A Text Analysis Approach for Cooking Recipe Classification Based on Brazilian Portuguese Documents

Larissa F. S. Britto[1], **Emilia G. Oliveira**[1], **Luciano D. S. Pacifico**[1], **Teresa B. Ludermir**[2]

[1]Departamento de Computação (DC) – Universidade Federal Rural de Pernambuco (UFRPE) – Recife – PE – Brazil

[2]Centro de Informática (CIn) – Universidade Federal de Pernambuco (UFPE) – Recife - PE - Brazil

{larissa.feliciana,emilia.galdino,luciano.pacifico}@ufrpe.br,

tbl@cin.ufpe.br

Abstract. *In this work, the cooking recipe classification problem is evaluated by means of the development of a new computational tool for Brazilian Portuguese text analysis. The proposed tool will be a fundamental piece to the development of more precise recipe recommendation systems for Brazilian people, as a manner to motivate such people to practice healthier eating habits. A new data set, obtained from Brazilian recipe websites, is proposed and tested by the use of classification algorithms from Machine Learning literature. Experiments have been performed towards the selection of the best classifiers to compose the recognition modules for the recipe recommendation systems to be developed as future works.*

Resumo. *Neste trabalho, a classificação de receitas culinárias é abordada através da elaboração de uma ferramenta computacional própria para a análise de documentos textuais escritos em Português. A ferramenta proposta será parte fundamental no desenvolvimento de sistemas de recomendação de receitas para os brasileiros, no intuito do incentivo à prática de hábitos alimentares saudáveis por essa população. Uma base de dados nova, obtida através de páginas web brasileiras, é elaborada e testada pelo uso de algoritmos obtidos da literatura de Aprendizagem de Máquina. Experimentos foram efetuados no intuito da seleção dos melhores classificadores para a composição dos módulos de reconhecimento dos sistemas de recomendação a serem desenvolvidos.*

1. Introdução

A escolha do alimento a ser consumido por um indivíduo é uma decisão importante e que envolve diversos fatores, como tempo disponível, local no qual deseja-se realizar a refeição, fatores sócio-culturais, disponibilidade dos alimentos, valor nutricional da comida, e gostos pessoais. Com o avanço da Internet, o acesso a uma grande variedade de alimentos e ingredientes vem tornando-se cada vez mais fácil, assim como a consulta a receitas culinárias através de sites especializados. Porém, mesmo com o fácil acesso proporcionado por esses sites, a busca por uma receita específica, desejada em um dado momento do dia, ainda é uma tarefa difícil, em decorrência da grande quantidade de dados armazenadas nesses repositórios. Com o objetivo de auxiliar os usuários na busca de

informações relevantes através de dados massivos, os Sistemas de Recomendação surgiram como uma subárea da Mineração de Dados.

Os Sistemas de Recomendação visam auxiliar seus usuários no processo de tomada de decisão em um determinado contexto, tendo tais sistemas também sido extendidos à área da recomendação de receitas culinárias [Gorbonos et al. 2018, Mokdara et al. 2018, Nezis et al. 2018, Nirmal et al. 2018]. Os Sistemas de Recomendação de Receitas (SRRs) são propostos como ferramentas para o auxílio aos usuários no encontro de dietas personalizadas e balanceadas nutricionalmente, assim como saborosas, de modo a promover melhores hábitos alimentares entre esses usuários, como uma forma de melhoria de suas saúdes, uma vez que maus hábitos alimentares estão diretamente relacionados ao desenvolvimento de várias doenças crônicas (como problemas cardíacos, diabetes e algumas formas de câncer) [Trattner and Elsweiler 2017]. A categorização automática de receitas tem sido empregada como peça fundamental na composição dos SRRs, servindo essa classificação como um filtro para a diminuição da quantidade de receitas retornadas por tais sistemas, assim como para a geração e recuperação de receitas que se adequem a um determinado conjunto de ingredientes já possuídos pelos usuários.

Sistemas de Recomendação de Receitas também podem ser vistos como ferramentas para a difusão cultural, uma vez que determinadas categorias de receitas apresentam elementos marcantes dos povos que as desenvolveram. Neste contexto, tendo como objetivo o oferecimento de maior acesso ao uso dos SRRs pela população brasileira, o presente trabalho terá como foco o desenvolvimento de uma base de dados de receitas obtida pela análise de documentos textuais em Português brasileiro, que servirá como ponto de partida para a elaboração de SRRs nesse idioma. Tendo em vista que a maioria dos SRRs existentes são propostos no idioma inglês, e dado que aproximadamente 95% da população brasileira não fala tal idioma, a recomendação de receitas em português será um elemento facilitador de integração social para que brasileiros tenham acesso aos benefícios oferecidos por esses sistemas.

