UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

ESCOLA DE ENGENHARIA

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

Filipe Rissieri Lucini

APLICAÇÕES DE MINERAÇÃO DE TEXTOS NA GESTÃO DE OPERAÇÕES

Porto Alegre

2018

Filipe Rissieri Lucini

Aplicações de Mineração de Textos na Gestão de Operações

Tese submetida ao Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal do Rio Grande do Sul como requisito parcial à obtenção do título de Doutor em Engenharia, modalidade Acadêmica, na área de concentração em Sistemas de Qualidade.

Orientadores:

Flávio Sanson Fogliatto, *Ph.D.*

Giovani José Caetano da Silveira, *Ph.D.*

Porto Alegre

2018

Filipe Rissieri Lucini

Aplicações de Mineração de Textos na Gestão de Operações

Esta tese foi julgada adequada para a obtenção do título de Doutor em Engenharia de Produção na modalidade Acadêmica e aprovada em sua forma final pelo Orientador e pela Banca Examinadora designada pelo Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal do Rio Grande do Sul.

_____

Prof. Flávio Sanson Fogliatto, Ph.D.
Orientador, PPGEP/UFRGS

_____

Prof. Giovani José Caetano da Silveira, Ph.D.
Orientador, HSB/UofC

_____

Prof. Michel José Anzanello, Ph.D.
Vice-Coordenador do PPG, PPGEP/UFRGS

Banca Examinadora:

Professor. Michel José Anzanello, Ph.D. (PPGEP / UFRGS)

Professor Leandro Miletto Tonetto, Dr. (PPGD / UNISINOS)

Professor Ricardo de Souza Kuchenbecker, Dr. (PPGE / UFRGS)

À minha esposa Deise

e aos meus pais Maria Regina e João Carlos

AGRADECIMENTOS

Meus sinceros agradecimentos a todos que, direta ou indiretamente, contribuíram para a realização desta tese. Especialmente, agradeço:

À minha esposa Deise, pelo companheirismo e apoio durante esta jornada.

À minha família, pelo incentivo ao longo da minha carreira.

Ao meu orientador, professor Flávio Sanson Fogliatto, pela orientação, incentivo e convivência.

Ao meu co-orientador, professor Giovani José Caetano da Silveira, pela acolhida na *University of Calgary*, orientação, incentivo e convivência.

A todos os professores e funcionários do Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal do Rio Grande do Sul, pelos conhecimentos compartilhados e pela convivência.

Aos colegas de PPGEP, pela convivência e troca de experiências.

E a Deus por me guiar na realização de mais um objetivo pessoal e profissional.

LUCINI, Filipe R. ***Aplicações de Mineração de Textos na Gestão de Operações***, 2018. Tese (Doutorado em Engenharia de Produção) – Universidade Federal do Rio Grande do Sul, Brasil.

RESUMO

A presente tese apresenta proposições para o desenvolvimento e aplicação de técnicas de mineração de textos, de modo a contribuir para a gestão de operações nas áreas médicas e de negócios. Os objetivos desta tese são: (*i*) identificar e estruturar técnicas de mineração de texto, de modo a elaborar um método para prever internações de pacientes provenientes de emergências hospitalares, tendo como base somente os registros textuais não estruturados escritos por médicos durante o primeiro encontro médico-paciente; (*ii*) comparar previsões realizadas pelo método proposto no objetivo (i) com análises médicas realizadas por humanos, de modo a verificar se computadores podem atuar de forma autônoma na tarefa de previsão de internações de pacientes provenientes de emergências hospitalares; e (*iii*) identificar e estruturar técnicas de mineração de texto, de modo a elaborar um método para prever a satisfação de clientes de companhias aéreas, tendo como base as avaliações escritas e publicadas por passageiros na internet. Os métodos propostos utilizaram diferentes técnicas de mineração de textos, sendo validados por estudos de caso. Em relação à área médica, o método proposto pode realizar previsões em tempo real sobre a necessidade de leitos, ajudando as equipes de gerenciamento de leitos a melhorar os processos de fluxo de pacientes. Além disso, verificou-se que tanto médicos (iniciantes ou experientes), quanto máquina, tiveram desempenhos semelhantes na tarefa de previsão de internação de pacientes. Já em relação à área de negócios, o método proposto permitiu extrair dimensões de satisfação de avaliações *online*, além dos sentimentos associados a elas, considerando diferentes perfis de passageiros, serviços e períodos de tempo. Desta forma, foi possível prever a recomendação de companhias aéreas baseado nas avaliações escritas por passageiros.

Palavras-chave: Mineração de texto; gestão de operações; apoio à decisão; departamento de emergência hospitalar; transporte aéreo de passageiros.

LUCINI, Filipe R. ***Applications of Text Mining Techniques in Operations Management***, 2018. Dissertation (Doctorate in Industrial Engineering) – Universidade Federal do Rio Grande do Sul, Brazil.

ABSTRACT

This dissertation presents propositions for the development and application of text mining techniques, in order to contribute to operations management in the medical and business areas. The objectives of this dissertation are: (*i*) identify and structure text mining techniques, in order to propose a method to predict admissions of patients from hospital emergencies, based only on unstructured textual records written by physicians during the first encounter with patients; (*ii*) compare predictions made by the method proposed in objective (i) with medical analyses carried out by humans, in order to verify if computers can work autonomously in predicting hospitalizations of patients coming from hospital emergencies; and (*iii*) identify and structure text mining techniques to develop a method for predicting airline customer satisfaction based on online customer reviews. The proposed methods used different text mining techniques, being validated by case studies. Regarding the medical area, the proposed method was able to perform real-time forecasts about the need for beds, helping bed management teams to improve patient flow processes. In addition, it was found that both physicians (novice or experienced) and machine had similar performances in predicting patient hospitalization. In relation to the business area, the proposed method allowed to extract satisfaction dimensions of online customer reviews, as well as sentiments associated to them, considering different profiles of passengers, services and time periods. It also enabled the prediction of airline recommendation based on online customer reviews.

Keywords: Text mining; operations management; decision support; hospital emergency department; airline industry.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A mineração de textos é um processo de descoberta de conhecimento que utiliza técnicas de análise e extração de dados a partir de textos não estruturados (ANANIADOU, 2008; LIU; CAO; HE, 2011; ROMERO; VENTURA, 2010; ZAFRA; VENTURA, 2009; ZENG et al., 2012). Ela envolve a aplicação de algoritmos computacionais que processam textos e identificam informações úteis e implícitas, que normalmente não poderiam ser recuperadas utilizando métodos tradicionais de consulta (MORAIS; AMBRÓSIO, 2007).

Os benefícios da mineração de textos podem se estender a qualquer domínio que utilize textos (LOH, 2001). Suas principais contribuições estão relacionadas à busca de informações específicas em documentos, à análise qualitativa e quantitativa, e à melhor compreensão do conteúdo disponível em documentos textuais (MORAIS; AMBRÓSIO, 2007). Ela foca na identificação de tendências, padrões ou regras em textos não estruturados (ABDOUS; HE, 2011; CHIANG; LIN; CHEN, 2011; LIN; HSIEH; CHUANG, 2009). Além disso, busca criar interpretações ou modelos que expliquem esses padrões e tendências (GUO; XIAO; GONG, 2012; ROMERO; VENTURA; GARCÍA, 2008).

Diversos estudos reportam resultados satisfatórios na aplicação de técnicas de mineração de textos para análise de dados textuais nas áreas de negócios (HE; ZHA; LI, 2013; INGVALDSEN; GULLA, 2012; LAU; LI; LIAO, 2014), ciências médicas (GENES et al., 2013; LI et al., 2012; YANG et al., 2009) e na área educacional (ABDOUS; HE, 2011; HUNG, 2012; TANE; SCHMITZ; STUMME, 2004). Por este motivo, acredita-se que a aplicação das técnicas de mineração de textos pode trazer benefícios à gestão de operações em diferentes áreas. Dentre elas, destacam-se oportunidades nas áreas médica e de negócios.

Em relação à área médica, um problema comum relacionado à gestão de operações diz respeito ao ajuste do fluxo de pacientes em hospitais. Neste sentido, gestores hospitalares procuram ajustar os processos de internação, alta e transferência de pacientes, de modo a proporcionar um tempo de permanência ideal nos hospitais (NAPHHS, 2005). Ineficiências nessa gestão podem ocasionar descompasso entre a oferta e a demanda de recursos. Verifica-se que as ineficiências mais comuns estão relacionadas ao baixo volume de recursos, incapacidade de atender a demanda de pico e falta de coordenação (HOOT; ARONSKY, 2008). Usualmente, as unidades de emergência, de cuidados pós-anestesia, de tratamento intensivo e de internação são os setores mais impactados por esse descompasso. Como consequência, os pacientes acabam aguardando longos períodos entre as etapas do seu

atendimento (LOVETT; ILLG; SWEENEY, 2014), o que pode levar a complicações nas suas condições de saúde (DERLET; RICHARDS, 2000).

A lotação de unidades de emergência está diretamente relacionada a diversos problemas, tais como: (i) insatisfação dos pacientes em relação ao atendimento; (ii) demora no controle dos sintomas; (iii) queda na qualidade do tratamento de doenças; e (iv) aumento no risco de óbito (DERLET; RICHARDS, 2000; PINES et al., 2009). Independentemente do tamanho do hospital, a unidade de emergência é uma área de interações complexas, pois envolve diversos agentes com processos relacionados, os quais podem possuir recursos desequilibrados (HOOT; ARONSKY, 2008). Sabe-se que a lotação da unidade de emergência ocorre em função de causas complexas e inter-relacionadas. As mais comuns incluem: (i) aumento do volume de pacientes, (ii) maior complexidade dos atendimentos, (iii) falta de leitos para internação, (iv) número reduzido de funcionários para atendimento, (v) atrasos na emissão de resultados de exames e (vi) pouco espaço físico na unidade de emergência (CRAWFORD et al., 2014; DERLET; RICHARDS, 2000; FATOVICH; HIRSCH, 2003).

Nos últimos anos, diversas abordagens foram propostas visando reduzir as lotações das unidades de emergência (FORERO et al., 2010). Pode-se citar (RUFFIN; HOOPER, 2003): (i) mudanças nos fluxos de pacientes, (ii) introdução de novas funções de enfermagem, (iii) gestão de grupos específicos de pacientes, (iv) ações para pacientes frequentes, (v) implantação de equipes coordenadas de atendimento, (vi) uso de salas de trânsito, e (vii) introdução de unidades de avaliação rápida de pacientes.

Neste contexto, a mineração de textos surge como uma potencial ferramenta de auxílio à tomada de decisões e gestão de recursos. Por exemplo, sabe-se que para todo atendimento em emergência são feitos registros em formato texto livre, desde a entrada do paciente, até os exames e consultas realizadas. A mineração desses registros textuais, considerando as primeiras horas de atendimento, pode viabilizar a antecipação da tomada de decisões, melhorando a gestão do fluxo de pacientes.

Já em relação à área de negócios, uma área que vem recebendo atenção diz respeito à inteligência competitiva. Ela pode ser definida como uma disciplina econômica dinâmica que busca reunir, analisar, comunicar e gerir o conhecimento sobre o ambiente econômico, gerando vantagens competitivas (ŠTEFÁNIKOVÁ; MASÁROVÁ, 2014). Segundo as autoras, gestão do conhecimento é uma importante parte da inteligência competitiva. Ela atua na aquisição sistemática, disseminação e na receptividade das informações, buscando o aumento da performance das organizações (HU, 2010).

Desde o surgimento da internet, a maneira como as pessoas interagem com a informação vem mudando. Um dos pontos que merece destaque é a maneira como as informações são compartilhadas. Nos dias de hoje, por exemplo, são comuns as publicações em *blogs*, *wikis*¸ fóruns de discussão e redes sociais. Nestes exemplos, os usuários postam informações, dão opiniões e recebem comentários de outros usuários. Desta forma, é estabelecida uma fonte rica de informações comportamentais sobre diferentes aspectos da vida humana (TSYTSARAU; PALPANAS, 2012).

A diversidade de opiniões, quando analisada, pode apresentar novas informações e contribuir para o conhecimento de um assunto específico. Essa ideia é especialmente reforçada quando o objeto de estudo está relacionado às atitudes das pessoas (TSYTSARAU; PALPANAS, 2012). Por exemplo, sabe-se que as avaliações de usuários postadas na internet podem influenciar a escolha de compra de determinados produtos (HE; ZHA; LI, 2013), ou ainda direcionar a indústria no desenvolvimento de novos produtos e de estratégias de mercado (HOFFMAN, 2008). Nesse contexto, acredita-se que a extração e análise de textos publicados em milhares de avaliações online possa fornecer orientações para as companhias melhorarem sua competitividade.

1.1 TEMA E OBJETIVOS

O tema de pesquisa desta tese contempla o uso de técnicas de mineração de textos aplicadas à gestão de operações. Mais especificamente, ela está focada em aplicações nas áreas médica e de negócios. Em relação à área médica, ela está relacionada à gestão do fluxo de pacientes em emergências hospitalares. Já em relação à área de negócios, ela está direcionada à gestão de informação. Mais especificamente, ela está direcionada à análise e previsão da satisfação de passageiros de companhias aéreas.

