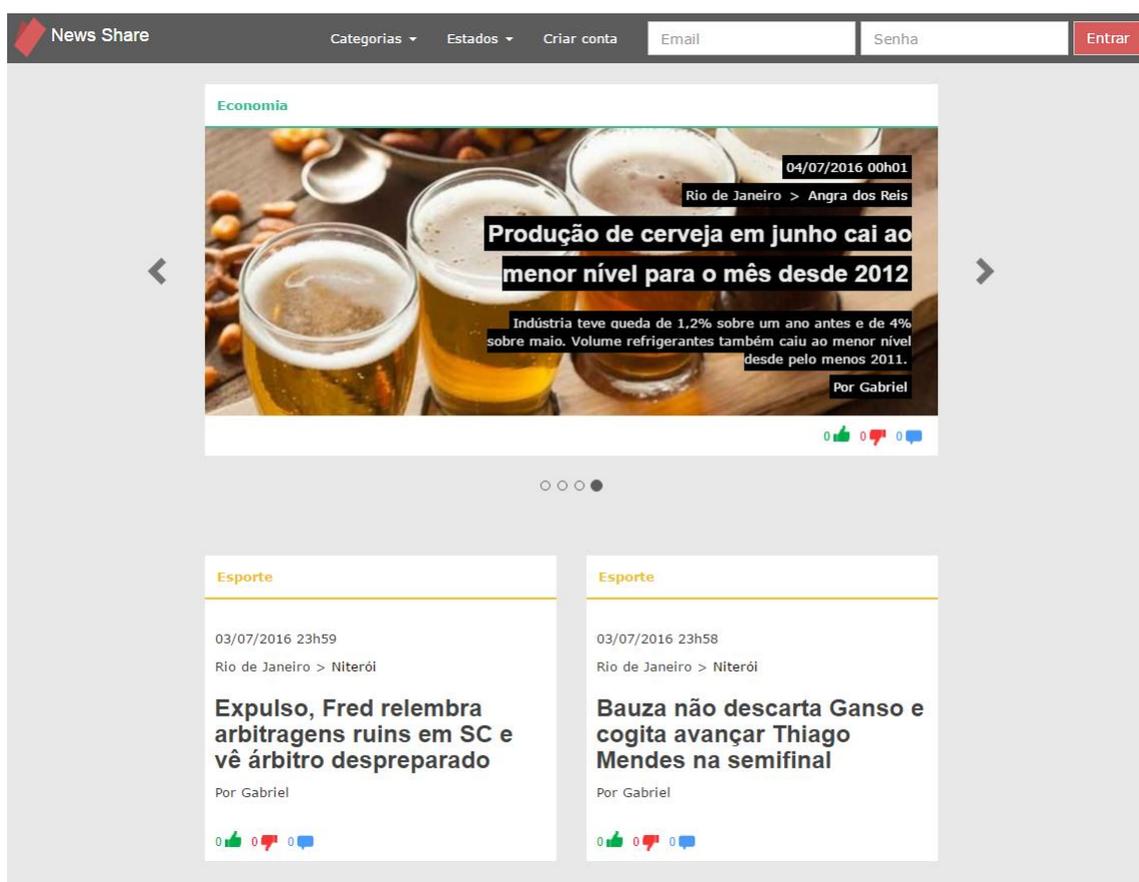


Figura 13 - Página inicial do sistema com algumas notícias publicada, retiradas do *site* Globo.com.



Economia



04/07/2016 00h01

Rio de Janeiro > Angra dos Reis

Produção de cerveja em junho cai ao menor nível para o mês desde 2012

Indústria teve queda de 1,2% sobre um ano antes e de 4% sobre maio. Volume refrigerantes também caiu ao menor nível desde pelo menos 2011.

Por Gabriel



Esporte

03/07/2016 23h59

Rio de Janeiro > Niterói

Expulso, Fred relembra arbitragens ruins em SC e vê árbitro despreparado

Por Gabriel



Figura 14 – Layout versão mobile.

A escolha por posicionar as notícias dessa maneira também foi influenciada por sites de notícias como o Globo.com (ver figura 15) e também por jornais tradicionais em sua versão impressa (ver figura 16) que possuem na sua capa a matéria principal que fica em destaque e as outras notícias ao seu redor e com tamanhos reduzidos.



Figura 15 - Site Globo.com. Fonte: print screen do site Globo.com.



Figura 16 - Capa do jornal The New York Times. Fonte: Diário do Rio.

4.6 Objeto Social

Como foi apresentado na seção 2.2.2, ao desenvolver um sistema voltado para a *web* social é importante identificar o objeto que irá provocar a interação entre os usuários. No caso do sistema descrito neste trabalho, o objeto principal ou social são as notícias e os elementos que a compõem: título, resumo, descrição, imagem, comentários, categoria e os votos positivos e negativos.

4.6.3 Funcionalidades

As funcionalidades do sistema foram desenvolvidas utilizando os padrões, práticas e princípios apresentados na seção 4.6.3. No sistema, o usuário pode realizar o cadastro e o *login/logout*. Também é possível comentar uma notícia, acrescentando mais informações. Na figura 17 é possível ver um exemplo de uma notícia que recebeu 247 (duzentos e quarenta e sete) comentários. A figura 18 mostra uma notícia publicada exibindo o campo para inserir comentários. Somente usuários cadastrados e que realizaram o *login* podem comentar uma notícia.



Figura 17 - Exemplo de uma notícia publicada, retirada do *site* Globo.com.

The image shows a screenshot of a news article on the website Globo.com. At the top left, the category 'Economia' is displayed in green. Below it, the date and time '03/07/2016 23h41' and the location 'Rio de Janeiro > Niterói' are shown. The main headline is 'Petrobras demitirá três pessoas após investigação com origem na Lava Jato' in large, bold black font. A sub-headline reads: 'Informação foi dada por comunicado interno da estatal, segundo a Reuters. Foi apurado o envolvimento de 26 pessoas nas irregularidades.' Below the headline, the author is listed as 'Por Gabriel'. The main body of text describes the investigation and the consequences for Petrobras employees. At the bottom of the article, there are two rows of social interaction metrics: '0 Aprovações' with a green thumbs-up icon and '0 Desaprovações' with a red thumbs-down icon. Below these is a green bar with '0 Comentários' and a speech bubble icon. At the very bottom, there is a white text input field with the placeholder text 'Comentar'.