Neste trabalho serão descritas as etapas realizadas para o desenvolvimento de uma base de dados de receitas em Português brasileiro. Para o desenvolvimento da base proposta, documentos textuais foram obtidos de forma automática, através da aplicação da técnica de *web scraping* a páginas web brasileiras. Para a análise das receitas em português, uma ferramenta própria de análise de documentos textuais foi desenvolvida. Com o intuito de promover uma avaliação inicial à base de dados proposta, uma abordagem de Classificação de Receitas foi realizada, através do uso de cinco classificadores de propósito geral, obtidos da literatura de Aprendizagem de Máquina. A avaliação experimental adotada está em conformidade ao estado da arte de classificação de receitas [Su et al. 2014, Jayaraman et al. 2017, Kalajdziski et al. 2018, Nirmal et al. 2018], e também servirá como fator determinante na escolha do classificador a ser adotado em etapas futuras do projeto, nas quais os SRRs serão desenvolvidos e disponibilizados ao público brasileiro, em conformidade com os objetivos da pesquisa sendo realizada.

O trabalho está organizado como segue. Uma breve revisão do estado da arte em classificação de receitas será oferecida na Seção 2. Logo após, a metodologia adotada para o desenvolvimento da base de dados será apresentada (Seção 3), seguida pela análise experimental (Seção 4). Por fim, as conclusões e algumas tendências para pesquisas futu-

ras serão apresentadas (Seção 5).

2. Trabalhos Relacionados

O problema de classificação de receitas culinárias pode ser facilmente mapeado em um problema tradicional de classificação, mais especificamente, em um problema de classificação de documentos textuais em múltiplas categorias (*Multi-Class Document Classification*), onde temos as seguintes relações:

1. Padrão = Documento = Receita;
2. Característica = Palavra = Ingrediente;
3. Classe = Classe do Documento = Categoria da Receita.

Nessa abordagem, a lista total de ingredientes (ou seja, o conjunto formado pela união de todos os ingredientes da base de receitas) é considerado o conjunto final de características do problema [Su et al. 2014, Jayaraman et al. 2017, Kalajdziski et al. 2018].

Uma vez que um problema de classificação de receitas é mapeado como um problema de classificação comum, podemos resolvê-lo através da execução de três etapas básicas: aquisição da base de dados; pré-processamento e extração de características (*representação*); e a etapa de classificação (*reconhecimento*).

A aquisição da base de dados geralmente é realizada pela recuperação automática de documentos através da aplicação da técnica de *web scraping* a páginas web especializadas em culinária e gastronomia [Su et al. 2014, Ooi et al. 2015, Mokdara et al. 2018, Nezis et al. 2018, Nirmal et al. 2018]. Também é comum o uso de bases de dados presentes em repositórios de dados públicos [Kalajdziski et al. 2018].

A etapa de pré-processamento é executada visando a correção e padronização dos documentos (*receitas*) antes da etapa de classificação. Alguns métodos de pré-processamento para documentos textuais comumente empregados no contexto de classificação de receitas são: transformação das palavras (por exemplo, conversão para letras minúsculas), correção de erros na representação das palavras (por exemplo, correção de erros de digitação), remoção de caracteres especiais, remoção de pontuação e acentuação, remoção de valores numéricos, conversão dos documentos textuais para a Matriz de Termos dos Documentos (*Document-Term Matrix*, ou DTM), etc [Jayaraman et al. 2017, Kalajdziski et al. 2018].

Como módulo de reconhecimento, algoritmos de Aprendizagem de Máquina supervisionada (ou seja, *classificadores*) são adotados. Dentre os principais classificadores empregados em classificação de receitas, podemos citar o classificador Naïve Bayes [Jayaraman et al. 2017, Kalajdziski et al. 2018], a Classificação Associativa [Su et al. 2014], a Regressão Logística Multinominal [Jayaraman et al. 2017], as Máquinas de Vetores de Suporte [Jayaraman et al. 2017, Kalajdziski et al. 2018, Su et al. 2014], o algoritmo de Floresta Aleatória [Jayaraman et al. 2017, Nirmal et al. 2018], as Redes Neurais Artificiais [Kalajdziski et al. 2018] e as Redes Neurais Profundas [Mokdara et al. 2018, Nezis et al. 2018].

Alguns trabalhos na área de classificação de receitas culinárias são discutidos brevemente na sequência.