O objetivo geral desta pesquisa é desenvolver e aplicar técnicas de mineração de textos, de modo a contribuir para a gestão de operações nas áreas médicas e de negócios. Em decorrência do objetivo principal, são relacionados três objetivos específicos:

a) Identificar e estruturar técnicas de mineração de texto, de modo a elaborar um método para prever internações de pacientes provenientes de emergências hospitalares, tendo como base somente os registros textuais não estruturados escritos por médicos durante o primeiro encontro médico-paciente;

b) Comparar previsões realizadas pelo método proposto no objetivo a) com análises médicas realizadas por humanos, de modo a verificar se computadores podem atuar de forma autônoma na tarefa de previsão de internações de pacientes provenientes de emergências hospitalares;
c) Identificar e estruturar técnicas de mineração de texto, de modo a elaborar um método para prever a satisfação de clientes de companhias aéreas, tendo como base as avaliações escritas e publicadas por passageiros na internet.

1.2 JUSTIFICATIVA DO TEMA E DOS OBJETIVOS

A lotação de emergências hospitalares (EH) é um problema sério para hospitais, representando riscos para a segurança dos pacientes (SPRIVULIS et al., 2006). Sabe-se que esse fenômeno ocorre principalmente quando a demanda por leitos excede a capacidade da EH (HOOT; ARONSKY, 2008). Nos últimos anos, o número de leitos hospitalares vem reduzindo (ALIJANI et al., 2003; MACKAY, 2005; OECD, 2011; UNIÃO, 2013), apesar do aumento da demanda e da expectativa por melhores níveis de atendimento (LUCAS et al., 2009; SCHMIDT et al., 2011). No Brasil, os hospitais têm um papel importante, visto que promovem um dos principais acessos ao sistema de saúde. Além do atendimento especializado, os hospitais oferecem atendimento emergencial, seja para pessoas que acessaram o atendimento por conta própria, seja para pessoas enviadas por outras unidades de atendimento de saúde (NAPHHS, 2005; WILER et al., 2012).

O uso de modelos de previsão de demanda é a abordagem clássica para realizar a gestão de capacidade e alocação de recursos em uma EH (FARMER; EMAMI, 1990; MACKAY, 2005; SWEENEY; ASHLEY, 1981). A demanda futura é estimada através da análise de preditores relevantes obtidos através dos registros dos pacientes na EH (BIRD et al., 2007; GARCÍA-GUBERN; COLÓN-ROLÓN; APONTE, 2005). Entretanto, os modelos tradicionais de previsão de demanda dependem de dados quantitativos, normalmente obtidos por resultados de laboratório e exames ambulatoriais, os quais raramente estão disponíveis no início do atendimento. Por outro lado, predições com o uso de mineração de textos (MT) dependem somente de dados textuais. Nesse sentido, podem gerar informações úteis a partir das anotações e registros realizados nas primeiras etapas do atendimento, mesmo antes que resultados de exames estejam disponíveis (MOWERY et al., 2012).

Nos últimos anos, a evolução de tecnologias de computação e de técnicas de programação vem permitindo que computadores emulem tarefas normalmente executadas por seres humanos (CANTU-ORTIZ, 2014). Nesse cenário, pelo menos cinco argumentos apoiam a ideia de que sistemas inteligentes podem superar os humanos em tarefas comuns da área médica. Primeiro, na maioria dos sistemas inteligentes, as regras são criadas pela máquina e não pelas pessoas, permitindo que padrões não triviais sejam descobertos (CHOCKLEY; EMANUEL, 2016). Segundo, os sistemas inteligentes têm menos limitações cognitivas (CHOCKLEY; EMANUEL, 2016), não sendo afetados por aspectos comuns do sistema de saúde, como altas cargas de trabalho, execução de multitarefas, interrupções, pressão de tempo e mudanças de turno (LAXMISAN et al., 2007). Terceiro, a crescente velocidade de processamento dos computadores torna os sistemas inteligentes significativamente mais rápidos que humanos na tomada de decisões (TAJIK, 2016). Quarto, sistemas inteligentes não apresentam variação entre observadores (LEE et al., 2017). Em outras palavras, dois computadores diferentes usando o mesmo algoritmo e treinados com o mesmo conjunto de dados darão a mesma resposta a um problema. Em contraste, médicos diferentes podem ter opiniões diferentes sobre o mesmo problema, mesmo compartilhando experiências ou histórico similares. Quinto, a base de dados utilizada por sistemas inteligentes para a tomada de decisões pode aumentar continuamente. Além disso, os computadores podem compartilhar os bancos de dados com outros computadores em uma rede, sem que ocorra perda de informações (TAJIK, 2016).

Sabe-se que a EH é um ambiente complexo com alta carga de trabalho, sensível ao tempo, altamente estressante, não-determinístico, repleto de interrupções e crítico para a vida dos pacientes (FRANKLIN et al., 2011). Nestas condições, a tomada de decisões eficientes é difícil ou mesmo impossível de ser alcançada pelos seres humanos (FRANKLIN et al., 2011). Apesar de alguns esforços no desenvolvimento de sistemas inteligentes para apoio de tomada de decisões em EH (MATOS; RODRIGUES, 2011), não existem muitas evidências de estudos avaliando a capacidade de sistemas inteligentes de operar de forma totalmente automatizada e autônoma. Ainda não se tem conhecimento suficiente para determinar se os sistemas baseados em computadores podem ser considerados adequados para substituir humanos na tomada de decisões em EH, sejam as decisões clínicas ou gerenciais. Mais especificamente em relação ao foco desta pesquisa na área médica, não existem evidências de comparações entre as performances de máquinas e humanos na tarefa de prever a internação de pacientes provenientes de uma EH.

Em relação à área de negócios, mais especificamente ao setor de transporte aéreo de passageiros, a aplicação de técnicas de mineração de textos pode auxiliar a extrair informações não explícitas, de modo a constituir vantagens competitivas para as companhias aéreas. Sabe-se que o setor de companhias aéreas opera em um cenário altamente competitivo, no qual as empresas têm de lidar com diversos desafios para ter sucesso (CALISIR; BASAK; CALISIR, 2016; DOLNICAR et al., 2011). Em um ambiente econômico tão desafiador, atingir e manter um alto nível de satisfação dos passageiros é visto como uma vantagem competitiva fundamental (CHEN, 2008; LI et al., 2017). Nesse sentido, é importante não apenas entender como os passageiros avaliam os serviços das companhias aéreas, mas também identificar as dimensões de satisfação consideradas mais relevantes por eles (PARK; ROBERTSON; WU, 2004).

Os passageiros tendem a ficar satisfeitos com uma companhia aérea quando os atributos de qualidade de serviço considerados mais importantes são atendidos ou superados (CHOW, 2015). Diversos estudos afirmam que a satisfação desempenha um papel importante na fidelização dos clientes. Passageiros satisfeitos tendem a realizar críticas positivas e repetir voos, além de recomendar as companhias a passageiros (FORGAS et al., 2010; GUO; BARNES; JIA, 2017; MATTILA, 2004; MORGAN; HUNT, 1994). Por outro lado, passageiros insatisfeitos podem reconsiderar a utilização da mesma companhia aérea em voos futuros (NAMUKASA, 2013), ou iniciar uma campanha de boca-a-boca negativa, eventualmente causando danos à reputação e imagem da empresa (BLODGETT; LI, 2007).

Nos últimos anos, diferentes abordagens foram propostas para identificar e medir a qualidade dos serviços aéreos e a satisfação dos passageiros (TSAFARAKIS; KOKOTAS; PANTOUVAKIS, 2017). Uma linha de pesquisa utiliza técnicas estatísticas e regressões para modelar a relação entre as dimensões da qualidade e a satisfação dos clientes (ALI; KIM; RYU, 2016; EBOLI; MAZZULLA, 2009; JOSEPHAT; ISMAIL, 2012). Outra linha de pesquisa utiliza métodos e ferramentas de decisão multicriterial para avaliar os níveis de serviço das companhias aéreas e gerar proposições para melhorias (LI et al., 2017; LIOU et al., 2011; LIOU; TZENG, 2007). No entanto, esses métodos são demorados, caros e muitas vezes imprecisos (WAN; GAO, 2015).

Como alternativa aos métodos convencionais, estudos de diferentes campos mostraram que o conteúdo gerado por usuários (CGU) pode ser usado como uma fonte de informação para entender as preferências e demandas do cliente (CHAU; XU, 2012). CGUs são resultado da ampla difusão de tecnologias da Web 2.0 (GUO; BARNES; JIA, 2017), que

permitiram a clientes, incluindo aqueles do setor de companhias aéreas, o compartilhamento de suas experiências e opiniões. Apesar da considerável atenção dedicada à análise de CGUs em diferentes setores da indústria, existe uma lacuna na literatura relacionada à mineração destas informações no setor de transporte aéreo.

1.3 DELINEAMENTO DO ESTUDO

Uma vez definidos os objetivos geral e específicos do trabalho e apresentadas as justificativas a respeito da importância desta pesquisa, é necessário delinear os meios através dos quais estes objetivos serão alcançados. Nesse sentido, são apresentados o método de pesquisa e o método de trabalho utilizados.

1.3.1 Método de Pesquisa

A pesquisa apresentada por esta tese segue uma abordagem quantitativa. Segundo Gerhardt e Silveira (2009) esta abordagem tende a enfatizar o raciocínio dedutivo, as regras da lógica e os atributos mensuráveis da experiência humana. Neste contexto, o pesquisador deve capturar evidências de pesquisa por meio da mensuração das variáveis, ocorrendo pequena ou nenhuma interferência nas variáveis de pesquisa (MIGUEL, 2010).

O desenvolvimento dos artigos apresentados segue a metodologia hipotética-dedutiva. Esta metodologia parte da percepção de uma lacuna nos conhecimentos, da qual se formulam as hipóteses originadas de problemas teóricos/práticos existentes, que devem ser submetidos à verificação com o propósito de serem corroboradas (LAKATOS; MARCONI, 2005).

Do ponto de vista da sua natureza, esta pesquisa pode ser classificada como pesquisa aplicada, pois objetiva gerar conhecimentos com aplicação prática e dirigidos à solução de problemas específicos (CERVO; BERVIAN, 2002). Em relação aos objetivos, esta pesquisa pode ser classificada como exploratória. Pesquisas exploratórias tem o objetivo de proporcionar maior familiaridade com um problema, de modo a torná-lo mais explícito, a construir hipóteses, a aprimorar ideias ou a descobrir intuições (GIL, 2002).

1.3.2 Método de Trabalho

O desenvolvimento do trabalho e a execução das atividades com vistas a alcançar o objetivo geral proposto ocorre através de três etapas. Cada etapa está vinculada a um objetivo específico, sendo representada por um artigo. A Tabela 1.1 apresenta a estrutura das etapas da pesquisa desenvolvida.

O Artigo 1 – "Text mining approach to predict hospital admissions using early medical records from the emergency department" – propõe o uso de técnicas de mineração de texto, de modo a elaborar um método para prever internações de pacientes provenientes de emergências hospitalares, tendo como base somente os registros textuais não estruturados escritos por médicos durante o primeiro encontro médico-paciente no atendimento em emergência.

Ao todo, 3.024 combinações de ajustes e técnicas de pré-processamento e processamento dos registros são analisadas. Nesse sentido, pode-se citar: três composições de termos (*unigrams, bigrams* e *trigrams*); três métricas para representar o peso dos termos (*binary representation, term frequency,* e *term frequency-inverse document frequency*); duas métricas para seleção de variáveis (χ^2 e *F-value*); 21 referências de percentil limite para a seleção de variáveis e oito algoritmos de previsão (*Decision Tree, Random Forest, Extremely Randomized Tree, AdaBoost, Logistic Regression, Multinomial Naïve Bayes, Support Vector Machine – kernel linear,* e *Nu-Support Vector Machine – kernel linear*). O banco de dados utilizado conta com 16.703 registros médicos, representando todos os pacientes atendidos na emergência do Hospital de Clínicas de Porto Alegre no ano de 2014. A performance das diferentes combinações é verificada com o uso de três diferentes métricas: *F_1-score, precision* e *recall.* Como resultado, a combinação que utiliza *trigrams, term frequency-inverse document frequency, F-value,* percentil de 65 e o algoritmo *Nu-Support Vector Machine – kernel linear* é identificada como aquela que possui o melhor desempenho.

O Artigo 2 – "Man vs. machine: predicting hospital bed demand from an emergency department" – propõe a comparação de previsões realizadas pelo método identificado no Artigo 1 com análises médicas realizadas por humanos. Desta forma, espera-se verificar se computadores podem atuar de forma autônoma na tarefa de previsão de internações de pacientes provenientes de emergências hospitalares.

Tabela 1.1: Estrutura das etapas da pesquisa desenvolvida

Estudos	Objetivos	Questões de Pesquisa	Revisão Teórica	Método de Pesquisa
Artigo 1	Propor um método de mineração de textos para prever a internação de pacientes provenientes de emergências hospitalares, tendo como base somente os registros textuais escritos por médicos no primeiro encontro médico-paciente do atendimento em emergência.	Quais técnicas de mineração de texto podem ser combinadas de modo a prever satisfatoriamente a internação de pacientes provenientes de emergências hospitalares, tendo como base somente os registros textuais escritos por médicos no primeiro encontro médico-paciente do atendimento em emergência?	1. Técnicas de pré-processamento de textos. 2. Técnicas de classificação/previsão de registros baseadas em textos (aprendizado supervisionado).	Pesquisa quantitativa: 1. Análise do método proposto. 2. Validação em estudo de caso.
Artigo 2	Comparar previsões realizadas pelo método de melhor desempenho identificado no Artigo 1 com análises médicas realizadas por humanos.	Computadores podem atuar de forma autônoma na tarefa de previsão de internações de pacientes provenientes de emergências hospitalares?	1. Uso de sistemas inteligentes na tomada de decisões gerenciais e clínicas na área médica.	Pesquisa quantitativa: 1. Análise via estudo de caso.
Artigo 3	Propor um método de mineração de textos para medir a satisfação de clientes na indústria de transporte aéreo de passageiros.	Quais técnicas de mineração de texto podem ser combinadas de modo a medir satisfatoriamente a satisfação de clientes na indústria de transporte aéreo de passageiros?	1. Técnicas de pré-processamento de textos. 2. Técnicas de classificação/previsão de registros baseadas em textos (aprendizado supervisionado e não supervisionado).	Pesquisa quantitativa: 1. Análise do método proposto. 2. Validação em estudo de caso.