Figura 18 - Exemplo de uma notícia publicada, retirada do *site* Globo.com.

Utilizando o princípio da classificação (“*rating*”) de publicações, no sistema é possível classificar uma notícia através de uma classificação do tipo sim/não, assim os usuários poderão diferenciar uma notícia com conteúdo relevante de uma que não é muito importante. Na figura 17 é possível verificar que a notícia recebeu 50 (cinquenta) votos positivos e 12 (doze) negativos. Somente usuários cadastrados e que fizeram o *login* podem votar uma notícia.

4.6.3.1 Publicação e Categorização de Notícias

Para publicar uma notícia, o usuário deve preencher o formulário de cadastro que possui os seguintes campos: título, resumo, descrição, imagem, estado, cidade e categoria. O processo de publicação e categorização da notícia pode ser visualizado na figura 19. O usuário preenche os campos do formulário e, ao enviar os dados, o

controlador realiza o pré-processamento (mesmo processo descrito nas seções 2.2.3 e 4.3) do título, resumo e descrição. Uma nova instancia do categorizador Naive Bayes Multinomial é criada, este realiza o treinamento e categoriza a notícia baseado nos dados pré-processados. O controlador gera uma nova tela contendo o campo com as opções de cidades (baseadas no estado que o usuário escolheu) e categorias, sendo que a primeira opção é a categoria retornada pelo categorizador. O usuário pode aceitar a sugestão e clicar em cadastrar ou escolher outra opção. O controlador cria um novo objeto do tipo notícia que é salvo no banco de dados.

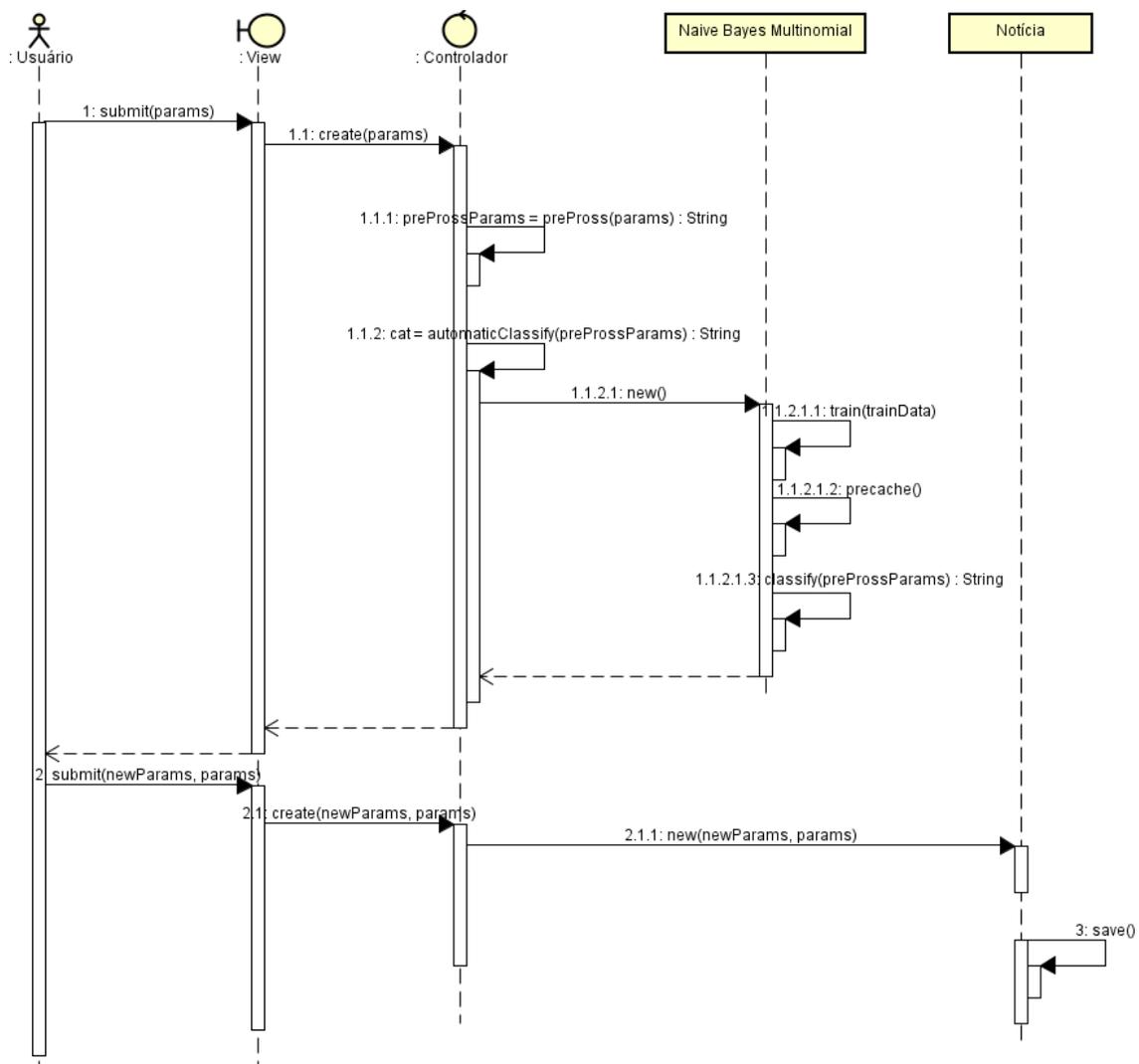
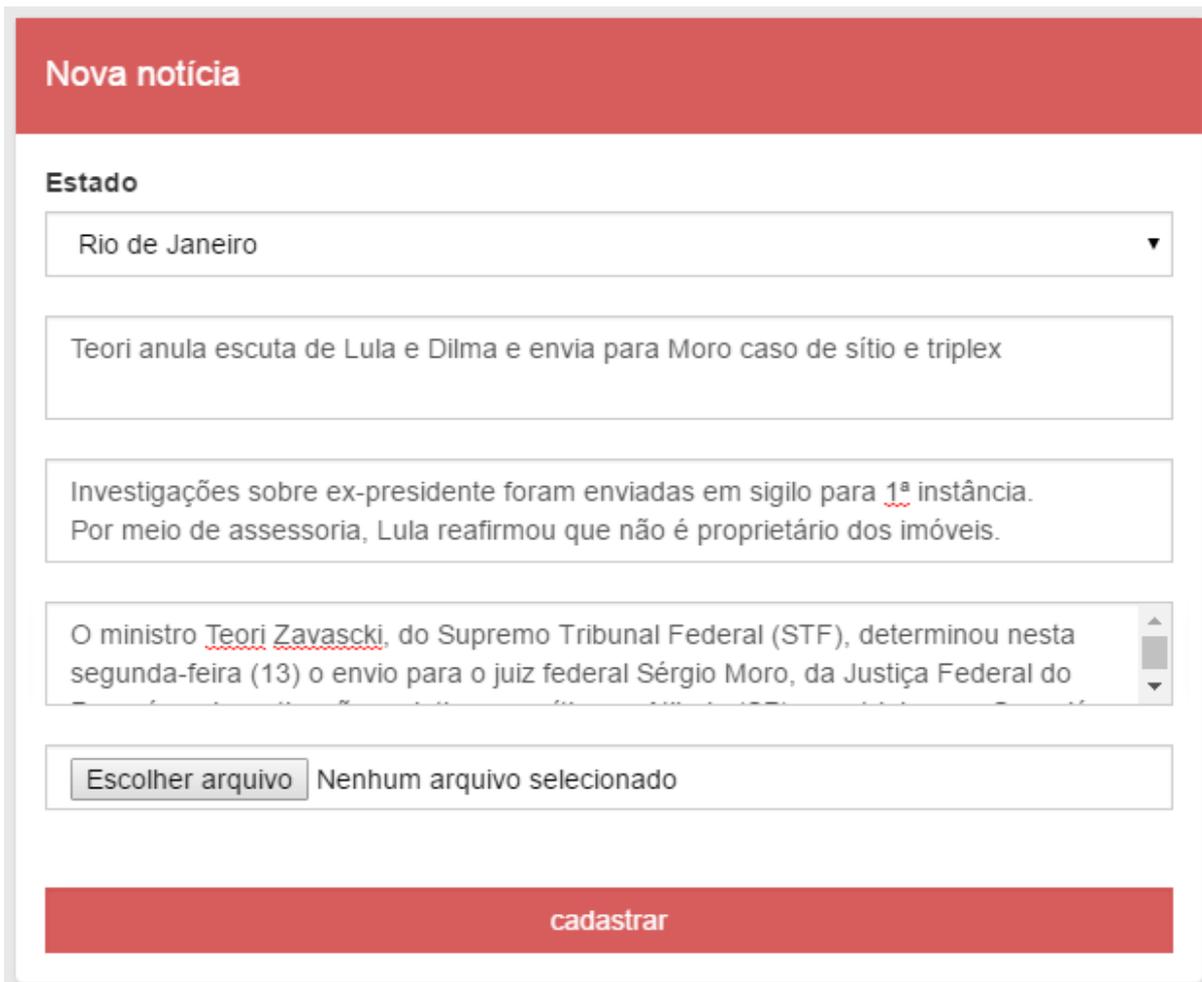