Em [Su et al. 2014], a correlação entre categorias de cozinha e suas listas de ingredientes é avaliada, em uma tentativa de mapear as conexões entre ingredientes e cozinhas. A Classificação Associativa e as Máquinas de Vetores de Suporte são empregadas como módulos de reconhecimento, e os autores tentam explicar as correlações entre diferentes categorias culinárias por meio da identificação dos ingredientes presentes em cada uma dessas categorias e da matriz de confusão gerada pelos classificadores.

Em [Jayaraman et al. 2017], os autores também ofereceram uma análise da correlação entre categorias culinárias e listas de ingredientes pertencentes às receitas nessas categorias. A metodologia proposta avaliou a performance de quatro algoritmos de classificação (Naïve Bayes, Regressão Logística Multinominal, Floresta Aleatória e uma Máquina de Vetores de Suporte Linear) quando empregados no problema de classificação de receitas. Os autores usaram diferentes técnicas de pré-processamento e extração de características de documentos textuais para padronizar a base de dados adotada na análise experimental realizada.

Em [Nirmal et al. 2018], um sistema de recomendação e geração de receitas é proposto, baseado na otimização tanto dos sabores dos ingredientes, quanto de seus valores nutricionais. Um algoritmo de Floresta Aleatória é usado para a classificação das receitas em onze classes culinárias, relacionadas às cozinhas de diferentes países, sendo essa classificação usada como um primeiro filtro do sistema proposto.

Em [Kalajdziski et al. 2018], também foram empregadas técnicas bem estabelecidas de pré-processamento, extração e seleção de características da área de análise de textos, como os métodos de *Bag of Words* e TF-IDF, para a realização da tarefa de classificação automática de textos. Os classificadores Naïve Bayes, uma Rede Neural Artificial e uma Máquina de Vetores de Suporte foram adotados como módulos de reconhecimento, e os melhores conjuntos de características extraídos foram empregados para a composição do sistema final proposto.

3. Metodologia

Nesta seção, as etapas adotadas para o desenvolvimento da base de dados proposta serão apresentadas em detalhes. As principais etapas da metodologia adotada estão representadas na Fig. 1.

3.1. Aquisição dos Documentos

A etapa de aquisição dos documentos foi realizada de forma automática, através da aplicação da técnica de *web scraping* a páginas web brasileiras especializadas em receitas culinárias. O presente trabalho fez a opção de uso de documentos apenas em língua portuguesa com o intuito de popularizar o acesso às ferramentas de recomendação de receitas e dietas pelo público brasileiro, como forma de auxílio à prática de hábitos alimentares saudáveis por essa população. No total, 4448 documentos foram obtidos. Após a etapa de aquisição dos documentos, os processos de pré-processamento e extração de características foram executados.

3.2. Pré-Processamento dos Documentos

A grande maioria das páginas web de receitas brasileiras é construída cooperativamente por sua comunidade de usuários (ou seja, qualquer usuário pode realizar a submissão de

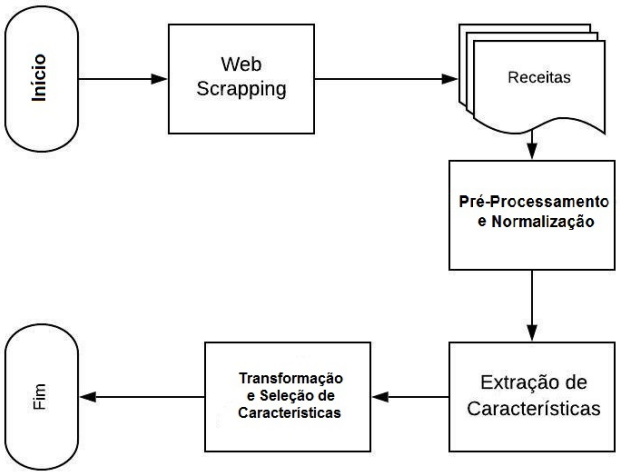

Figura 1. Etapas da Metodologia Adotada para a Geração da Base de Dados Proposta.

receitas). Essa abordagem tem como vantagens principais tornar maior o engajamento de seus usuários à ferramenta, e, do ponto de vista de conteúdo, tornar maior e mais diversificada a quantidade de receitas disponíveis nesses repositórios. Porém, como os usuários desses sitemas tendem a escrever de forma livre (ou seja, não seguem a norma culta do idioma), a qualidade final dos documentos submetidos tende a ser baixa, do ponto de vista do formalismo da linguagem [Yasukawa and Scholer 2017]. Com o intuito de corrigir eventuais problemas ocasionados por essa falta de formalismo, técnicas de pré-processamento e normalização são aplicadas aos documentos. Em decorrência da existência de poucas ferramentas para a análise automática de documentos escritos em Português Brasileiro, e dado que as ferramentas existentes para esse idioma são bastante limitadas e passíveis a erros [de Souza et al. 2018], optamos pelo desenvolvimento de uma ferramenta própria para a análise das receitas obtidas.