O banco de dados utilizado neste artigo é o mesmo utilizado no Artigo 1. O subconjunto de testes, utilizado para comparar o desempenho de homens e máquinas, é composto por 230 registros selecionados aleatoriamente. A representação humana da comparação é formada por quatro médicos, sendo dois considerados experientes e dois considerados novatos (com mais de dez anos de experiência e até dois anos de experiência em emergências hospitalares, respectivamente). Todos os 230 registros são avaliados por médicos experientes e novatos. Ao final, as previsões são reunidas e comparadas com aquelas indicadas pelo método de melhor desempenho, conforme análise do Artigo 1.

O Artigo 3 – "Text mining approach to explore dimensions of airline customer satisfaction using spontaneous generated content" – propõe elaborar um método para prever a satisfação de clientes de companhias aéreas, tendo como base os textos de avaliações escritas e publicadas por passageiros na internet.

O banco de dados utilizado no Artigo 3 conta com mais de 55.000 avaliações online dos serviços prestados por mais de 400 companhias aéreas a passageiros de mais de 170 nacionalidades. Técnicas de mineração de textos supervisionadas e não-supervisionadas são aplicadas de modo a identificar as dimensões de satisfação existentes nas avaliações, além dos sentimentos a elas relacionados. Complementarmente, o uso combinado destas informações é utilizado para prever a recomendação das companhias aéreas, sendo identificadas as dimensões mais relevantes para a previsão de recomendação. O método mostra sensibilidade para identificar variações na distribuição de dimensões de acordo com as diferentes características e preferências dos passageiros (por exemplo, nacionalidade, tipo de viajante e cabine voada), bem como tendências nas preferências dos clientes.

1.4 DELIMITAÇÕES DO ESTUDO

A presente pesquisa incide sobre aspectos da descoberta de conhecimento em textos não estruturados com o uso de técnicas de mineração de textos. Todas as análises e métodos apresentados nesta tese estão baseados exclusivamente em análises de textos, sendo excluídas quaisquer análises de usos combinados ou comparações com técnicas que utilizam outros tipos de dados.

No tocante à abrangência, a pesquisa concentra-se em duas áreas: médica e de negócios. Em relação à área médica, são apresentadas análises relacionadas exclusivamente aos registros médicos textuais gerados no primeiro encontro médico-paciente em emergências hospitalares. Não são observadas outras aplicações além da previsão de internação ou alta dos pacientes atendidos. Cabe ressaltar que as análises da área médica são apresentadas considerando somente a língua portuguesa, não sendo analisadas aplicações em outras línguas. Adicionalmente, nesta área de aplicação somente técnicas supervisionadas foram analisadas. Já em relação à área de negócios, são apresentadas análises relacionadas exclusivamente às avaliações online escritas e registradas por passageiros de companhias aéreas. Todos os registros analisados estão escritos em língua inglesa, não sendo contempladas análises que envolvam outras línguas. A medida de satisfação utilizada neste estudo é a recomendação das companhias aéreas. Eventuais métricas adicionais de satisfação não estão contempladas nesse estudo. Complementarmente, não são apresentadas análises de diferenças de desempenho entre diferentes técnicas de identificação de dimensões de satisfação, de sentimentos ou de previsão de recomendação de companhias aéreas.

1.5 ESTRUTURA DA TESE

A tese está organizada em cinco capítulos principais. Neste primeiro capítulo foram apresentados a contextualização do trabalho, os objetivos, além de justificativas referentes à importância desta pesquisa sob os pontos de vista acadêmico e prático. Adicionalmente, este capítulo apresentou o método de trabalho, a estrutura e as delimitações do estudo. Os capítulos seguintes, de dois a quatro, apresentam os artigos contendo os desenvolvimentos propostos, seguindo a estrutura previamente apresentada. O capítulo final apresenta as conclusões da tese e sugestões de futuras pesquisas a serem realizadas com base nos resultados apresentados.

1.6 REFERÊNCIAS

ABDOUS, M.; HE, W. Using text mining to uncover students' technology-related problems in live video streaming. **British Journal of Educational Technology**, v. 42, n. 1, p. 40–49, jan. 2011.

ALI, F.; KIM, W. G.; RYU, K. The effect of physical environment on passenger delight and satisfaction: Moderating effect of national identity. **Tourism Management**, v. 57, p. 213–224, 2016.

ALIJANI, A. et al. Instrument for objective assessment of appropriateness of surgical bed occupancy : validation study Representation of South Asian people in randomised clinical trials : analysis of trials ' data. p. 1243–1244, 2003.

ANANIADOU, S. **What is Text Mining ?** NCRM Research Methods Festival. **Anais**...2008

BIRD, S. R. et al. Integrated care facilitation for older patients with complex health care needs reduces hospital demand. v. 31, n. 3, p. 451–461, 2007.

BLODGETT, J. G.; LI, H. Assessing the effects of post-purchase dissatisfaction and complaining behavior on profitability: a Monte Carlo simulation. **Journal of Consumer Satisfaction, Dissatisfaction & Complaining Behavior**, v. 20, p. 1–14, 2007.

CALISIR, N.; BASAK, E.; CALISIR, F. Key drivers of passenger loyalty: A case of Frankfurt-Istanbul flights. **Journal of Air Transport Management**, v. 53, p. 211–217, 2016.

CANTU-ORTIZ, F. J. Advancing artificial intelligence research and dissemination through

conference series: Benchmark, scientific impact and the MICAI experience. **Expert Systems with Applications**, v. 41, n. 3, p. 781–785, 2014.

CERVO, A. L.; BERVIAN, P. A. **Metodologia científica**. São Paulo: Prentice Hall, 2002.

CHAU, M.; XU, J. Business intelligence in blogs: Understanding consumer interactions and communities. **MIS Quarterly: Management Information Systems**, v. 36, n. 4, p. 1189–1216, 2012.

CHEN, C.-F. Investigating structural relationships between service quality, perceived value, satisfaction, and behavioral intentions for air passengers: Evidence from Taiwan. **Transportation Research Part A**, v. 42, n. 4, p. 709–717, maio 2008.

CHIANG, D. M.-H.; LIN, C.-P.; CHEN, M.-C. The adaptive approach for storage assignment by mining data of warehouse management system for distribution centres. **Enterprise Information Systems**, v. 5, n. 2, p. 219–234, maio 2011.

CHOCKLEY, K.; EMANUEL, E. The End of Radiology? Three Threats to the Future Practice of Radiology. **Journal of the American College of Radiology**, v. 13, n. 12, p. 1415–1420, 2016.

CHOW, C. K. W. On-time performance, passenger expectations and satisfaction in the Chinese airline industry. **Journal of Air Transport Management**, v. 47, p. 39–47, 2015.

CRAWFORD, K. et al. Initiatives to reduce overcrowding and access block in Australian emergency departments: A literature review. **Collegian**, v. 21, n. 4, p. 359–366, dez. 2014.

DERLET, R. W.; RICHARDS, J. R. Overcrowding in the nation's emergency departments: complex causes and disturbing effects. **Annals of emergency medicine**, v. 35, n. 1, p. 63–68, 2000.

DOLNICAR, S. et al. Key drivers of airline loyalty. **Tourism Management**, v. 32, n. 5, p. 1020–1026, 2011.

EBOLI, L.; MAZZULLA, G. An ordinal logistic regression model for analysing airport passenger satisfaction. **EuroMed Journal of Business**, v. 4, n. 1, p. 40–57, 2009.

FARMER, R. D. T.; EMAMI, J. Models for forecasting hospital bed requirements in the acute sector. v. 365, n. 1, p. 307–312, 1990.

FATOVICH, D. M.; HIRSCH, R. L. Entry overload, emergency department overcrowding, and ambulance bypass. **EmergencyMedicine Journal**, v. 20, n. 5, p. 406–409, 2003.

FORERO, R. et al. Access block and ED overcrowding. **Emergency medicine Australasia : EMA**, v. 22, n. 2, p. 119–35, abr. 2010.

FORGAS, S. et al. Antecedents of airline passenger loyalty: Low-cost versus traditional airlines. **Journal of Air Transport Management**, v. 16, n. 4, p. 229–233, 2010.

FRANKLIN, A. et al. Opportunistic decision making and complexity in emergency care. **Journal of Biomedical Informatics**, v. 44, n. 3, p. 469–476, 2011.

GARCÍA-GUBERN, C.; COLÓN-ROLÓN, L.; APONTE, M. M. Workforce projections for emergency medicine in Puerto Rico: a five-year follow-up of an evident demand. **The Journal of emergency medicine**, v. 29, n. 1, p. 107–10, jul. 2005.

GENES, N. et al. Validating Emergency Department Vital Signs Using a Data Quality Engine for Data Warehouse. p. 34–39, 2013.

GERHARDT, T. E.; SILVEIRA, D. T. **Métodos de Pesquisa**. 1. ed. Porto Alegre: Editora da UFRGS, 2009.

GIL, A. C. **Como elaborar projetos de pesquisa**. São Paulo: ATLAS, 2002.

GUO, J.; XIAO, G.; GONG, Z. Improving Multilingual Semantic Interoperation in Cross-Organizational Enterprise Systems Through Concept Disambiguation. **IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS**, v. 8, n. 3, p. 647–658, 2012.

GUO, Y.; BARNES, S. J.; JIA, Q. Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. **Tourism Management**, v. 59, p. 467–483, 2017.

HE, W.; ZHA, S.; LI, L. Social media competitive analysis and text mining: A case study in the pizza industry. **International Journal of Information Management**, v. 33, n. 3, p. 464–472, jun. 2013.

HOFFMAN, T. **Online reputation management is hot -- but is it ethical ?**

HOOT, N. R.; ARONSKY, D. Systematic review of emergency department crowding: causes, effects, and solutions. **Annals of emergency medicine**, v. 52, n. 2, p. 126–36, ago. 2008.

HU, Y. **The Role of Marketing Performance in the Relationship between Knowledge Management and Financial Performance**. 2010 Third International Conference on Knowledge Discovery and Data Mining. **Anais**...Phuket, Thailand: 2010Disponível em: <http://ieeexplore.ieee.org/document/5432553/>

HUNG, J. Trends of e-learning research from 2000 to 2008: Use of text mining and bibliometrics. **British Journal of Educational Technology**, v. 43, n. 1, p. 5–16, jan. 2012.

INGVALDSEN, J. E.; GULLA, J. A. Industrial application of semantic process mining. **Enterprise Information Systems**, v. 6, n. 2, p. 139–163, maio 2012.

JOSEPHAT, P.; ISMAIL, A. A Logistic Regression Model of Customer Satisfaction of Airline. **International Journal of Human Resource Studies**, v. 2, n. 4, p. 255, 2012.

LAKATOS, E. M.; MARCONI, M. A. **Fundamentos de metodologia científica**. São Paulo: ATLAS, 2005.

LAU, R. Y. K.; LI, C.; LIAO, S. S. Y. Social analytics: Learning fuzzy product ontologies for aspect-oriented sentiment analysis. **Decision Support Systems**, v. 65, p. 80–94, set. 2014.

LAXMISAN, A. et al. The multitasking clinician: Decision-making and cognitive demand during and after team handoffs in emergency care. **International Journal of Medical Informatics**, v. 76, n. 11, p. 801–811, 2007.

LEE, L. H. et al. Diagnosing colorectal medullary carcinoma: interobserver variability and clinicopathological implications. **Human Pathology**, v. 62, p. 74–82, 2017.

LI, L. et al. Guest Editorial Integrated Healthcare Information Systems. **IEEE Transactions on Information Technology in Biomedicine**, v. 16, n. 4, p. 515–517, 2012.

LI, W. et al. A hybrid approach based on fuzzy AHP and 2-tuple fuzzy linguistic method for evaluation in-flight service quality. **Journal of Air Transport Management**, v. 60, p. 49–64, 2017.

LIN, F.-R.; HSIEH, L.-S.; CHUANG, F.-T. Discovering genres of online discussion threads via text mining. **Computers & Education**, v. 52, n. 2, p. 481–495, fev. 2009.

LIOU, J. J. H. et al. Using a modified grey relation method for improving airline service quality. **Tourism Management**, v. 32, n. 6, p. 1381–1388, 2011.

LIOU, J. J. H.; TZENG, G.-H. A non-additive model for evaluating airline service quality. **Journal of Air Transport Management**, v. 13, n. 3, p. 131–138, 2007.

LIU, B.; CAO, S. G.; HE, W. Distributed data mining for e-business. **Information Technology and Management**, v. 12, n. 2, p. 67–79, mar. 2011.

LOH, S. **Abordagem baseada em conceitos para descoberta de conhecimento em textos**. [s.l.] UFRGS, 2001.

LOVETT, P. B.; ILLG, M. L.; SWEENEY, B. E. A Successful Model for a Comprehensive Patient Flow Management Center at an Academic Health System. **American journal of medical quality : the official journal of the American College of Medical Quality**, 30 dez. 2014.