Figura 19 - Diagrama de sequência do processo de categorização e publicação das notícias.

Segundo Ferreira e Leite (2003), uma boa interface minimiza o número de ações necessárias para a entrada de dados, reduzindo a tarefa do usuário, e ao atribuir parte da tarefa de categorização ao sistema, o usuário acaba possuindo uma experiência mais

agradável pois, caso o categorizador automático tenha êxito, não será preciso escolher uma das opções.

As figuras 20 e 21 ilustram um teste realizado com uma notícia retirada do *site* Globo.com sobre política. Ao preencher os campos e avançar com o cadastro, o sistema apresenta outra tela com as opções de cidade e categoria, a categoria marcada foi “Política”, ou seja, o sistema sugeriu a categoria esperada.



The image shows a web form for registering a new news item. At the top, there is a red header with the text "Nova notícia". Below this, the form is organized into several sections:

- Estado:** A dropdown menu with "Rio de Janeiro" selected.
- Teori anula escuta de Lula e Dilma e envia para Moro caso de sítio e triplex:** A text input field containing a headline.
- Investigações sobre ex-presidente foram enviadas em sigilo para 1ª instância. Por meio de assessoria, Lula reafirmou que não é proprietário dos imóveis:** A text input field containing a sub-headline.
- O ministro Teori Zavascki, do Supremo Tribunal Federal (STF), determinou nesta segunda-feira (13) o envio para o juiz federal Sérgio Moro, da Justiça Federal do:** A text input field containing a lead sentence, with a scrollbar on the right side.
- Escolher arquivo Nenhum arquivo selecionado:** A button labeled "Escolher arquivo" and a text field showing "Nenhum arquivo selecionado".

At the bottom of the form, there is a prominent red button labeled "cadastrar".

Figura 20 - Formulário de cadastro de notícia com os campos preenchidos.

Nova notícia

Cidade

Angra dos Reis ▼

*** Categoria**

Política ▼

cadastrar

Figura 21 - Formulário com as opções de cidades e categorias.

5 Conclusões e Trabalhos Futuros

As redes sociais, os *blogs* e os *microblogs* não possuem funcionalidades típicas de *sites* de divulgação de notícias, tornando o compartilhamento e busca por informações em um processo difuso e algumas vezes restrito à alguns usuários. Os *sites* de jornais, por sua vez, não são abertos aos usuários, somente os jornalistas podem publicar as notícias. O sistema proposto fornece uma página de divulgação de notícias publicadas pelos próprios usuários, organizando os artigos em categorias, estados e cidades, tornando o compartilhamento de informações em algo mais público, democrático e organizado.

Com relação as técnicas de KDD e de mineração de dados, foi possível observar que os sistemas analisados no capítulo 3 as utilizam para categorizar suas notícias, alguns inclusive aplicando o categorizador Naive Bayes Multinomial. O sistema desenvolvido aplica essas técnicas para facilitar o processo de publicação e categorização das notícias, melhorando sua usabilidade.