A primeira etapa do pré-processamento dos documentos aquiridos foi a remoção de documentos redundantes, ou seja, a remoção de documentos idênticos recuperados automaticamente pelo *web scraper*. Como as páginas web brasileiras estão geralmente dividias em categorias próprias (não há padronização na nomenclatura das categorias em diferentes páginas web), uma mesma receita pode estar associada a categorias diferentes, assim como também é comum que uma mesma receita seja referenciada em mais de uma página da mesma categoria, o que faz com que o *web scraper* recupere o mesmo documento em mais de uma ocasião. Desta forma, a remoção de redundâncias foi executada pela comparação direta entre os links dos documentos recuperados. Como o intuito do presente trabalho é a análise de documentos apenas em Português brasileiro, documentos escritos em outros idiomas também foram removidos nesta etapa. De modo semelhante, optou-se pela exclusão de documentos que continham mais de uma receita, o que resultou na elaboração de uma base de dados contendo 3106 documentos (*receitas*) finais.

Após a obtenção dos documentos, os principais problemas encontrados nas receitas foram: uso de abreviações, erros de digitação, ingredientes que possuem mais de uma nomenclatura, modos de preparo completos listados como ingredientes, e receitas com

preparos completos listados como um ingrediente (como por exemplo, *massa de pastel* e *biscoito de chocolate* sendo listados como ingredientes, e não como outros tipos de preparo - ou seja, outras receitas). Visando a minimização desses problemas, os seguintes passos foram realizados:

1. Conversão das palavras para letras minúsculas;
2. Separação da lista dos ingredientes (como listado em cada documento) do modo de preparo;
3. Expansão de abreviações;
4. Correções ortográficas;

Como o interesse do presente trabalho é na análise da lista de ingredientes que compõem cada receita e suas relações com as categorias culinárias existentes, outros elementos contidos nas receitas (como quantidades dos ingredientes, modos de preparo, lista de instrumentos necessários, etc.) serão desconsiderados, de modo que cada receita final será representada apenas por sua lista de ingredientes [Jayaraman et al. 2017, Su et al. 2014, Kalajdziski et al. 2018]. Para isso, dois léxicos foram criados, sendo um deles para a identificação de expressões relacionadas a medidas (por exemplo, *gramas*, *colher de sopa*, *xícara*), e o segundo para indicações relacionadas ao estado do ingrediente (por exemplo, *picado*, *morno*, *cozido*). Além disso, informações numéricas foram removidas.

Alguns ingredientes possuem ainda diversas nomenclaturas, fazendo com que muitas vezes um mesmo ingrediente seja visto pela ferramenta de análise dos textos desenvolvida como ingredientes diferentes, (por exemplo, *carne de boi* e *carne bovina*). A normalização foi empregada para que esses ingredientes sejam escritos de forma padronizada. Outros problemas, como ingredientes no plural, aumentativo ou diminuitivo, também foram corrigidos pela normalização.

Após o pré-processamento e normalização das receitas, cada receita será representada por sua lista de ingredientes ainda em formato textual. Uma lista completa formada por todos os ingredientes da base de receitas é obtida, com um total de 1312 ingredientes diferentes.

3.3. Base de Dados Proposta

Com a obtenção da lista total de ingredientes, as receitas, que após o pré-processamento e a normalização são representados por uma lista textual de ingredientes (ou seja, palavras), é convertida para uma matriz de dados numéricos (a DTM). Como para o trabalho proposto o único interesse é saber se um dado ingrediente está ou não contido em uma receita, a DTM será codificada como uma matriz *binária*. Em seguida, a técnica de TF-IDF [Hamoud et al. 2018] é aplicada à DTM para a obtenção da relevância de cada ingrediente em cada receita [Jayaraman et al. 2017], sendo a base de dados final constituída pelas frequências do TF-IDF.

Após a análise das categorias de receitas propostas pelas páginas web das quais as receitas foram extraídas, chegou-se a categorização da base de dados em 7 classes distintas (sem sobreposição), listadas na Tabela 1.

4. Análise Experimental

Nesta seção, a base de dados proposta é testada através do uso de cinco classificadores de propósito geral, obtidos da literatura de Aprendizagem de Máquina: o Classificador

Tabela 1. Distribuição das Classes para a Base de Dados Proposta.