LUCAS, R. et al. Emergency department patient flow: the influence of hospital census variables on emergency department length of stay. **Academic emergency medicine : official journal of the Society for Academic Emergency Medicine**, v. 16, n. 7, p. 597–602, jul. 2009.

MACKAY, M. Choice of Models for the Analysis and Forecasting of Hospital Beds. p. 221–230, 2005.

MATOS, J.; RODRIGUES, P. P. **Modeling decisions for hospital bed management - a review**. HEALTHINF. **Anais**...2011Disponível em: <http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0003135005040507>

MATTILA, A. S. The impact of service failures on customer loyalty. **International Journal of Service Industry Management**, v. 15, n. 2, p. 134–149, 2004.

MIGUEL, P. A. C. ET AL. **Metodologia de pesquisa em engenharia de produção e gestão de operações**. Rio de Janeiro: [s.n.].

MORAIS, E. A. M.; AMBRÓSIO, A. P. L. **Mineração de Textos**. [s.l: s.n.].

MORGAN, R. M.; HUNT, S. D. The commitmemt-trust theory of relationship marketing. **Journal of Marketing**, v. 58, n. 3, p. 20–38, 1994.

MOWERY, D. et al. Building an automated SOAP classifier for emergency department reports. **Journal of biomedical informatics**, v. 45, n. 1, p. 71–81, fev. 2012.

NAMUKASA, J. The influence of airline service quality on passenger satisfaction and loyalty. **The TQM Journal**, v. 25, n. 5, p. 520–532, 2013.

NAPHHS. **Perfecting Patient Flow**. [s.l: s.n.].

OECD. **Hospital Beds**. Health at a Glance. **Anais**...2011

PARK, J.-W.; ROBERTSON, R.; WU, C.-L. The effect of airline service quality on passengers' behavioural intentions: A Korean case study. **Journal of Air Transport Management**, v. 10, n. 6, p. 435–439, nov. 2004.

PINES, J. M. et al. The association between emergency department crowding and adverse

cardiovascular outcomes in patients with chest pain. **Academic emergency medicine : official journal of the Society for Academic Emergency Medicine**, v. 16, n. 7, p. 617–25, jul. 2009.

ROMERO, C.; VENTURA, S. Educational Data Mining : A Review of the State of the Art. **IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)**, v. 40, n. 6, p. 601–618, 2010.

ROMERO, C.; VENTURA, S.; GARCÍA, E. Data mining in course management systems: Moodle case study and tutorial. **Computers & Education**, v. 51, n. 1, p. 368–384, ago. 2008.

RUFFIN, R. E.; HOOPER, J. K. Responses to access block in Australia: The Queen Elizabeth Hospital Medical Division. **Medical Journal of Australia**, v. 178, n. 3, p. 104–105, 2003.

SCHMIDT, M. I. et al. Chronic non-communicable diseases in Brazil: burden and current challenges. **Lancet (London, England)**, v. 377, n. 9781, p. 1949–61, 4 jun. 2011.

SPRIVULIS, P. C. et al. The association between hospital overcrowding and mortality among patients admitted via western australian emergency departments. 2006.

ŠTEFÁNIKOVÁ, Ľ.; MASÁROVÁ, G. The Need of Complex Competitive Intelligence. **Procedia - Social and Behavioral Sciences**, v. 110, p. 669–677, 2014.

SWEENEY, T. K.; ASHLEY, S. A. Forecasting hospital bed needs. v. 283, n. July, 1981.

TAJIK, A. J. Machine learning for echocardiographic imaging: embarking on another incredible journey. **Journal of the American College of Cardiology**, v. 68, n. 21, p. 2296–2298, 2016.

TANE, J.; SCHMITZ, C.; STUMME, G. **Semantic Resource Management for the Web : An E-Learning Application**. Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters. **Anais**...ACM, 2004

TSAFARAKIS, S.; KOKOTAS, T.; PANTOUVAKIS, A. A multiple criteria approach for airline passenger satisfaction measurement and service quality improvement. **Journal of Air Transport Management**, p. 1–15, 2017.

TSYTSARAU, M.; PALPANAS, T. Survey on mining subjective data on the web. **Data Mining and Knowledge Discovery**, v. 24, n. 3, p. 478–514, 2012.

UNIÃO. **Relatório Sistêmico de Fiscalização da Saúde - FICSAÚDE**, 2013.

WAN, Y.; GAO, Q. An Ensemble Sentiment Classification System of Twitter Data for

Airline Services Analysis. **Proceedings - 15th IEEE International Conference on Data Mining Workshop, ICDMW 2015**, p. 1318–1325, 2015.

WILER, J. L. et al. Predictors of patient length of stay in 9 emergency departments. **The American journal of emergency medicine**, v. 30, n. 9, p. 1860–4, nov. 2012.

YANG, H. et al. A text mining approach to the prediction of disease status from clinical discharge summaries. **Journal of the American Medical Informatics Association : JAMIA**, v. 16, n. 4, p. 596–600, 2009.

ZAFRA, A.; VENTURA, S. Predicting Student Grades in Learning Management Systems with Multiple Instance Genetic Programming. **Educational Data Mining**, p. 307–314, 2009.

ZENG, L. et al. Distributed data mining: a survey. **Information Technology and Management**, v. 13, n. 4, p. 403–409, maio 2012.

2 ARTIGO 1 – TEXT MINING APPROACH TO PREDICT HOSPITAL ADMISSIONS USING EARLY MEDICAL RECORDS FROM THE EMERGENCY DEPARTMENT

Artigo publicado no International Journal of Medical Informatics 100 (2017)

Abstract

Objective: Emergency department (ED) overcrowding is a serious issue for hospitals. Early information on short-term inward bed demand from patients receiving care at the ED may reduce the overcrowding problem, and optimize the use of hospital resources. In this study, we use text mining methods to process data from early ED patient records using the SOAP framework, and predict future hospitalizations and discharges.

Design: We try different approaches for pre-processing of text records and to predict hospitalization. Sets-of-words are obtained via binary representation, term frequency, and term frequency-inverse document frequency. Unigrams, bigrams and trigrams are tested for feature formation. Feature selection is based on χ^2 and F-score metrics. In the prediction module, eight text mining methods are tested: Decision Tree, Random Forest, Extremely Randomized Tree, AdaBoost, Logistic Regression, Multinomial Naïve Bayes, Support Vector Machine (Kernel linear) and Nu-Support Vector Machine (Kernel linear).

Measurements: Prediction performance is evaluated by F1-scores. Precision and Recall values are also informed for all text mining methods tested.

Results: Nu-Support Vector Machine was the text mining method with the best overall performance. Its average F1-score in predicting hospitalization was 77.70%, with a standard deviation (SD) of 0.66%.

Conclusions: The method could be used to manage daily routines in EDs such as capacity planning and resource allocation. Text mining could provide valuable information and facilitate decision-making by inward bed management teams.

Key words: text mining; emergency departments; clinical decision support

2.1 INTRODUCTION

ED overcrowding is a serious contemporary issue for hospitals, posing a well-defined risk to patient safety (SPRIVULIS et al., 2006). It has been shown that ED overcrowding mainly occurs when demand for inpatient beds exceeds capacity, and patients remain in the ED as 'boarded' patients (HOOT; ARONSKY, 2008). In recent years, the number of inpatient beds in hospitals has declined (ALIJANI et al., 2003; MACKAY, 2005; OECD, 2011) despite increasing service level expectations and demand due to longer life expectancy (LUCAS et al., 2009; SCHMIDT et al., 2011). In Brazil, hospitals play an important role in promoting access to the health care system; in addition to specialized care, hospitals provide emergency care for people who are either sent by another health unit or walk in, without being formally referred (NAPHHS, 2005; WILER et al., 2012).

Using forecasting models is the classical approach to manage capacity and allocate resources such as hospital beds in overcrowded EDs (FARMER; EMAMI, 1990; MACKAY, 2005; SWEENEY; ASHLEY, 1981). Future demand is estimated by analyzing relevant predictors of patient entry and stay in the ED (BIRD et al., 2007; GARCÍA-GUBERN; COLÓN-ROLÓN; APONTE, 2005). However, traditional forecasting models demand quantitative entry data, usually results from laboratory and ambulatory exams, which are rarely available when patients start receiving care. On the other hand, predictions using text-mining (TM) methods rely entirely on textual data, and thus may provide useful information based on

clinical records of patients while they are under observation at the ED. Here, we hypothesize that using TM to analyze free-text medical records from ED patients may provide relevant and clinically reliable data to predict hospital admissions soon after patients have their first contact with ED physicians, even before complementary laboratory information is available (MOWERY et al., 2012).

Text mining (TM) is a method that bridges the gap between free-text and structured representation of information (FELDMANN; SANGER, 2007). It uses techniques from natural language processing, knowledge management, data mining, and machine learning to process free-text documents. The method can be applied towards information retrieval, document classification, information extraction, terminology extraction, and named entity recognition (i.e. identifying entities from predefined categories), among other applications (KARYSTIANIS; BUCHAN; NENADIC, 2014; SPASIĆ et al., 2014). TM has been used in medical studies with various objectives including automatic disease-specific classification of clinical discharges (SOLT et al., 2009), identification of patients' obesity status (AMBERT; COHEN, 2008; YANG et al., 2009) and analysis of biomedical and clinical documents to identify disease-drug associations (CHEN et al., 2008), among others (AYVAZ et al., 2015; DINAKARPADIAN; LEE; VISHWANATH, 2006; LU et al., 2006; MICHELSON; PARISEAU; PAGANELLI, 2014; PEREIRA et al., 2013; VALLMUUR, 2015; WANG et al., 2009; YANG; KIANG; SHANG, 2015; YONGJING; LI; CHEN, 2007).

In the ED context, TM has been used to convert vital sign data from plain text to numerical format so they can be exported to data warehouses and analyzed with conventional methods (GENES et al., 2013). For example, St-Maurice e Kuo (2012) converted primary care information from text to numerical data for logistic regression, predicting inappropriate use of ED. Mowery *et al*. (MOWERY et al., 2012) assessed ED medical records through an automated Subjective, Objective, Assessment, and Plan (SOAP) framework. The study showed that the

use of support vector machine-based classifiers to automatically map sentences from clinical narratives onto an SOAP framework resulted in high inter-annotator agreement and excellent coverage for sentences in ED reports (MOWERY et al., 2012).

2.2 METHODS

In this study, we apply TM methods to evaluate consultation records from the ED of a tertiary teaching hospital, to develop a prediction model of inpatient bed demand. We also use a large dataset made available by the hospital to compare different configurations of free text pre-processing strategies and TM classification methods.

This is a retrospective study using TM methods to forecast inpatient bed demand based on medical records generated in the first doctor-patient encounter at the ED, before the information from laboratory exams is available. The study was performed at the ED of Hospital de Clínicas de Porto Alegre (HCPA), an 842-bed, tertiary care teaching public hospital in Porto Alegre, Brazil. We obtained electronic medical records from the ED in a database that contained relevant information on clinical care provided to individuals, including medical records, prescriptions, and exams of ambulatory and ED patients. The HCPA's Ethical Committee has approved this study, and authors have complied with the recommendations of the Declaration of Helsinki.

To assess the ability of TM approaches in forecasting bed demands, we used written records of the first medical assessment in the ED. Figure 2.1 gives an overview of the method, which is divided in three modules: (*i*) pre-processing, (*ii*) prediction, and (*iii*) performance evaluation.

Analyses were carried out in Python 3.4 (PYTHON SOFTWARE FOUNDATION, 2017). The package included libraries to support records processing and information retrieval,

besides application and validation of the prediction method. We used the NLTK (BIRD; KLEIN; LOPER, 2009) and scikit-learn (HACKELING, 2014) libraries.

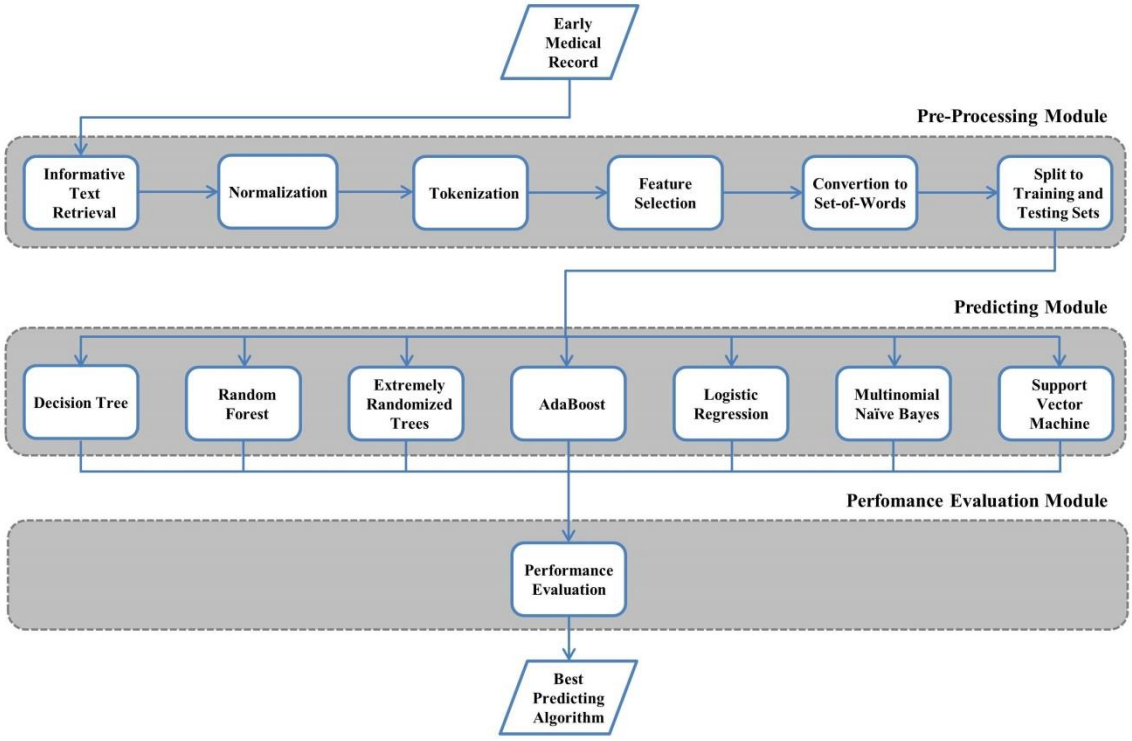

Figure 2.1: Overview of proposed method

2.2.1 Pre-Processing Module

In this module, the dataset is prepared to allow the use of classification methods in the prediction module. The module has six steps including informative text retrieval, normalization, tokenization, feature selection, conversion to set-of-words, and split to training and test sets.