Com o Weka, foi possível realizar os testes iniciais com os algoritmos e avaliar os resultados, para então escolher um dos categorizadores e aplicá-lo no sistema. Corroborando com os estudos apresentados na subseção 2.1.4.1, o categorizador Naive Bayes Multinomial obteve resultados mais satisfatórios do que o Naive Bayes (modelo Multivariado Bernoulli), J48 (árvore de decisão) e o PART (regra de decisão).

Para avaliar a precisão da ferramenta de categorização semiautomática utilizada pelo protótipo, devem ser realizados testes com usuários e então uma avaliação para verificar se é preciso voltar o processo de KDD para uma das etapas descritas no capítulo 2, ou não. Também deve ser feita uma análise de uma funcionalidade que permita atualizar a base de treinamento quando um usuário escolher uma categoria diferente da que foi sugerida pelo sistema. É preciso fazer um estudo sobre essa funcionalidade uma vez que sua implementação pode ser tanto útil como também ruim, caso em que o usuário escolhe uma categoria errada.

Além disso, o protótipo não garante que as publicações sejam confiáveis e que possuam conteúdos relevantes para os demais usuários, configurando assim uma limitação do sistema. Como trabalho futuro, para tentar contornar este problema, devem ser utilizadas técnicas e ferramentas de análise de sentimento e processamento de linguagem natural.

Outra melhoria que será aplicada no protótipo, e que também utiliza o processo e as técnicas de KDD e mineração de dados apresentadas, é a aplicação da categorização múltipla, por exemplo uma notícia pertencente a categoria “política” pode também pertencer a categoria “economia”. O usuário poderá decidir qual é a categoria principal além da possibilidade de escolher outras categorias.

Com relação ao desenvolvimento das interfaces e funcionalidades, foi possível avaliar que as ferramentas utilizadas são úteis para a implementação de sistemas voltados para a *web* social. Com o Sass foi possível tornar o desenvolvimento mais ágil uma vez que possui várias funções e facilidades que o CSS não possui. O Bootstrap também foi útil por possuir vários elementos prontos e com a sua *grid* foi possível tornar o sistema acessível para vários tamanhos de telas. O Ruby on Rails, com seu modelo MVC bem estruturado e suas extensões e *plugins*, reduziu o tempo de desenvolvimento.

Os padrões, práticas, princípios propostos por Bell (2009) e Crumlish e Malone (2009) foram essenciais na elaboração das funcionalidades do *site* com o intuito de torná-lo social e atrativo. Como continuação do trabalho de desenvolvimento do protótipo, além de melhorias na interface, serão implementadas outras funcionalidades, algumas delas inclui:

- Lista de notícias filtrada por popularidade: Ou seja, as notícias mais comentadas, compartilhadas e com mais classificações positivas terão posições mais altas na lista e também perderão relevância com o passar do tempo;
- Perfil de usuário: Uma vez que existam usuários em um *site*, é útil para eles ter a habilidade de identificar um do outro; uma forma comum de se fazer isso é disponibilizando uma página pessoal, um perfil (Bell, 2009). No *site* em questão, a página pessoal de um usuário será como o seu jornal pessoal, nela estarão as notícias publicadas pela pessoa e outras notícias que a pessoa “recortou e colou” no seu “jornal”. O usuário também poderá definir um nome para o seu “jornal” e este ficará exibido para os demais usuários;

- Sistema de busca e busca avançada: A busca costuma ser um dos primeiros lugares explorados por novos visitantes (Bell, 2009). Também estará disponível um sistema de busca avançada que pode ser muito útil, especialmente para conteúdos temporais (Bell, 2009). A busca avançada possuirá um campo para o usuário definir um limite de tempo facilitando assim a busca por conteúdos mais antigos. Além disso, existirão outros campos para filtrar a busca por estado, cidade e categoria;
- Compartilhamento de notícias: Segundo Crumlish e Malone (2009), no espaço virtual, os objetos são replicados e refletidos mais facilmente do que no mundo real. No *site* um usuário poderá adicionar uma notícia criada por outro usuário ao seu perfil e essa ficará exposta em sua página pessoal juntamente com as notícias criadas por ele próprio.

Referências Bibliográficas

BARBOSA, E. F. et al. (2004) “Inclusão das Tecnologias de Informação e Comunicação na Educação Através de Projetos”. Disponível em: <http://www.tecnologiadeprojetos.com.br/banco_objetos/%7BC36C8E12-B78C-4FFB-AB60-C428F2EBFD62%7D_inclus%C3%A3o%20das%20tecnologias.pdf>. Acesso em: 17 de jun. de 2016.

BELL, G. (2009) Building Social Web Applications. 1ª ed. Estados Unidos da América, O'Reilly Media, Inc., 438 p.

CAMARGO, Y. B. L. Abordagem Lingüística na Classificação de Textos em Português. 2007. 99 f. Dissertação (Mestrado) - Universidade Federal do Rio de Janeiro, COPPE, Rio de Janeiro. 2007.

CARVALHO, J. L. (2011) Twitter: Uma questão de confiabilidade em 140 caracteres. João Pessoa, UFPB, p. 19 – 20.

CETENFOLHA. Disponível em: <http://www.linguateca.pt/cetenfolha/index_info.html>. Acesso em: 15 de jun. de 2015.

CHHAJED, V. et al. (2015) Data Classification Algorithms. *International Journal of Pure and Applied Research in Engineering and Technology*, CS Dept, NMIMS University Shirpur, v. 3, p. 1762-1773. Disponível em: <<http://www.ijpret.com/publishedarticle/2015/4/IJPRET%20-%20CSIT%20287.pdf>>. Acesso em: 17 de jun. de 2016.

CORRÊA, R. F. Categorização de Documentos Utilizando Redes Neurais: Análise comparativa com técnicas não-conexionistas. 2002. 136 f. Dissertação (Mestrado) - Universidade Federal de Pernambuco, Centro de Informática, Recife. 2002.

CRUMLISH, C., MALONE, E. (2009) Designing Social Interfaces. 1ª ed. Canada, O'Reilly Media, Inc., 517 p.