Rótulo da Classe	Nome da Categoria	Número de Receitas
1	Alimentação Saudável	111
2	Acompanhamentos	620
3	Carnes e Omeletes	1107
4	Doces e Sobremesas	808
5	Entradas e Salgados	199
6	Massas e Risotos	244
7	Molhos	17

de Regressão Logística (*Logistic Regression Classifier*, ou LRC) [Jayaraman et al. 2017], uma Rede Neural Artificial do tipo Perceptron de Múltiplas Camadas (MLP) treinada com o algoritmo *Backpropagation* [Haykin 2001, Rumelhart et al. 1985], o Classificador Naïve Bayes (NB) [De Stefano et al. 2012], o Algoritmo de Floresta Aleatória (*Random Forest*, ou RFC) [Criminisi et al. 2011] e uma Máquina de Vetores de Suporte (*Support Vector Machine*, ou SVM) [Haykin 2001] com função de *kernel* Linear. Todos os algoritmos foram implementados na linguagem de programação Python, através do uso da biblioteca *scikit-learn* [Pedregosa et al. 2011, Buitinck et al. 2013], sendo executados com as configurações *default* dessa biblioteca (com exceção do RFC, que fez uso de 50 estimadores, e da MLP, que fez uso de 1000 épocas de treinamento). Os testes foram realizados em um computador com uma CPU i7-7700K, com uma GPU NVIDIA GeForce GTX 1060 de 6 GB, e 32 GB de memória RAM.

Os experimentos foram conduzidos através de um *framework* do tipo validação cruzada 10-*fold*: a base de dados proposta é dividida aleatoriamente em dez partes (sem sobreposição entre essas partes), e em cada uma das dez etapas da repetição do experimento, uma dessas partes é usada como *conjunto de teste*, enquanto as outras nove partes são usadas como *conjunto de treinamento* dos modelos. Visando a obtenção de um conjunto maior de amostras, o processo de validação cruzada 10-*fold* foi repetido 10 vezes, sendo em cada uma dessas vezes os dados redestribuídos aleatoriamente para a formação das *folds*. O objetivo dessa reamostragem é evitar resultados obtidos por sorte pelos classificadores, promovendo uma análise mais justa dos experimentos.

Como métricas de avaliação dos experimentos, quatro índices comumente aplicados a problemas de classificação foram adotados: a Acurácia, a Precisão (*Precision*), a Revocação (*Recall*) e a F-*Measure*. A avaliação inclui uma análise empírica dos resultados obtidos para o conjunto de teste dos experimentos, assim como uma análise do tempo médio de execução de cada um dos algoritmos. Com o objetivo de complementar a análise empírica, a avaliação levará em consideração ainda um sistema de *ranks* obtidos através de testes de hipóteses pareados do tipo t-teste, com um grau de confiança $\alpha = 0.05$, em relação a cada uma das métricas escolhidas. O sistema de *ranks* é implementado de modo que sempre que um algoritmo for considerado estatisticamente superior (de acordo com o t-teste pareado) a algum outro algoritmo, o mesmo obterá 1 (um) ponto no *rank*, enquanto sempre que tal algoritmo for considerado com desempenho inferior a algum outro algoritmo, ao mesmo será atribuído um ponto negativo (-1). Algoritmos considerados estatisticamente equivalentes em relação a uma determinada métrica recebem pontuação 0 (zero) na comparação. O *rank* final do algoritmo considerando uma métrica será obtido pela soma total de seus *ranks* em relação a todos os demais algoritmos.

Tabela 2. Resultados Experimentais: Média e Desvio Padrão (Std) para cada métrica.

Métrica	LRC		MLP		NB		RFC		SVM	
	Média	Std	Média	Std	Média	Std	Média	Std	Média	Std
Acurácia	0.700	0.022	0.677	0.024	0.643	0.023	0.695	0.023	**0.711**	**0.023**
F-$Measure$	0.659	0.027	0.675	0.025	0.562	0.030	0.656	0.027	**0.688**	**0.026**
Precisão	0.689	0.029	0.679	0.025	0.611	0.054	0.682	0.030	**0.702**	**0.027**
Revocação	0.700	0.022	0.677	0.024	0.643	0.023	0.695	0.023	**0.711**	**0.023**
Tempo	0.370	0.067	204.6	14.1	**0.129**	**0.084**	0.419	0.052	8.094	0.054

Figura 2. Lista de Ingredientes Mais Frequentes na Base de Receitas: (a) Lista Global (Ingredientes Mais Frequentes na Base de Dados como um Todo); (b) Ingredientes com Maiores Frequência em Múltiplas Classes.