Data from ED patient records were made available by HCPA in *comma-separated values* (.csv) format. Records were written in Brazilian Portuguese. The dataset was structured with two fields of information. The first was a binary variable representing patient final status (inpatient or discharged); this variable was used to classify the studied records. The second contained free-form text entered by HCPA healthcare providers, with no intermediaries (i.e.

scribes). Only the first medical records, consisting of textual assessments made by physicians on patients, were included in the analysis, regardless of the time elapsed between admission and record entry. Only data from the second field were treated in the pre-processing module. Numerical entries (corresponding, for example, to exam results) were not available in the dataset; classification was based exclusively on textual medical records.

Preliminary analysis indicated that all records had similar content, which included visit record number, urgency type – an internal scale to determine type of treatment rather than acuity (adult clinical, adult surgical, pediatric clinical, or pediatric surgical), assisting healthcare teams, and SOAP framework notes. SOAP was used to structure progress records to facilitate problem-specific, clinical decision-making by attending ED physicians and nurses (MOWERY et al., 2012). All records ended with the statement, "Prepared and signed by (…)". Most of the text was automatically generated by the electronic charting system. Only the SOAP notes were deemed informative, and thus used in the analyses to follow. Remaining contents were disregarded for not providing discriminant information on patient's progress.

Next, SOAP notes were analyzed. To improve the reliability of tests, we wanted to exclude records in which the patient's final status was explicitly described in the text (e.g. texts informing the decision to admit or discharge patients, as well as discharges that took place against medical advice) Using records with such unambiguous information would give obvious clues for the text mining methods to predict patient classification.

Once records were reduced to parts carrying relevant information, normalization took place. Punctuation marks and stop words were removed, capital letters substituted, and words reduced to radicals excluding information on gender and grade (morphological normalization) (FELDMANN; SANGER, 2007).

Next, tokenization was carried out. This consists of breaking down a continuous text into linguistic units (tokens) such as words or sentences (FELDMANN; SANGER, 2007). Since

all records were written in Portuguese, and numbers and punctuation marks had been removed in the previous step, tokens were delimited by blank spaces in the continuous text. After this step, each record contained a set of tokens.

Once tokens were identified, features could be defined. We defined three types of features. The first was comprised exclusively of unigrams (i.e. single words); the second included bigrams, and the third included trigrams. Unigrams, bigrams and trigrams were defined as sequences of one, two or three adjacent words from a list of tokens (FELDMANN; SANGER, 2007).

In the feature selection step, features were ranked per an index, and those above a given percentile (of largest values) were retained. Evaluations were based on χ^2 and F-values. These statistics were chosen because of their good performance in previous studies (BASHIR et al., 2016; ILANGO, 2010; MEESAD; BOONRAWD; NUIPIAN, 2011; MOWERY et al., 2012; OHEN et al., 2006). Twenty-one percentiles were tested, ranging from 1 to 100.

The χ^2 test is used in statistics to verify the independence between a pair of events, among other uses (UYSAL; GUNAL, 2012). In feature selection, χ^2 was used to test whether the occurrence of features in classes were independent. It was based on the following equation:

$$\chi^2(t, C) = \sum_{t\in\{0,1\}} \sum_{C\in\{0,1\}} \frac{(N_{t,C} - E_{t,C})^2}{E_{t,C}} \quad (1)$$

where *N* is the observed occurrence frequency, and *E* is the expected occurrence frequency of feature *t* in class *C*. For the independence hypothesis (H_0) to hold, $N_{t,C}$ and $E_{t,C}$ should converge to 0.5, and the difference to zero. In opposition, large χ^2 values indicate that H_0 should be rejected; in this case, feature *t* should be selected (UYSAL; GUNAL, 2012).

For a given set of features, the F-value of a particular feature is calculated by (BASHIR et al., 2016; CHEN; LIN, 2006):

$$F\_value(i) = \frac{(X_i^{imp} - X'_i)^2 + (X_i^{dis} - X_i)^2}{\frac{1}{n_{imp}-1}\sum_{k=1}^{n_{imp}}(X_{k,i}^{imp} - X_i^{imp})^2 + \frac{1}{n_{dis}-1}\sum_{k=1}^{n_{dis}}(X_{k,i}^{dis} - X_i^{dis})^2} \quad (2)$$

where X_i, X_i^{imp} and X_i^{dis} is the average of the *i*-th feature in the complete, inpatient, and discharged datasets respectively; $X_{k,i}^{imp}$ is the *i*-th feature of the *k*-th inpatient instance, and $X_{k,i}^{dis}$ is the *i*-th feature of *k*-th discharge instance. The numerator indicates the discrimination between the inpatient and discharged sets, and the denominator indicates the one within each of the two sets. A threshold value is used to select appropriate features from a set. If the F-value of a given feature is greater than a threshold value, the feature is added to the selected feature space; otherwise, it is removed from that space. The threshold value is obtained by calculating the average of F-values overall features. Eq. (2) is used to calculate the F-value of *i*-th feature, where n_{imp} is number of inpatient instances, and n_{dis} is number of discharge instances. X_k is a given vector where $k = 1,\ldots, n$. High F-values indicate discriminating features (BASHIR et al., 2016).

Once all features were selected, the conversion to set-of-words representation took place. For that, a matrix was created indicating the occurrence of features in records. Matrix columns corresponded to record features and rows corresponded to visit records. We filled out matrix cells using three different indicators: binary representation (BR), term frequency (TF), and term frequency-inverse document frequency (TF-IDF). In BR, whenever a term appeared in a record, a value of 1 was assigned to the cell in which column and row intersect; otherwise, a value of 0 was assigned to that cell. TF gives the frequency in which a feature appears in a record. TF-IDF reflects the importance of a feature in a record from a collection of records. The

TF-IDF value increases proportionally to the frequency of the feature in a record, but is compensated by the frequency of the feature in the collection of records. Mathematically, it is expressed as (FELDMANN; SANGER, 2007):

$$TFIDF(t,d,D) = \frac{f(t,d)}{\max\{f(t,d):t\in d\}} \times \log\frac{|D|}{|\{d\in D:t\in d\}|} \quad (3)$$

where t denotes the feature, d denotes the record, D is the total number of records in the collection, and $f(t,d)$ is the number of occurrences of feature t in record d (FELDMANN; SANGER, 2007).

In the last step of the pre-processing module, a set-of-words was divided in training and test sets to allow k-fold cross-validation. The method consists in dividing the dataset in k mutually exclusive subsets of equal size, such that one subset is used for testing and $k-1$ subsets are used for parameter estimation. This process is carried out k times alternating the test subset; performance statistics are calculated from the results (HASTIE; TIBSHIRANI; FRIEDMAN, 2008). In this study, k was set to 10.

2.2.2 Prediction Module

Different classification methods were tested in this module. In this study, the need for inpatient beds depended on inpatient class, i.e. if a record was classified as inpatient, a new bed was required, and if a record was classified as discharged, no hospitalization took place.

Classification methods were selected based on their reported performances in previous text classification studies involving many features. Methods included Decision Tree (BASHIR et al., 2016; DELEN; OZTEKIN; TOMAK, 2012; ESFANDIARI et al., 2014), Random Forest (BREIMAN, 2001; PU et al., 2015), Extremely Randomized Tree (GEURTS; ERNST;

WEHENKEL, 2006; SCALZO et al., 2012), AdaBoost (GABER et al., 2016; NAYAK; DASH; MAJHI, 2016; NIE; JIN; FEI, 2014; REDDY; SODHI, 2016), Logistic Regression (BASHIR et al., 2016; MORID et al., 2016), Multinomial Naïve Bayes (BASHIR et al., 2016; ESFANDIARI et al., 2014; MORID et al., 2016) and Support Vector Machines (COHEN; AMBERT; MCDONAGH, 2009; ESFANDIARI et al., 2014; LU et al., 2006; YANG et al., 2009; YANG; KIANG; SHANG, 2015). Table 2.1 provides a summary description of the methods, including references for further reading.

2.2.3 Performance Evaluation

Based on results from the prediction module, performance statistics were calculated for the different combinations of pre-processing strategies and prediction methods. Mean values, standard-deviations, and confidence intervals were calculated for three indicators (Precision, Recall, and F1-score) that had been frequently used in similar studies (AMBERT; COHEN, 2008; HUANG et al., 2005; LU et al., 2006; MORID et al., 2016; OHEN et al., 2006; WELLMER et al., 2007; YANG et al., 2009).

Precision gives the percentage of classified records that are correct (i.e. the ratio between true positives and the sum of true positives and false positives). Recall gives the percentage of correct records in the classification (i.e. the ratio between true positives and the sum of true positives and false negatives). In this study, we estimated the F1-score, which is the weighted harmonic mean of precision and recall estimates (HASTIE; TIBSHIRANI; FRIEDMAN, 2008). The best prediction method combines the pre-processing and prediction module alternatives such that the F1-score is maximized. Whenever combinations of pre-processing and prediction alternatives led to F1-score that were not statistically different, the best combination was the one with the highest mean.

Table 2.1: Text classification methods in study

Method	Description	References
Decision tree	Classifies an entry starting from the root node of a decision tree, applying the first test considering the attribute under consideration. The process is carried out until a leave is found, representing the class to which the entry belongs.	(BASHIR et al., 2016; DELEN; OZTEKIN; TOMAK, 2012; ESFANDIARI et al., 2014)
Random forest	Ensemble classifier comprised of several decision trees. Each tree is recursively generated by running a binary classification test to each non-leaf node, using training set of entries. The final classification of an entry is the most popular among those obtained in all trees.	(BREIMAN, 2001; PU et al., 2015)
Extremely randomized tree (Extra-Tree)	Ensemble classifier that extends the Decision Tree method introducing randomness in tree generation process.	(GEURTS; ERNST; WEHENKEL, 2006; SCALZO et al., 2012)
Adaptive boosting (AdaBoost)	Machine learning method that combines several weak classifiers (i.e. those that perform better when compared to random classifiers) in a single, more accurate ensemble classifier.	(NAYAK; DASH; MAJHI, 2016; PU et al., 2015)
Logistic regression	Probabilistic classification method widely used when the objective is to classify entries in one of two classes. The best logistic regression model is obtained by minimizing a regularized negative log-likelihood function.	(BASHIR et al., 2016; MORID et al., 2016)
Multinomial Naïve Bayes (Multinomial NB)	Classification method based on Bayes theorem. It considers a priori (known) probabilities of entries belonging to one of two classes in training sets. Determining a posteriori probabilities based on most frequent classification, the method defines the final assignment of entries to classes.	(BASHIR et al., 2016; ESFANDIARI et al., 2014; MORID et al., 2016)
Support Vector Machine methods (Linear SVC and nuSVC)	Grounded on statistical learning theory, Support Vector Machine (SVM) methods set principles to be followed towards obtaining generalizable classifiers based on their ability to correctly predict, on the same domain in which learning took place, the class to which new entries belong. Two SVM-based classifiers were considered in this work: Linear Support Vector Classifier (Linear SVC) and its extension nuSVC.	(COHEN; AMBERT; MCDONAGH, 2009; ESFANDIARI et al., 2014; LU et al., 2006; YANG et al., 2009; YANG; KIANG; SHANG, 2015)

2.3 RESULTS

Data included textual medical records from 16,703 patients who visited the hospital ED between January and December 2014. After excluding records with duplicated and empty reports, and with patient final status explicitly described in the text, the number of records reduced to 11,175, of which 5,584 were classified as *discharged* and 5,592 as *inpatient* (Table 2.2).

The ED provided care to high complexity cases, which includes only patients in the immediate (I), very urgent (VU) and urgent (U) categories of the Manchester Triage System (MTS). As seen in Table 2.2, the raw database presented a 48:52 proportion of cases in the inpatient and discharged categories, which shifted to 50:50 after selection criteria were applied. Both raw and treated databases were balanced, reflecting the type of care provided by the ED.

Table 2.3 shows the performance of each classification test. Performance indicators with respect to the *inpatient* class indicated that nuSVC gave the best prediction with the following pre-processing options: feature selection carried out using the F-value, best percentile of 65, and set-of-words using indicator TF-IDF and trigrams. This combination of pre-processing and prediction alternatives resulted in an average F1-score of 77.70% (SD = 0.66%), average Recall of 82.48% (SD = 1.05%), and average Precision of 73.47% (SD = 1.14%). A summary of parameter settings for nuSVC (the best performing method) is given in Table 2.4. Python recommends the best settings for parameters in each classification method based on characteristics of the dataset. We also tested whether the F1-score resulting from the best combination of Pre-Processing and Prediction alternatives was different from F1-scores obtained from all other combinations (Table 2.3).