DESAI, H. J. et al. (2014) “News Article Categorization”. Disponível em: <http://sifaka.cs.uiuc.edu/~wang296/Course/IR_Fall/docs/Projects/Samples/6.pdf>. Acesso em: 17 de jun. de 2016.

DIÁRIO DO RIO. #ProtestoRJ é capa no New York Times. Disponível em: <<http://diariodorio.com/wp-content/uploads/2013/06/NY-Times.jpg>>. Acesso em jun. 2015.

EMM NEWS EXPLORER. Disponível em: <<http://emm.newsexplorer.eu/NewsExplorer/home/en/latest.html>>. Acesso em: 17 de jun. de 2016.

FAYYAD, U. et al. (1996a) “The KDD Process for Extracting Useful Knowledge from Volumes of Data”. Disponível em: <<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.27.2315&rep=rep1&type=pdf>>. Acesso em: 17 de jun. de 2016.

FAYYAD, U. et al. (1996b) “From Data Mining to Knowledge Discovery in Databases”. Disponível em: <<https://www.aaai.org/ojs/index.php/aimagazine/article/viewFile/1230/1131>>. Acesso em: 17 de jun. de 2016.

FOGELMAN-SOULIÉ, F. et al. (2007) Mining Massive Data Sets for Security: Advances in Data Mining, Search, Social Networks and Text Mining and their Applications to Security. Amsterdam: IOS Press. v. 19, p. 336.

HAN, J., et al. (2011) Data Mining: Concepts and Techniques. 3ª ed. Estados Unidos da América, Morgan Kaufmann, p. 10.

IDRIS (2012). NaiveBayes: Multinomial Naive Bayes. Disponível em: <http://github.com/sld/naive_bayes>. Acesso em: 17 de jun. de 2016.

KANTARDZIC, M. (2011) Data Mining: Concepts, Models, Methods, and Algorithms. 2ª ed. Estados Unidos da América, John Wiley & Sons, Inc., p. 2 – 3.

KIBRIYA, A. M. et al. (2004) “Multinomial Naive Bayes for Text Categorization Revisited”. Disponível em: <http://www.cs.waikato.ac.nz/ml/publications/2004/kibriya_et_al_cr.pdf>. Acesso em:

FERREIRA, S. B. L., LEITE, J. C. S. P. (2003) “Avaliação da Usabilidade em Sistemas de Informação: O Caso do Sistema Submarino”. Revista de Administração Contemporânea - RAC. Publicação quadrimestral da ANPAD, v. 7, n. 2, p. 115 - 137.

LINOFF, G. S., BERRY, M. J. A. (2011) Data Mining Techniques: For Marketing, Sales, and Customer. 2ª ed. Estados Unidos da América, Wiley Publishing, Inc., 643 p.

MANNING, C. D. et al. (2009) An Introduction to Information Retrieval. Inglaterra, Cambridge University Press, p. 253 – 288.

MCCALLUM, A., NIGAM, K. A (1998) “Comparison of Event Models for Naive Bayes Text Classification”. Disponível em: <<http://www.cs.cmu.edu/~knigam/papers/multinomial-aaaiws98.pdf>>. Acesso em: 17 de jun. de 2016.

MCCALLUM, A. et al. (2000) “Automating the Construction of Internet Portals with Machine Learning”. Disponível em: <<http://www.kamalnigam.com/papers/corajnl.pdf>>. Acesso em: 17 de jun. de 2016.

MCCALLUM, A, PENG, F. (2004) “Accurate Information Extraction from Research Papers using Conditional Random Fields”. Disponível em: <<http://people.cs.umass.edu/~mccallum/papers/hlt2004.pdf>>. Acesso em: 17 de jun. de 2016.

MELO, L. B. S. Reconhecimento de Padrões Textuais para Categorização Automática de Documentos. 2007. 83 f. Dissertação (Mestrado) - Universidade Federal do Rio de Janeiro, COPPE, Rio de Janeiro. 2007.

MICHIE, D. et al. (1994) “Machine Learning, Neural and Statistical Classification”. Disponível em: <<https://www1.maths.leeds.ac.uk/~charles/statlog/whole.pdf>>. Acesso em: 19 de jun. 2016.

MITCHELL, T. M. (1997) Machine Learning. Estados Unidos da América, McGraw-Hill, p. 154 - 200.

MOHIUDDIN, U. et al. (2015) NEWS-D: A Realtime News Classification Engine for Web Streaming Data. *International Conference on Recent Advances in Computer Systems*, Arábia Saudita, p. 61, 1 dez. 2015.

OLIVEIRA, D. M. B. (2010) Reconhecimento de Padrões no Diagnóstico de Distúrbios Vocais de Docentes. 92 f. Dissertação (Mestrado em Ciências). Universidade Federal do Paraná, Curitiba. 2010. Disponível em: <<http://acervodigital.ufpr.br/bitstream/handle/1884/24908/DISSERTACAO%20DIVANETE.pdf>>. Acesso em: 17 de jun. de 2016.

PARDO, T. A. S., NUNES, M. G. V. (2002) “Aprendizado Bayesiano Aplicado ao Processamento de Línguas Naturais”. Disponível em: <<http://www.icmc.usp.br/~tasparado/NILCTR0225-PardoNunes.pdf>>. Acesso em:

PIMENTEL, M., FUKS, H. (2011) Sistemas Colaborativos. Rio de Janeiro, Elsevier, p. 3 – 15.

PRASS, F. S. (2012) “Uma Visão Geral Sobre as Fases do Knowledge Discovery in Databases (KDD)”. Disponível em: <<http://fp2.com.br/blog/index.php/2012/um-visao-geral-sobre-fases-kdd>>. Acesso em: 17 de jun. de 2016.

RASCHKA, S. (2014) “Naive Bayes and Text Classification – Introduction and Theory”. Disponível em: <<http://arxiv.org/pdf/1410.5329.pdf>>. Acesso em: 17 de jun. de 2016.

REFAEILZADEH, P. et al. (2008) “Cross-Validation”. Disponível em: <<http://leitang.net/papers/ency-cross-validation.pdf>>. Acesso em: 17 de jun. de 2016.