Os resultados para o conjunto total de ingredientes (1312) são apresentados na Tabela 2. Em uma análise empírica, podemos observar que os melhores modelos obtiveram uma Acurácia média acima de 70%. Tendo em vista que a base proposta apresenta um alto grau de desbalanceamento entre as classes (por exemplo, a classe *Carnes e Omeletes* possui 1107 receitas, enquanto a classe *Molhos* possui apenas 17 receitas), assim como avaliando a prevalência de certos ingredientes em mais de uma das categorias de receitas elaboradas (por exemplo, os ingredientes *sal*, *cebola*, *azeite de oliva* e *alho*, aparecem com alta frequência em muitas das classes da base de dados - vide Fig. 2), podemos considerar que os resultados obtidos foram bastante promissores.

Em uma análise empírica, os melhores resultados obtidos em termos da Acurácia no conjunto de testes foram encontrados pelo SVM, LRC e RFC, respectivamente. De acordo com os t-testes pareados (Tabela 3), houve diferença estatística significativa em relação aos resultados obtidos pelo SVM e os resultados obtidos pelos demais modelos. Embora o NB tenha obtido o melhor tempo de execução médio dentre os algoritmos testados (tendo em vista sua simplicidade), seus resultados médios em relação às métricas de classificação adotadas foram os piores.

A Fig. 3 apresenta a variação no comportamento dos classificadores quando apenas um conjunto dos ingredientes mais frequentes é utilizado. De modo geral, os classificadores apresentaram acurácias significativamente menores quando o conjunto de ingredientes mais frequentes era formado por menos do que 400 ingredientes, atingindo desempenhos mais estáveis quando ao menos 500 dos ingredientes mais frequentes eram

Tabela 3. t-testes Pareados: t-$value$ e p-$value$ para cada métrica. † significa que os valores das amostras apresentaram diferenças estatísticas significativas. R: valor do *rank* final obtido pelo algoritmo para a métrica em avaliação.

Acurácia											
Alg.	LRC		MLP		NB		RFC		SVM		R
	t-$val.$	p-$val.$	t-$val.$	p-$val.$	t-$val.$	p-$val.$	t-$val.$	p-$val.$	t-$val.$	p-$val.$	
LRC	N/A	N/A	7.042†	0.0	17.57†	0.0	1.362	0.175	-3.467†	0.0002	1
MLP	-7.042†	0.0	N/A	N/A	10.01†	0.0	-5.540†	0.0	-10.55†	0.0	-2
NB	-17.57†	0.0	-10.01†	0.0	N/A	N/A	-15.75†	0.0	-21.00	0.0	-4
RFC	-1.362	0.175	5.540†	0.0	15.75†	0.0	N/A	N/A	-4.985†	0.0	1
SVM	3.749†	0.0002	10.55†	0.0	21.00†	0.0	4.985†	0.0	N/A	N/A	4
F-$Measure$											
Alg.	LRC		MLP		NB		RFC		SVM		R
	t-$val.$	p-$val.$	t-$val.$	p-$val.$	t-$val.$	p-$val.$	t-$val.$	p-$val.$	t-$val.$	p-$val.$	
LRC	N/A	N/A	-4.496†	0.0	24.40†	0.0	0.788	0.432	-7.651	0.0	-1
MLP	4.496†	0.0	N/A	N/A	29.65†	0.0	5.274†	0.0	-3.463†	0.0007	2
NB	-24.40†	0.0	–29.65†	0.0	N/A	N/A	-23.45†	0.0	-31.86†	0.0	-4
RFC	-0.788	0.432	-5.274†	0.0	23.45†	0.0	N/A	N/A	-8.375†	0.0	-1
SVM	7.651†	0.0	3.463†	0.0007	31.86†	0.0	8.375†	0.0	N/A	N/A	4
Precisão											
Alg.	LRC		MLP		NB		RFC		SVM		R
	t-$val.$	p-$val.$	t-$val.$	p-$val.$	t-$val.$	p-$val.$	t-$val.$	p-$val.$	t-$val.$	p-$val.$	
LRC	N/A	N/A	2.590†	0.103	12.67†	0.0	1.577	0.116	-3.347†	0.001	1
MLP	-2.590†	0.103	N/A	N/A	11.35†	0.0	-0.883	0.378	-6.292†	0.0	-1
NB	-12.67†	0.0	-11.35†	0.0	N/A	N/A	-11.55†	0.0	-15.10†	0.0	-4
RFC	-1.577	0.116	0.883	0.378	11.55†	0.0	N/A	N/A	-4.950†	0.0	0
SVM	3.347†	0.001	6.292†	0.0	15.10†	0.0	4.950†	0.0	N/A	N/A	4
Revocação											
Alg.	LRC		MLP		NB		RFC		SVM		R
	t-$val.$	p-$val.$	t-$val.$	p-$val.$	t-$val.$	p-$val.$	t-$val.$	p-$val.$	t-$val.$	p-$val.$	
LRC	N/A	N/A	7.045†	0.0	17.57†	0.0	1.362	0.175	-3.749†	0.0002	1
MLP	-7.042†	0.0	N/A	N/A	10.01†	0.0	-5.540†	0.0	-10.55†	0.0	-2
NB	-17.57†	0.0	-10.01	0.0	N/A	N/A	-15.75†	0.0	-21.00	0.0	-4
RFC	-1.362	0.175	5.540†	0.0	15.75†	0.0	N/A	N/A	-4.985†	0.0	1
SVM	3.749†	0.0002	10.55†	0.0	21.00†	0.0	4.985†	0.0	N/A	N/A	4