Table 2.2: Selection of text reports included in the database

Description	Inpatient	Discharged	Total
Number of text reports in the database	8038 (48%)	8665 (52%)	16703
Duplicated reports	-10	-5	-15
Information on final status explicit in report. Examples: HAA (hospital admission authorized), patient left, PD (patient discharged), patient not located, patient did not answer when called.	-191	-2009	-2200
Empty reports. No text between headings "Subjective" and "Prepared and Signed by".	-2246	-1067	-3313
Number of text reports used in the study	5591 (50%)	5584 (50%)	11175

Feature selection influenced the classification performance of methods such as Extra Tree and Random Forest (Table 2.5). With nuSVM, feature selection did not significantly increase the classification performance; however, it minimized the F1-score standard deviation, as shown in Table 2.5, and had an strong positive impact on processing time.

The total number of features depends on the number of *n*-grams used for classification. Considering the database under analysis, if only unigrams were used, there would be 724,856 features; considering unigrams and bigrams, there would be 1,526,642 features; considering unigrams, bigrams and trigrams the total number of features would increase to 2,330,813. After feature selection, the best performing classification method (nuSVM) uses uni-, bi-, and trigrams and only 1,538,956 features; i.e. a 33% reduction.

The most predictive features for admission were, as expected, terms associated with the presence of symptoms (e.g. "edema", "non-measured fev", "paroxysmal nocturnal dyspnea") indicated during physical examination or in the patient referral letter. The most predictive features for discharge were those pertaining to a normal physical examination or the absence of symptoms (such as "well perfused extrem", "normal breath sound", and "no history dysuria").

Table 2.3: Best performance for each classification method

Method	FS	P	SoW	NG	F1-score					Recall				Precision			
					Mean	SD	LB-CI	UB-CI	PV	Mean	SD	LB-CI	UB-CI	Mean	SD	LB-CI	UB-CI
nuSVC	F-score	65	TF-IDF	3	77.70%	0.66%	77.23%	78.18%	1.000	82.48%	1.05%	81.73%	83.23%	73.47%	1.14%	72.65%	74.28%
SVC Linear	F-score	45	TF-IDF	2	77.59%	0.60%	77.16%	78.02%	0.693	79.70%	1.25%	78.80%	80.59%	75.62%	1.21%	74.75%	76.49%
Extra Tree	F-score	30	TF	3	77.03%	1.10%	76.24%	77.82%	0.117	82.04%	1.83%	80.73%	83.35%	72.62%	0.98%	71.91%	73.32%
Logistic Regression	χ^2	65	BR	3	76.95%	0.74%	76.42%	77.48%	0.028	80.16%	1.50%	79.09%	81.23%	74.02%	1.34%	73.06%	74.98%
Random Forest	χ^2	35	BR	3	76.90%	0.97%	76.21%	77.59%	0.046	84.82%	1.39%	83.83%	85.81%	70.35%	1.24%	69.46%	71.23%
Multinomial NB	χ^2	60	BR	2	76.55%	0.91%	75.90%	77.20%	0.005	85.30%	1.71%	84.08%	86.53%	69.44%	0.86%	68.83%	70.05%
AdaBoost	χ^2	1	BR	1	73.72%	0.97%	73.03%	74.41%	0.000	70.86%	1.47%	69.81%	71.90%	76.84%	1.22%	75.97%	77.71%
Decision Tree	χ^2	20	BR	2	67.86%	1.29%	66.93%	68.78%	0.000	68.00%	1.70%	66.79%	69.21%	67.73%	1.18%	66.88%	68.57%

FS: Feature Selection, P: Percentile, SoW: Set-of-Words, NG: n-grams, SD: Standard Deviation, LB-CI: Lower Boundary – Confidence Interval, UB-CI: Upper Boundary – Confidence Interval, PV: p-value

Table 2.4: Parameter settings for the nuSVC method

Parameter	Setting	Description
nu	0.5	An upper bound on the fraction of training errors and a lower bound of the fraction of support vectors. Should be in the interval (0, 1].
kernel	linear	Specifies the kernel type to be used in the algorithm.
probability	False	Whether to enable probability estimates.
shrinking	True	Whether to use the shrinking heuristic.
tol	1e-3	Tolerance for stopping criterion.
class_weight	None	If not given, all classes are supposed to have weight one.
verbose	False	Enable verbose output.
max_iter	-1	Hard limit on iterations within solver, or -1 for no limit.
decision_function_shape	ovr	Whether to return a one-vs-rest ('ovr') decision function of shape (n_samples, n_classes)
random_state	1	The seed of the pseudo random number generator

The worst average performance was with the method Decision Tree, with the following pre-processing options: feature selection carried out using the χ^2 indicator, best percentile of 20, and set-of-words using indicator BR and bigrams. This combination resulted in an average F1-score of 67.86% (SD = 1.29%), average Recall of 68.00% (SD = 1.70%), and average Precision of 67.72% (SD = 1.18%).

Figure 2.2 presents confidence intervals of F1-scores for each combination of pre-processing and prediction alternatives. Visual analysis indicated that intervals associated with the first six methods in the Figure 2.2 overlapped. However, analyzing their corresponding *p*-values (Table 2.2), the hypothesis of different F1-scores was rejected only for nuSVC, SVC Linear, and Extremely Randomized Tree methods. Recall that overlapping confidence intervals

not necessarily imply that corresponding means are not statistically different. That is the case since the distance from the mean is calculated differently for the t-statistic and the mean confidence interval; see (KNEZEVIC, 2008).

Table 2.5: Feature selection impact on the performance of some classifiers

Classifier	Approach	Average	St. Dev.	p-value	Reduction in Processing time
Extra Tree	TF; trigrams; F-value; Percentile 30	77.03%	1.10%	0.000	48.68%
	TF; trigrams; No feature selection	74.02%	1.23%		
Random Forest	BR; trigrams; F-value; Percentile 35	76.90%	0.97%	0.000	34.55%
	BR; trigrams; No feature selection	73.45%	1.39%		
nuSVC	TF-IDF; trigrams; F-value; Percentile 65	77.70%	0.66%	0.242	33.33%
	TF-IDF; trigrams; No feature selection	77.29%	0.87%		
Logistic Regression	BR; trigrams; χ^2; Percentile 65	76.95%	0.74%	0.029	7.69%
	BR; trigrams; No feature selection	75.86%	1.25%		

The nuSVC method with the pre-processing options previously described was the best (based on F1-score) in predicting hospitalizations using ED records. However, several alternative combinations of pre-processing and prediction methods resulted in F1-scores that were not statistically different from our choice (see Appendix).

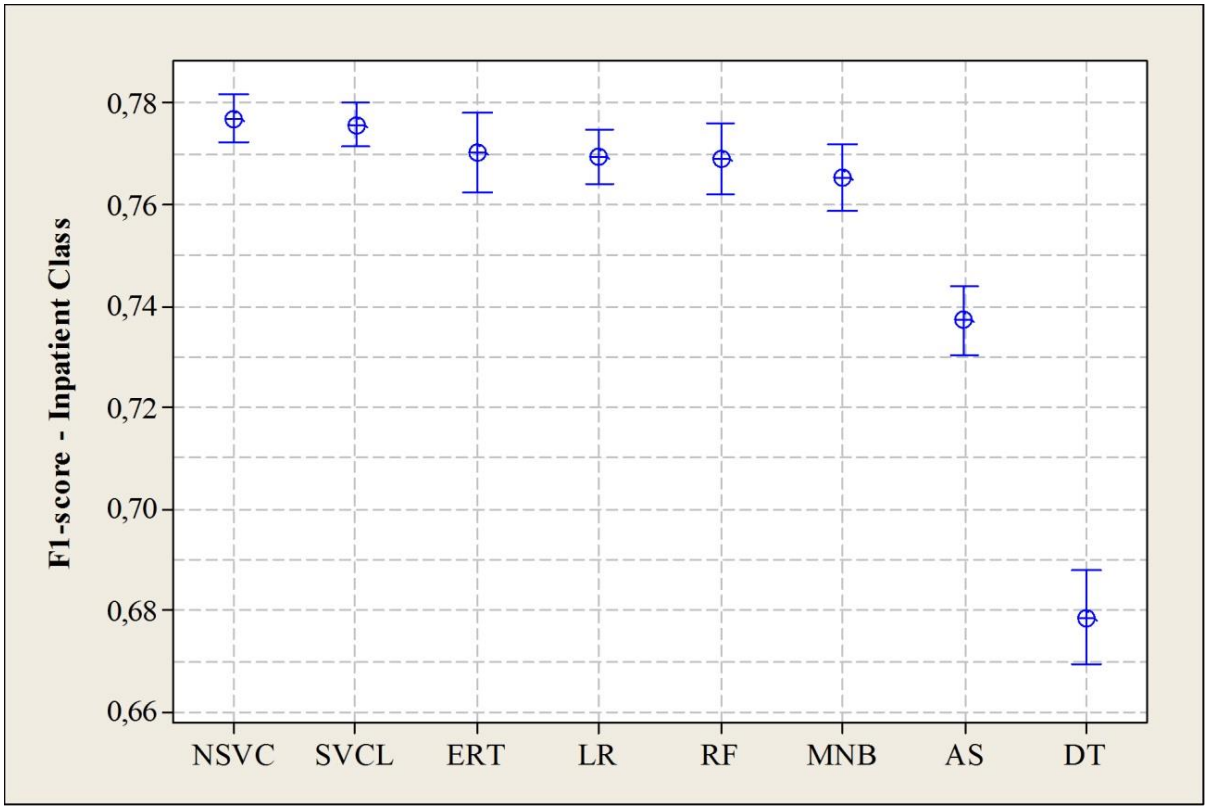

Figure 2.2: Comparison of F1-score 95% confidence intervals for different prediction methods

NSVC: nuSVC, SVCL: SVC Liner, ERT: Extremely Randomized Tree, LR: Logistic Regression, RF: Random Forest, MNB: Multinomial Naïve Bayes, AB: AdaBoost, DT: Decision Tree.

2.4 DISCUSSION

To the best of our knowledge, this is the first study that used TM to estimate bed demand based on medical records generated in the first patient-doctor encounter at an ED, even before information from laboratory exams was available. Further, this is one of the few existing studies that compared pre-processing and classification strategies using ED medical text records (MOWERY et al., 2012; ST-MAURICE; KUO, 2012; ZMIRI; SHAHAR; TAIEB-MAIMON, 2012).

Mowery *et al*. (2012) used TM to build an automated SOAP classifier, which allowed mapping passages from free form clinical text narratives onto one of the classes of the SOAP framework based on ED clinical reports. For that, they trained a support vector machine-based

classifier using ED text records that were previously classified by human experts. In a ten-fold cross validation experiment, they obtained an average F1-score of 85.30% in their classifications. Building up from Mowery *et al*.'s efforts (MOWERY et al., 2012), our results show the potential benefits of adopting a TM approach in a chronically overcrowded ED with a large number of low-acuity patients (data not shown) and elevated boarding times. Although using the SOAP framework was not mandatory in the ED, it was still used in most patient clinical records; this could have contributed to the success of TM in predicting resource utilization.

In terms of practical implications, the study provides an effective tool for ED managers to analyze the commonly abundant text information from medical reports, leading to better management of a healthcare operations system.

The database used included initial patient assessments by emergency clinicians using the SOAP structure, before laboratory or image test results were available. Mining the database, we found that the decision to admit or discharge patients was made on average 6 hours and 18 minutes after the SOAP report was entered in the ED system. So, at that point the clinicians did not have yet much information to support their decision.

The ED often rerouted patients to other hospitals once its bed capacity was exhausted. However, some might stay and wait for beds to be vacated by inpatients nearing discharge. In this scenario, accurate prediction of inward bed demand can help decision making in at least two ways. First, estimating how many inpatients will be admitted at the end of the day may help to decide how many new patients should stay and wait for available beds. Second, inward bed capacity planning contemplating demand from the ED can be completed at least one shift earlier using the TM system.

Some remarks on the nature of the text reports used in our study and difficulties that may arise from their use in practice are also noteworthy. First, all text reports analyzed used the

SOAP framework, which includes some standard information. Thus, our results may be generalizable to similar application contexts. The SOAP framework has been used in most medical schools and allied hospitals (MAKI; PETTERSON, 2013). It is considered a standard tool for reporting in many areas including nursing, psychiatry, and emergency medicine (BORCHERDING; MORREALE, 2007; CAMERON; TURTLE-SONG, 2002; KOZIER et al., 1993; MOWERY et al., 2012). Second, we tested our method in a high complexity, overcrowded ED. In such context, clinicians' reports are likely to be shortened due to time constraints. Based on the positive results in this scenario, we conjecture that the method might be equally or even more successful under more favorable circumstances.

Implementing our method in real-time requires adherence to the use of standardized sentences to describe events that anticipate the final status of the patient (e.g. "patient is hospitalized", "discharged", "not found when called for consultation", or "departed from the premises"); that is the case of the ED under analysis. Whenever those standardized features appear in the reports, the classifier automatically directs them to one of the two classes. Clinicians thus should include discharge or admission information in their reports, using the standardized sentences prepared for that, since the method may be programmed to recognize reports containing those sentences as pertaining to one class or the other, with 100% confidence. Finally, it should be noted that quality inspection of texts is not feasible in a real-time application, and that the method may perform poorly classifying reports presenting typos. However, this problem did not appear to be frequent in the study database, and could be further reduced by instructing clinicians to run a spell check before uploading reports to the ED system.