STEINER, M. T. A. et al. (2006) “Abordagem de um Problema Médico por meio do Processo de KDD com Ênfase à Análise exploratória dos Dados”. Disponível em: <<http://www.scielo.br/pdf/gp/v13n2/31177.pdf>>. Acesso em: 17 de jun. de 2016.

SOUZA, G. T. et al. (2003) “PATI-MVC: Padrões MVC para Sistemas de Informação”. Disponível em: <http://www.cin.ufpe.br/~sugarloafplop/final_articles/13_Pati-MVC.pdf>. Acesso em: 17 de jun. de 2016.

ZAKI, M. J., WAGNER, M. (2014) Data Mining and Analysis: Fundamental Concepts and Algorithms. Estados Unidos da América, Cambridge University Press, p. 1 – 30.

APÊNDICE A – Modelo Conceitual do Banco de Dados

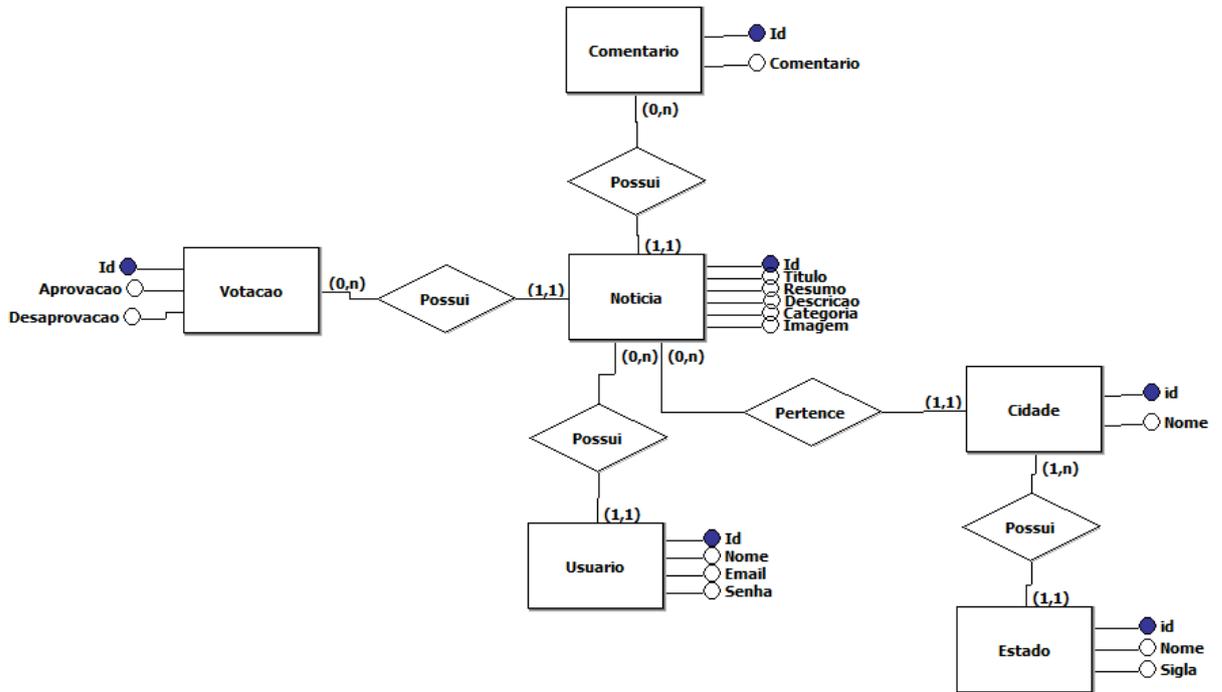


Figura 22 - Modelo conceitual do banco de dados.