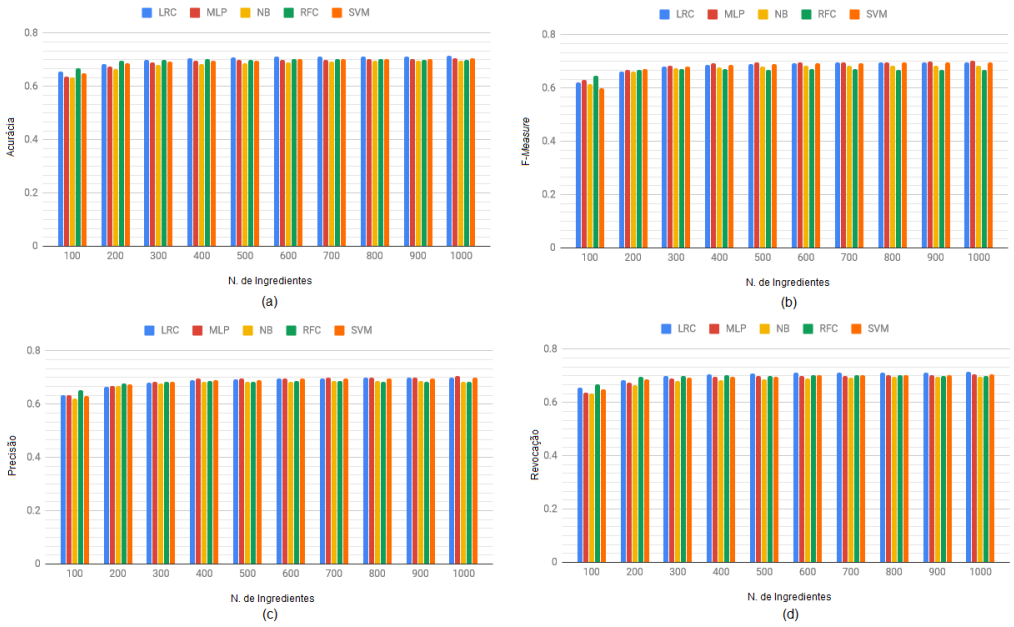

Figura 3. Variação dos Resultados Médios para Um Conjunto de Ingredientes Mais Frequentes: (a) Acurácia, (b) F-$Measure$, (c) Precisão, (d) Revocação.

utilizados. Mais uma vez, esse fato pode ser atribuído à existência de ingredientes que possuem alta frequência em múltiplas categorias do problema, o que torna a fronteira de decisão entre as classes mais difusa quando poucos ingredientes são usados. Os classificadores que apresentaram maior instabilidade com a variação no número de ingredientes considerados foram o NB e a MLP, enquanto os classificadores mais estáveis nessa análise foram o RFC e o SVM.

5. Conclusões

Neste trabalho, uma nova base de dados é proposta para o problema de classificação automática de receitas culinárias, baseada na análise de documentos escritos em Português brasileiro. A base de dados proposta é composta de 3106 receitas, formadas por um total 1312 ingredientes diferentes, e divididas em 7 classes que apresentam um alto grau de desbalanceamento entre si. A base de dados proposta servirá como base para o desenvolvimento de sistemas de recomendação de receitas em língua portuguesa, de modo a facilitar o uso de tais sistemas por brasileiros.

Como forma de análise da base de dados proposta, cinco algoritmos obtidos através da literatura de Aprendizagem de Máquina foram testados, no intuito da seleção do melhor modelo para compor o módulo de classificação dos sistemas de recomendação de receitas a serem desenvolvidos. Os testes foram realizados como o auxílio de quatro métricas aplicadas comumente a problemas de classificação (acurácia, f-$measure$, precisão e revocação), assim como de uma análise estatística obtida pela aplicação de testes de hipóteses do tipo t-testes pareados. Os resultados experimentais indicaram que as técnicas de Máquinas de Vetores de Suporte (SVM), Classificador de Regressão Logística (LRC) e Algoritmos de Florestas Aleatórias (RFC) foram capazes de obter, nessa ordem, os melhores desempenhos médios para a base de dados proposta.