The study has some limitations that may be overcome in future extensions. Few annotation studies in the literature included the SOAP framework in EDs, which could make it difficult to compare our results with theirs. Still, our analysis included a very large sample, which minimized the reliability risks associated with studies of smaller samples. Future

research should provide further insights on the contribution of the SOAP framework's embodied temporal progression, especially in the context of estimating resources for patients with chronic diseases presenting at overcrowded EDs with low acuity demands.

2.5 REFERENCES

ALIJANI, A. et al. Instrument for objective assessment of appropriateness of surgical bed occupancy : validation study Representation of South Asian people in randomised clinical trials : analysis of trials ' data. p. 1243–1244, 2003.

AMBERT, K. H.; COHEN, A. M. A system for classifying disease comorbidity status from medical discharge summaries using automated hotspot and negated concept detection. **Journal of the American Medical Informatics Association : JAMIA**, v. 16, n. 4, p. 590–5, 2008.

AYVAZ, S. et al. Toward a complete dataset of drug-drug interaction information from publicly available sources. **Journal of biomedical informatics**, v. 55, p. 206–17, jun. 2015.

BASHIR, S. et al. HMV: A medical decision support framework using multi-layer classifiers for disease prediction. **Journal of Computational Science**, v. 13, p. 10–25, mar. 2016.

BIRD, S.; KLEIN, E.; LOPER, E. **Natural Language Processing with Python**. 1º edition ed. Sebastopol, CA: O'Reilly Media, 2009.

BIRD, S. R. et al. Integrated care facilitation for older patients with complex health care needs reduces hospital demand. v. 31, n. 3, p. 451–461, 2007.

BORCHERDING, S.; MORREALE, M. J. **The OTA's guide to writing SOAP notes**. New Jersey: Slack Incorporated, 2007.

BREIMAN, L. E. O. Random Forests. p. 5–32, 2001.

CAMERON, S.; TURTLE-SONG, I. Learning to write case notes using the SOAP format. **Journal of Counseling and Development: JCD**, v. 80, n. 3, p. 286, 2002.

CHEN, E. S. et al. Automated Acquisition of Disease – Drug Knowledge from Biomedical and Clinical Documents : An Initial Study. v. 15, n. 1, p. 87–98, 2008.

CHEN, Y.; LIN, C. Combining SVMs with Various Feature Selection Strategies. In: **Studies in Fuzziness and Soft Computing**. [s.l.] Springer Berlin Heidelberg, 2006. p. 315–324.

COHEN, A. M.; AMBERT, K.; MCDONAGH, M. Cross-topic learning for work prioritization in systematic review creation and update. **Journal of the American Medical Informatics Association : JAMIA**, v. 16, n. 5, p. 690–704, 2009.

DELEN, D.; OZTEKIN, A.; TOMAK, L. An analytic approach to better understanding and management of coronary surgeries. **Decision Support Systems**, v. 52, n. 3, p. 698–705, fev. 2012.

DINAKARPADIAN, R. L. D.; LEE, Y.; VISHWANATH, K. MachineProse : An Ontological Framework for Scientific Assertions. **Journal of the American Medical Informatics Association : JAMIA**, v. 13, p. 220–232, 2006.

ESFANDIARI, N. et al. Knowledge discovery in medicine: Current issue and future trend. **Expert Systems with Applications**, v. 41, n. 9, p. 4434–4463, jul. 2014.

FARMER, R. D. T.; EMAMI, J. Models for forecasting hospital bed requirements in the acute sector. v. 365, n. 1, p. 307–312, 1990.

FELDMANN, R.; SANGER, J. **The text mining handbook: advanced approaches in analyzing unstructured data**. New York: Cambridge university press, 2007.

GABER, T. et al. Biometric cattle identification approach based on Weber's Local Descriptor and AdaBoost classifier. **Computers and Electronics in Agriculture**, v. 122, p. 55–66, mar. 2016.

GARCÍA-GUBERN, C.; COLÓN-ROLÓN, L.; APONTE, M. M. Workforce projections for emergency medicine in Puerto Rico: a five-year follow-up of an evident demand. **The Journal of emergency medicine**, v. 29, n. 1, p. 107–10, jul. 2005.

GENES, N. et al. Validating Emergency Department Vital Signs Using a Data Quality Engine for Data Warehouse. p. 34–39, 2013.

GEURTS, P.; ERNST, D.; WEHENKEL, L. Extremely randomized trees. **Machine Learning**, v. 63, n. 1, p. 3–42, 2 mar. 2006.

HACKELING, G. **Mastering Machine Learning With scikit-learn**. Birmingham, UK: Packt Publishing Ltd., 2014.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning: data mining, inference, and prediction**. Second edi ed. [s.l.] Springer, 2008.

HOOT, N. R.; ARONSKY, D. Systematic review of emergency department crowding: causes,

effects, and solutions. **Annals of emergency medicine**, v. 52, n. 2, p. 126–36, ago. 2008.

HUANG, Y. et al. Improved Identification of Noun Phrases in Clinical Radiology Reports Using a High-Performance Statistical Natural Language Parser Augmented with the UMLS Specialist Lexicon. **Journal of the American Medical Informatics Association : JAMIA**, v. 12, n. 3, p. 275–285, 2005.

ILANGO, N. R. . S. **A Hybrid Prediction Model with F-score Feature Selection for Type II Diabetes Databases**. 2CWiC Amrita ACM-W Celebration on Women in Computing in India. **Anais**...2010

KARYSTIANIS, G.; BUCHAN, I.; NENADIC, G. Mining characteristics of epidemiological studies from Medline: a case study in obesity. **Journal of biomedical semantics**, v. 5, n. 1, p. 22, jan. 2014.

KNEZEVIC, A. **Overlapping Confidence Intervals and Statistical Significance**. [s.l: s.n.].

KOZIER, B. et al. **Techniques in clinical nursing**. 4th ed ed. New York: Addison-Wesley, 1993.

LU, X. et al. Enhancing text categorization with semantic-enriched representation and training data augmentation. **Journal of the American Medical Informatics Association : JAMIA**, v. 13, n. 5, p. 526–35, 2006.

LUCAS, R. et al. Emergency department patient flow: the influence of hospital census variables on emergency department length of stay. **Academic emergency medicine : official journal of the Society for Academic Emergency Medicine**, v. 16, n. 7, p. 597–602, jul. 2009.

MACKAY, M. Choice of Models for the Analysis and Forecasting of Hospital Beds. p. 221–230, 2005.

MAKI, S. E.; PETTERSON, B. **Using the electronic health record in the health care provider practice**. [s.l.] Nelson Education, 2013.

MEESAD, P.; BOONRAWD, P.; NUIPIAN, V. A Chi-Square-Test for Word Importance Differentiation in Text Classification. v. 6, p. 110–114, 2011.

MICHELSON, J. D.; PARISEAU, J. S.; PAGANELLI, W. C. Assessing surgical site infection risk factors using electronic medical records and text mining. **American journal of infection control**, v. 42, n. 3, p. 333–6, mar. 2014.

MORID, M. A. et al. Classification of Clinically Useful Sentences in Clinical Evidence

Resources. **Journal of biomedical informatics**, v. 60, p. 14–22, 13 jan. 2016.

MOWERY, D. et al. Building an automated SOAP classifier for emergency department reports. **Journal of biomedical informatics**, v. 45, n. 1, p. 71–81, fev. 2012.

NAPHHS. **Perfecting Patient Flow**. [s.l: s.n.].

NAYAK, D. R.; DASH, R.; MAJHI, B. Brain MR image classification using two-dimensional discrete wavelet transform and AdaBoost with random forests. **Neurocomputing**, v. 177, p. 188–197, fev. 2016.

NIE, Q.; JIN, L.; FEI, S. Probability estimation for multi-class classification using AdaBoost. **Pattern Recognition**, v. 47, n. 12, p. 3931–3940, dez. 2014.

OECD. **Hospital Beds**. Health at a Glance. **Anais**...2011

OHEN, A. M. C. et al. Reducing Workload in Systematic Review Preparation Using Automated Citation Classification. **Journal of the American Medical Informatics Association : JAMIA**, v. 13, p. 206–219, 2006.

PEREIRA, L. et al. ICD9-based Text Mining Approach to Children Epilepsy Classification. **Procedia Technology**, v. 9, p. 1351–1360, 2013.

PU, X. et al. Facial expression recognition from image sequences using twofold random forest classifier. **Neurocomputing**, v. 168, p. 1173–1180, nov. 2015.

PYTHON SOFTWARE FOUNDATION. **The Python Language Reference**, 2017. Disponível em: <https://www.python.org/>

REDDY, M. V.; SODHI, R. A rule-based S-Transform and AdaBoost based approach for power quality assessment. **Electric Power Systems Research**, v. 134, p. 66–79, maio 2016.

SCALZO, F. et al. Intracranial hypertension prediction using extremely randomized decision trees. **Medical engineering & physics**, v. 34, n. 8, p. 1058–65, out. 2012.

SCHMIDT, M. I. et al. Chronic non-communicable diseases in Brazil: burden and current challenges. **Lancet (London, England)**, v. 377, n. 9781, p. 1949–61, 4 jun. 2011.

SOLT, I. et al. Semantic classification of diseases in discharge summaries using a context-aware rule-based classifier. **Journal of the American Medical Informatics Association : JAMIA**, v. 16, n. 4, p. 580–4, 2009.

SPASIĆ, I. et al. Text mining of cancer-related information: review of current status and future directions. **International journal of medical informatics**, v. 83, n. 9, p. 605–23, set. 2014.

SPRIVULIS, P. C. et al. The association between hospital overcrowding and mortality among patients admitted via western australian emergency departments. 2006.

ST-MAURICE, J.; KUO, M. H. Analyzing Primary Care Data to Characterize Inappropriate Emergency Room Use. p. 990–994, 2012.

SWEENEY, T. K.; ASHLEY, S. A. Forecasting hospital bed needs. v. 283, n. July, 1981.

UYSAL, A. K.; GUNAL, S. A novel probabilistic feature selection method for text classification. **Knowledge-Based Systems**, v. 36, p. 226–235, dez. 2012.

VALLMUUR, K. Machine learning approaches to analysing textual injury surveillance data: a systematic review. **Accident; analysis and prevention**, v. 79, p. 41–9, jun. 2015.

WANG, X. et al. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. **Journal of the American Medical Informatics Association : JAMIA**, v. 16, n. 3, p. 328–37, 2009.

WELLMER, L. H. BEN et al. Rapidly Retargetable Approaches to De-identification in Medical Records. **Journal of the American Medical Informatics Association : JAMIA**, v. 14, p. 564–573, 2007.

WILER, J. L. et al. Predictors of patient length of stay in 9 emergency departments. **The American journal of emergency medicine**, v. 30, n. 9, p. 1860–4, nov. 2012.

YANG, H. et al. A text mining approach to the prediction of disease status from clinical discharge summaries. **Journal of the American Medical Informatics Association : JAMIA**, v. 16, n. 4, p. 596–600, 2009.

YANG, M.; KIANG, M.; SHANG, W. Filtering big data from social media--Building an early warning system for adverse drug reactions. **Journal of biomedical informatics**, v. 54, p. 230–40, abr. 2015.

YONGJING, Y. L.; LI, W.; CHEN, K. A Document Clustering and Ranking System for Exploring MEDLINE Citations. **Journal of the American Medical Informatics Association : JAMIA**, v. 14, n. 5, p. 651–661, 2007.

ZMIRI, D.; SHAHAR, Y.; TAIEB-MAIMON, M. Classification of patients by severity grades during triage in the emergency department using data mining methods. **Journal of evaluation in clinical practice**, v. 18, n. 2, p. 378–88, abr. 2012.