APÊNDICE B – Modelo Lógico do Banco de Dados

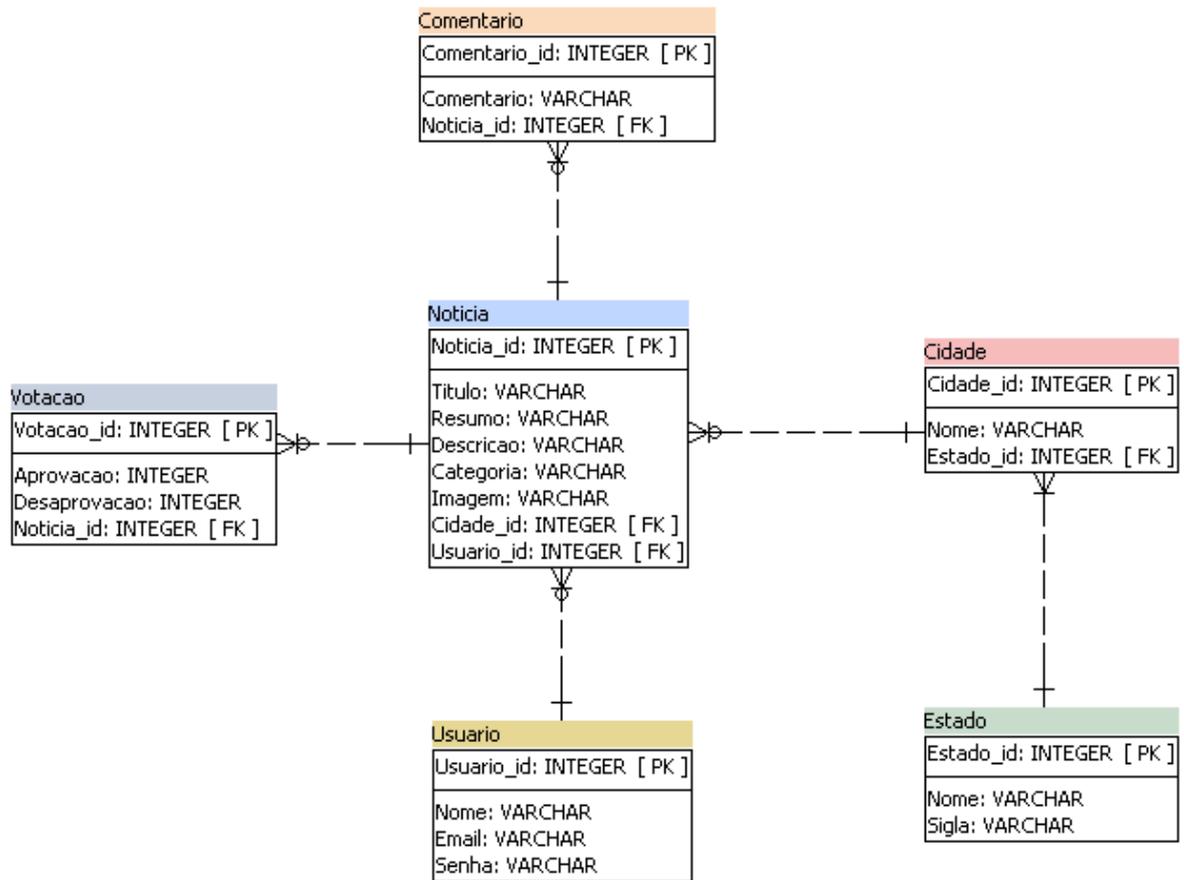


Figura 23 - Modelo lógico do banco de dados.

APÊNDICE C – Diagrama de Casos de Uso

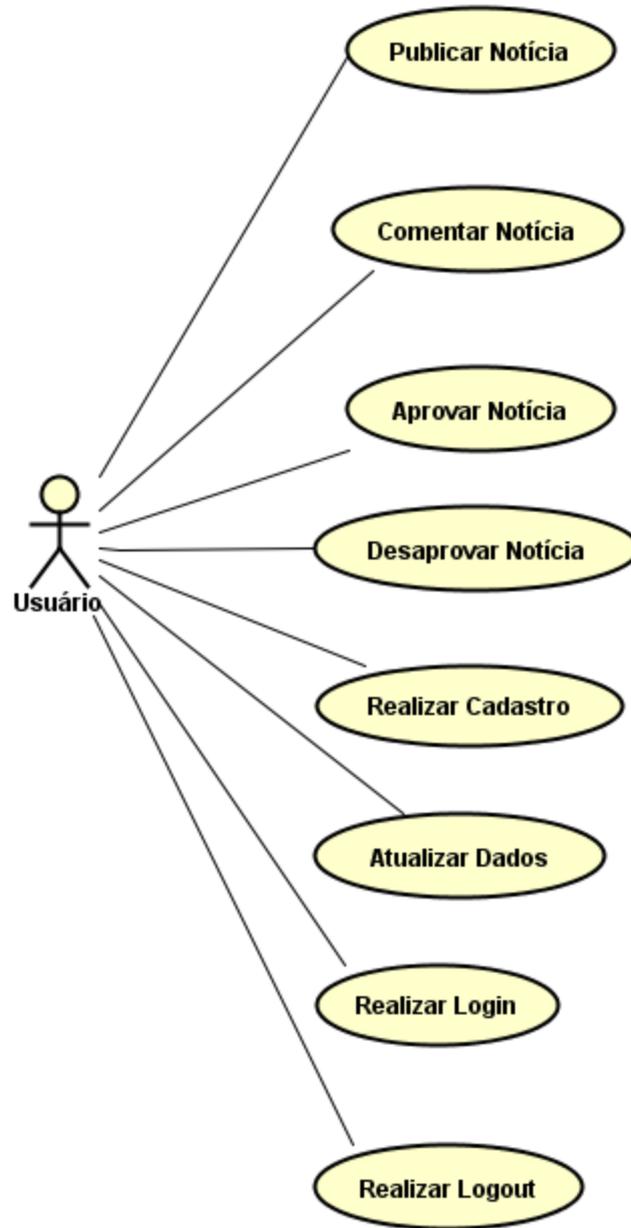


Figura 24 - Diagrama de casos de uso.

APÊNDICE D – Descrição dos Casos de Uso

| | |
|-------------------------------------|---|
| Caso de Uso | Realizar Cadastro |
| Objetivo | Usuário deseja realizar o seu cadastro no sistema |
| Ator | Usuário |
| Tipo | Secundário e essencial |
| Descrição | O usuário acessa a tela de cadastro do sistema e então preenche os campos do formulário. O sistema verifica se os dados estão corretamente preenchidos ou se já existe um usuário cadastrado com o endereço de e-mail. Se existirem problemas, o sistema retorna para a tela de cadastro e informa o erro ao usuário, do contrário o sistema realiza o cadastro do usuário. |
| Pré-Condições | N/A |
| Pós-Condições | Usuário cadastrado no sistema |
| Fluxo Principal de Eventos | <ol style="list-style-type: none"> 1. O sistema apresenta a tela de cadastro com os seguintes campos: nome de usuário, e-mail e senha. 2. O usuário preenche os campos e clica no botão “cadastro”. 3. O sistema cadastra o novo usuário. [F.A. 1] [F.A. 2] |
| Fluxo Alternativo de Eventos | <ol style="list-style-type: none"> 1. Caso o usuário não tenha preenchido um dos campos corretamente, o sistema retorna para a tela de cadastro e avisa o erro. 2. Caso já exista um usuário com o mesmo e-mail, o sistema retorna para a tela de cadastro e avisa o erro. |

Tabela 7 - Descrição do caso de uso "Realizar Cadastro".

| | |
|-------------------------------------|--|
| Caso de Uso | Atualizar Dados |
| Objetivo | Usuário deseja atualizar seus dados cadastros no sistema |
| Ator | Usuário |
| Tipo | Secundário e essencial |
| Descrição | O usuário acessa a tela de atualização e então preenche os campos do formulário. O sistema verifica se os dados estão corretamente preenchidos. Se existirem problemas, o sistema retorna para a tela de cadastro e informa o erro ao usuário, do contrário o sistema realiza a atualização. |
| Pré-Condições | Usuário logado no sistema |
| Pós-Condições | Dados do usuário atualizados no sistema |
| Fluxo Principal de Eventos | <ol style="list-style-type: none"> 1. O sistema apresenta a tela de atualização com os seguintes campos: nome de usuário e senha. 2. O usuário preenche os campos e clica no botão para realizar a atualização. 3. O sistema atualiza os dados do usuário. [F.A. 1] |
| Fluxo Alternativo de Eventos | <ol style="list-style-type: none"> 1. Caso o usuário não tenha preenchido um dos campos corretamente, o sistema retorna para a tela de atualização e avisa o erro. |