Como trabalhos futuros, pretendemos aprofundar a análise da base proposta

através de um estudo mais detalhado da prevalência dos ingredientes em cada categoria de receitas, de modo a tentar determinar qual o conjunto mínimo de ingredientes que seriam capazes de caracterizar corretamente uma determinada classe culinária. Tal análise é importante para que um sistema de recomendação de receitas seja capaz de, dada uma lista de ingredientes fornecida como entrada pelo usuário, inferir qual a categoria culinária mais próxima à essa lista, fazendo sugestões de receitas mais adequadas nessa categoria. Também pretendemos expandir a base de dados apresentada pelo acréscimo de novas receitas, a serem adquiridas pelo uso de métodos automáticos (*web scraping*) de páginas web brasileiras especializadas em culinária. Por fim, os sistemas de recomendação a serem desenvolvidos serão disponibilizados ao público brasileiro, como ferramentas de auxílio à elaboração de dietas saudáveis e nutritivas.

Agradecimentos

Os autores gostariam de agradecer ao CNPq e a CAPES pelo suporte financeiro.

Referências

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., and Varoquaux, G. (2013). API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122.

Criminisi, A., Shotton, J., and Konukoglu, E. (2011). Decision forests for classification, regression, density estimation, manifold learning and semi-supervised learning [internet]. *Microsoft Research*.

de Souza, J. G. R., de Paiva Oliveira, A., and Moreira, A. (2018). Development of a brazilian portuguese hotel's reviews corpus. In *International Conference on Computational Processing of the Portuguese Language*, pages 353–361. Springer.

De Stefano, C., Fontanella, F., and Di Freca, A. S. (2012). A novel naive bayes voting strategy for combining classifiers. In *Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on*, pages 467–472. IEEE.

Gorbonos, E., Liu, Y., and Hoàng, C. T. (2018). Nutrec: Nutrition oriented online recipe recommender. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 25–32. IEEE.

Hamoud, A. A., Alwehaibi, A., Roy, K., and Bikdash, M. (2018). Classifying political tweets using naïve bayes and support vector machines. In *Recent Trends and Future Technology in Applied Intelligence*, pages "736–744". Springer International Publishing.

Haykin, S. S. (2001). *Neural networks: a comprehensive foundation*. Tsinghua University Press.

Jayaraman, S., Choudhury, T., and Kumar, P. (2017). Analysis of classification models based on cuisine prediction using machine learning. In *2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon)*, pages 1485–1490. IEEE.

Kalajdziski, S., Radevski, G., Ivanoska, I., Trivodaliev, K., and Stojkoska, B. R. (2018). Cuisine classification using recipe's ingredients. In *2018 41st International Conven-*

tion on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pages 1074–1079. IEEE.

Mokdara, T., Pusawiro, P., and Harnsomburana, J. (2018). Personalized food recommendation using deep neural network. In *2018 Seventh ICT International Student Project Conference (ICT-ISPC)*, pages 1–4. IEEE.

Nezis, A., Papageorgiou, H., Georgiadis, P., Jiskra, P., Pappas, D., and Pontiki, M. (2018). Towards a fully personalized food recommendation tool. In *Proceedings of the 2018 International Conference on Advanced Visual Interfaces*, page 77. ACM.

Nirmal, I., Caldera, A., and Bandara, R. D. (2018). Optimization framework for flavour and nutrition balanced recipe: A data driven approach. In *2018 5th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, pages 1–9. IEEE.

Ooi, A., Iiba, T., and Takano, K. (2015). Ingredient substitute recommendation for allergy-safe cooking based on food context. In *2015 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM)*, pages 444–449. IEEE.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1985). Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science.

Su, H., Lin, T.-W., Li, C.-T., Shan, M.-K., and Chang, J. (2014). Automatic recipe cuisine classification by ingredients. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing: adjunct publication*, pages 565–570. ACM.

Trattner, C. and Elsweiler, D. (2017). Investigating the healthiness of internet-sourced recipes: implications for meal planning and recommender systems. In *Proceedings of the 26th international conference on world wide web*, pages 489–498. International World Wide Web Conferences Steering Committee.

Yasukawa, M. and Scholer, F. (2017). Concurrence of word concepts in cooking recipe search. In *Proceedings of the 9th Workshop on Multimedia for Cooking and Eating Activities in conjunction with The 2017 International Joint Conference on Artificial Intelligence*, pages 25–30. ACM.