Tabela 8 - Descrição do caso de uso "Atualizar Dados".

| | |
|-------------------------------------|---|
| Caso de Uso | Realizar Login |
| Objetivo | Usuário deseja realizar o login no sistema |
| Ator | Usuário |
| Tipo | Secundário e essencial |
| Descrição | O usuário preenche os campos do formulário. O sistema verifica se os dados estão corretamente preenchidos. Se existirem problemas, o sistema retorna para a tela de login e informa o erro ao usuário, do contrário o sistema realiza o login do usuário. |
| Pré-Condições | Usuário cadastrado no sistema |
| Pós-Condições | Usuário logado no sistema |
| Fluxo Principal de Eventos | <ol style="list-style-type: none"> 1. O sistema apresenta a tela de login com os seguintes campos: e-mail e senha. 2. O usuário preenche os campos e clica no botão para realizar o login. 3. O sistema realiza o login do usuário. [F.A. 1] |
| Fluxo Alternativo de Eventos | <ol style="list-style-type: none"> 1. Caso o usuário não tenha preenchido um dos campos corretamente, o sistema retorna para a tela de login e avisa o erro. |

Tabela 9 - Descrição do caso de uso "Realizar Login".

| | |
|-------------------------------------|---|
| Caso de Uso | Realizar Logout |
| Objetivo | Usuário deseja realizar o logout no sistema |
| Ator | Usuário |
| Tipo | Secundário e essencial |
| Descrição | O usuário clica no botão para realizar o logout. O sistema realiza o logout do usuário. |
| Pré-Condições | Usuário logado no sistema |
| Pós-Condições | Usuário deslogado no sistema |
| Fluxo Principal de Eventos | <ol style="list-style-type: none"> 1. O usuário clica no botão para realizar o logout 2. O sistema realiza o logout do usuário. |
| Fluxo Alternativo de Eventos | N/A |

Tabela 10 - Descrição do caso de uso "Realizar Logout".

| | |
|-------------------------------------|---|
| Caso de Uso | Publicar Notícia |
| Objetivo | Usuário deseja publicar uma notícia |
| Ator | Usuário |
| Tipo | Primário e essencial |
| Descrição | O usuário acessa a página de cadastro de nova notícia e preenche os campos do formulário. O sistema verifica se os dados estão corretamente preenchidos. Se existirem problemas, o sistema retorna para a tela de cadastro e informa o erro ao usuário, do contrário o sistema realiza o cadastro da nova notícia. |
| Pré-Condições | Usuário logado no sistema |
| Pós-Condições | Notícia cadastrada no sistema |
| Fluxo Principal de Eventos | <ol style="list-style-type: none"> 1. O sistema apresenta a tela de cadastro de nova notícia. 2. O usuário preenche os campos e clica no botão de cadastrar. [F.A. 1] 3. O sistema realiza a categorização da notícia e apresenta a tela para o usuário confirmar a categoria ou escolher outra e para escolher a cidade. 4. O usuário mantém a sugestão dada pelo sistema ou seleciona outra categoria, seleciona a cidade e clica no botão de cadastrar. 5. O sistema realiza o cadastro da notícia. |
| Fluxo Alternativo de Eventos | <ol style="list-style-type: none"> 1. Caso o usuário não tenha preenchido um dos campos corretamente, o sistema retorna para a tela de cadastro e avisa o erro. |

Tabela 11 - Descrição do caso de uso "Publicar Notícia".

| | |
|-------------------------------------|---|
| Caso de Uso | Comentar Notícia |
| Objetivo | Usuário deseja comentar uma notícia |
| Ator | Usuário |
| Tipo | Primário e real |
| Descrição | O usuário acessa uma notícia publicada no sistema e preenche o campo de comentários e aperta o botão “Enter” no teclado. O sistema publica o comentário na notícia. |
| Pré-Condições | Usuário logado no sistema |
| Pós-Condições | Notícia comentada pelo usuário |
| Fluxo Principal de Eventos | <ol style="list-style-type: none"> 1. O sistema apresenta a tela contendo o conteúdo da notícia e com o campo de comentário. 2. O usuário preenche o campo e aperta o botão “Enter” no teclado. 3. O sistema publica o novo comentário na notícia. |
| Fluxo Alternativo de Eventos | N/A |

Tabela 12 - Descrição do caso de uso "Comentar Notícia".

| | |
|-------------------------------------|---|
| Caso de Uso | Aprovar Notícia |
| Objetivo | Usuário deseja aprovar uma notícia |
| Ator | Usuário |
| Tipo | Primário e essencial |
| Descrição | O usuário acessa uma notícia publicada no sistema e clica no botão “Aprovar”. O sistema registra a nova aprovação na notícia. |
| Pré-Condições | Usuário logado no sistema |
| Pós-Condições | Notícia aprovada pelo usuário |
| Fluxo Principal de Eventos | <ol style="list-style-type: none"> 1. O sistema apresenta a tela contendo o conteúdo da notícia e com o botão “Aprovar”. 2. O usuário clica no botão “Aprovar”. 3. O sistema registra a nova aprovação na notícia. |
| Fluxo Alternativo de Eventos | N/A |

Tabela 13 - Descrição do caso de uso "Aprovar Notícia".

| | |
|-------------------------------------|--|
| Caso de Uso | Desaprovar Notícia |
| Objetivo | Usuário deseja desaprovar uma notícia |
| Ator | Usuário |
| Tipo | Primário e essencial |
| Descrição | O usuário acessa uma notícia publicada no sistema e clica no botão “Desaprovar”. O sistema registra a nova desaprovação. |
| Pré-Condições | Usuário logado no sistema |
| Pós-Condições | Notícia desaprovaada pelo usuário |
| Fluxo Principal de Eventos | <ol style="list-style-type: none"> 1. O sistema apresenta a tela contendo o conteúdo da notícia e com o botão “Desaprovar”. 2. O usuário clica no botão “Desaprovar”. 3. O sistema registra a nova desaprovação na notícia. |
| Fluxo Alternativo de Eventos | N/A |

Tabela 14 - Descrição do caso de uso "Desaprovar Notícia".