2.6 APPENDIX

Table 2.6: Complete list of alternative combinations of pre-processing and prediction methods that yielded *F*1 values statistically not different from our elected method

Algorithm	FS	P	SoW	NG	F1-score					Recall				Precision			
					Mean	SD	LB-CI	UB-CI	PV	Mean	SD	LB-CI	UB-CI	Mean	SD	LB-CI	UB-CI
NuSVC	F-value	65	TF-IDF	3	77.70%	0.66%	77.23%	78.18%	1.000	82.48%	1.05%	81.73%	83.23%	73.47%	1.14%	72.65%	74.28%
NuSVC	Chi2	80	TF-IDF	3	77.70%	0.69%	77.21%	78.19%	0.988	80.30%	1.49%	79.24%	81.37%	75.28%	1.03%	74.55%	76.02%
NuSVC	Chi2	85	TF-IDF	3	77.69%	0.83%	77.10%	78.29%	0.971	80.09%	1.69%	78.88%	81.30%	75.46%	1.02%	74.73%	76.19%
NuSVC	F-value	75	TF-IDF	3	77.66%	0.73%	77.13%	78.18%	0.881	81.39%	1.32%	80.45%	82.34%	74.27%	1.27%	73.37%	75.18%
NuSVC	F-value	80	TF-IDF	3	77.66%	0.83%	77.06%	78.25%	0.887	80.52%	1.33%	79.57%	81.47%	75.02%	1.35%	74.05%	75.98%
NuSVC	Chi2	50	TF-IDF	3	77.64%	0.55%	77.24%	78.04%	0.816	81.07%	1.32%	80.12%	82.02%	74.51%	1.11%	73.72%	75.30%
NuSVC	Chi2	70	TF-IDF	3	77.64%	0.54%	77.25%	78.03%	0.809	80.70%	1.35%	79.73%	81.66%	74.83%	0.98%	74.13%	75.52%
NuSVC	F-value	85	TF-IDF	3	77.63%	0.82%	77.04%	78.22%	0.825	80.41%	1.55%	79.30%	81.52%	75.06%	1.14%	74.24%	75.88%
NuSVC	F-value	55	TF-IDF	3	77.63%	0.40%	77.34%	77.91%	0.755	83.55%	1.28%	82.64%	84.47%	72.50%	0.85%	71.89%	73.11%
NuSVC	F-value	25	TF-IDF	3	77.62%	0.55%	77.22%	78.01%	0.749	82.96%	1.06%	82.20%	83.72%	72.93%	1.00%	72.21%	73.65%
NuSVC	F-value	70	TF-IDF	3	77.61%	0.75%	77.07%	78.14%	0.763	81.59%	1.36%	80.62%	82.56%	74.03%	1.42%	73.01%	75.04%
NuSVC	Chi2	65	TF-IDF	3	77.60%	0.60%	77.18%	78.03%	0.726	80.91%	1.42%	79.90%	81.92%	74.58%	1.02%	73.85%	75.31%
NuSVC	F-value	50	TF-IDF	3	77.60%	0.45%	77.28%	77.92%	0.679	84.21%	1.13%	83.41%	85.02%	71.96%	0.83%	71.37%	72.55%
SVC_linear	F-value	45	TF-IDF	2	77.59%	0.60%	77.16%	78.02%	0.693	79.70%	1.25%	78.80%	80.59%	75.62%	1.21%	74.75%	76.49%
NuSVC	Chi2	90	TF-IDF	3	77.59%	1.01%	76.86%	78.31%	0.766	79.91%	2.04%	78.46%	81.37%	75.43%	0.97%	74.74%	76.12%
NuSVC	F-value	30	TF-IDF	3	77.57%	0.41%	77.28%	77.86%	0.601	83.29%	0.90%	82.64%	83.93%	72.61%	0.91%	71.96%	73.26%
NuSVC	Chi2	75	TF-IDF	3	77.57%	0.70%	77.06%	78.07%	0.655	80.39%	1.47%	79.34%	81.45%	74.96%	1.12%	74.16%	75.76%
NuSVC	F-value	45	TF-IDF	3	77.56%	0.36%	77.30%	77.81%	0.543	84.88%	0.99%	84.17%	85.58%	71.41%	0.81%	70.83%	71.99%
NuSVC	F-value	60	TF-IDF	3	77.55%	0.48%	77.20%	77.89%	0.555	82.98%	1.01%	82.26%	83.70%	72.80%	0.98%	72.10%	73.50%
SVC_linear	F-value	55	TF-IDF	2	77.54%	0.65%	77.07%	78.00%	0.571	79.36%	1.37%	78.38%	80.34%	75.82%	1.13%	75.02%	76.63%
NuSVC	F-value	35	TF-IDF	3	77.53%	0.36%	77.27%	77.79%	0.471	83.79%	1.09%	83.01%	84.57%	72.16%	0.85%	71.55%	72.76%
NuSVC	Chi2	95	TF-IDF	3	77.53%	0.89%	76.89%	78.16%	0.617	79.89%	1.80%	78.60%	81.18%	75.32%	0.86%	74.71%	75.93%
NuSVC	F-value	40	TF-IDF	3	77.53%	0.24%	77.35%	77.70%	0.440	84.39%	1.10%	83.61%	85.18%	71.71%	0.84%	71.11%	72.31%
NuSVC	Chi2	55	TF-IDF	3	77.52%	0.49%	77.17%	77.87%	0.479	81.14%	1.27%	80.24%	82.05%	74.22%	1.04%	73.48%	74.97%

NuSVC	F-value	50	TF-IDF	2	77.52%	0.58%	77.10%	77.93%	0.507	80.68%	1.31%	79.74%	81.61%	74.62%	1.15%	73.80%	75.44%
SVC_linear	F-value	80	TF-IDF	3	77.52%	0.74%	76.98%	78.05%	0.556	79.86%	1.48%	78.80%	80.92%	75.33%	1.03%	74.59%	76.07%
NuSVC	F-value	90	TF-IDF	3	77.51%	0.70%	77.01%	78.01%	0.530	79.68%	1.60%	78.54%	80.82%	75.48%	1.00%	74.77%	76.20%
NuSVC	F-value	45	TF-IDF	2	77.51%	0.41%	77.22%	77.80%	0.437	80.98%	1.16%	80.15%	81.81%	74.34%	1.04%	73.60%	75.08%
SVC_linear	F-value	40	TF-IDF	2	77.51%	0.60%	77.08%	77.94%	0.494	79.46%	1.28%	78.55%	80.38%	75.67%	1.27%	74.76%	76.58%
SVC_linear	F-value	50	TF-IDF	3	77.50%	0.70%	77.00%	78.00%	0.515	81.34%	1.47%	80.29%	82.39%	74.03%	0.88%	73.40%	74.66%
SVC_linear	F-value	55	TF-IDF	3	77.50%	0.74%	76.97%	78.02%	0.514	81.09%	1.53%	79.99%	82.18%	74.23%	1.00%	73.51%	74.95%
NuSVC	F-value	55	TF-IDF	2	77.50%	0.56%	77.10%	77.89%	0.456	80.38%	1.33%	79.43%	81.32%	74.84%	1.04%	74.09%	75.58%
SVC_linear	F-value	70	TF-IDF	3	77.50%	0.74%	76.97%	78.02%	0.513	80.38%	1.42%	79.36%	81.39%	74.84%	1.05%	74.09%	75.58%
NuSVC	Chi2	60	TF-IDF	3	77.49%	0.57%	77.09%	77.90%	0.453	80.82%	1.24%	79.93%	81.71%	74.45%	1.07%	73.69%	75.21%
SVC_linear	F-value	50	TF-IDF	2	77.49%	0.60%	77.07%	77.92%	0.463	79.41%	1.36%	78.44%	80.39%	75.69%	1.12%	74.89%	76.50%
SVC_linear	F-value	90	TF-IDF	3	77.47%	0.81%	76.89%	78.05%	0.493	79.48%	1.65%	78.30%	80.66%	75.59%	1.02%	74.86%	76.32%
SVC_linear	F-value	85	TF-IDF	3	77.47%	0.84%	76.87%	78.07%	0.496	79.71%	1.66%	78.53%	80.90%	75.37%	1.02%	74.64%	76.10%
SVC_linear	F-value	65	TF-IDF	3	77.47%	0.77%	76.92%	78.02%	0.470	80.68%	1.43%	79.66%	81.70%	74.52%	1.03%	73.79%	75.26%
NuSVC	Chi2	15	TF-IDF	3	77.46%	0.59%	77.04%	77.88%	0.397	80.59%	1.37%	79.61%	81.57%	74.59%	1.00%	73.88%	75.30%
NuSVC	F-value	20	TF-IDF	3	77.46%	0.42%	77.15%	77.76%	0.334	82.13%	1.11%	81.33%	82.92%	73.31%	0.87%	72.68%	73.93%
SVC_linear	F-value	75	TF-IDF	3	77.45%	0.74%	76.93%	77.98%	0.435	80.20%	1.52%	79.11%	81.29%	74.92%	0.92%	74.26%	75.57%
SVC_linear	F-value	45	TF-IDF	3	77.45%	0.64%	76.99%	77.91%	0.397	81.59%	1.35%	80.62%	82.55%	73.73%	0.84%	73.13%	74.33%
NuSVC	Chi2	45	TF-IDF	3	77.45%	0.59%	77.03%	77.87%	0.375	80.89%	1.19%	80.04%	81.75%	74.31%	1.05%	73.56%	75.05%
SVC_linear	Chi2	100	TF-IDF	3	77.44%	0.68%	76.95%	77.93%	0.399	79.41%	1.50%	78.34%	80.48%	75.60%	1.07%	74.83%	76.37%
SVC_linear	F-value	100	TF-IDF	3	77.44%	0.68%	76.95%	77.93%	0.399	79.41%	1.50%	78.34%	80.48%	75.60%	1.07%	74.83%	76.37%
SVC_linear	F-value	75	TF-IDF	2	77.44%	0.74%	76.91%	77.97%	0.408	78.59%	1.64%	77.41%	79.77%	76.36%	1.26%	75.46%	77.26%
SVC_linear	F-value	60	TF-IDF	2	77.44%	0.65%	76.97%	77.90%	0.376	79.07%	1.43%	78.05%	80.10%	75.90%	1.17%	75.06%	76.74%
SVC_linear	F-value	60	TF-IDF	3	77.43%	0.71%	76.92%	77.94%	0.381	80.84%	1.43%	79.82%	81.86%	74.32%	1.00%	73.60%	75.03%
NuSVC	Chi2	30	TF-IDF	3	77.43%	0.60%	77.00%	77.86%	0.342	80.77%	1.33%	79.81%	81.72%	74.38%	1.05%	73.62%	75.13%
SVC_linear	Chi2	75	TF-IDF	3	77.43%	0.65%	76.96%	77.89%	0.357	79.82%	1.47%	78.77%	80.87%	75.20%	1.11%	74.41%	75.99%
SVC_linear	F-value	30	TF-IDF	2	77.43%	0.57%	77.02%	77.84%	0.328	79.11%	1.45%	78.07%	80.14%	75.85%	1.28%	74.94%	76.76%
NuSVC	F-value	95	TF-IDF	3	77.41%	0.80%	76.83%	77.98%	0.380	79.71%	1.64%	78.54%	80.89%	75.26%	0.92%	74.60%	75.91%
SVC_linear	F-value	40	TF-IDF	3	77.40%	0.67%	76.93%	77.88%	0.324	81.36%	1.21%	80.49%	82.22%	73.83%	0.93%	73.17%	74.49%
SVC_linear	Chi2	85	TF-IDF	3	77.40%	0.70%	76.90%	77.89%	0.323	79.46%	1.46%	78.42%	80.51%	75.46%	1.05%	74.70%	76.21%
SVC_linear	Chi2	55	TF-IDF	3	77.39%	0.74%	76.86%	77.92%	0.333	80.30%	1.60%	79.16%	81.45%	74.71%	1.10%	73.93%	75.50%
SVC_linear	F-value	25	TF-IDF	2	77.39%	0.66%	76.92%	77.86%	0.298	78.88%	1.49%	77.81%	79.94%	75.99%	1.27%	75.08%	76.90%
SVC_linear	Chi2	90	TF-IDF	3	77.38%	0.88%	76.75%	78.01%	0.365	79.41%	1.77%	78.15%	80.68%	75.48%	0.99%	74.77%	76.19%
SVC_linear	Chi2	65	TF-IDF	3	77.38%	0.67%	76.89%	77.86%	0.283	79.98%	1.62%	78.82%	81.14%	74.96%	1.14%	74.15%	75.78%
NuSVC	F-value	60	TF-IDF	2	77.37%	0.60%	76.94%	77.80%	0.254	79.95%	1.31%	79.01%	80.88%	74.99%	1.31%	74.05%	75.92%

SVC_linear	F-value	70	TF-IDF	2	77.37%	0.68%	76.88%	77.86%	0.282	78.71%	1.59%	77.58%	79.85%	76.11%	1.17%	75.27%	76.95%
SVC_linear	F-value	35	TF-IDF	2	77.37%	0.73%	76.85%	77.89%	0.298	79.16%	1.50%	78.09%	80.23%	75.69%	1.28%	74.78%	76.60%
SVC_linear	F-value	65	TF-IDF	2	77.37%	0.71%	76.86%	77.87%	0.284	78.84%	1.59%	77.70%	79.98%	75.98%	1.22%	75.10%	76.86%
NuSVC	F-value	15	TF-IDF	3	77.37%	0.31%	77.14%	77.59%	0.165	81.18%	1.11%	80.39%	81.97%	73.91%	0.76%	73.36%	74.45%
SVC_linear	F-value	80	TF-IDF	2	77.37%	0.90%	76.72%	78.01%	0.351	78.38%	1.64%	77.20%	79.55%	76.41%	1.38%	75.43%	77.40%
SVC_linear	Chi2	60	TF-IDF	3	77.36%	0.76%	76.82%	77.90%	0.298	80.05%	1.64%	78.88%	81.23%	74.88%	1.18%	74.04%	75.72%
NuSVC	F-value	65	TF-IDF	2	77.36%	0.62%	76.92%	77.80%	0.245	79.59%	1.26%	78.69%	80.49%	75.29%	1.43%	74.26%	76.31%
NuSVC	F-value	40	TF-IDF	2	77.36%	0.52%	76.99%	77.73%	0.212	80.61%	1.35%	79.64%	81.57%	74.39%	1.05%	73.64%	75.14%
SVC_linear	F-value	85	TF-IDF	2	77.36%	0.83%	76.76%	77.95%	0.313	78.30%	1.53%	77.21%	79.40%	76.46%	1.29%	75.54%	77.39%
SVC_linear	Chi2	80	TF-IDF	3	77.35%	0.59%	76.93%	77.77%	0.226	79.57%	1.37%	78.59%	80.55%	75.28%	1.15%	74.46%	76.11%
SVC_linear	F-value	15	TF-IDF	3	77.34%	0.86%	76.72%	77.95%	0.300	79.68%	1.83%	78.37%	80.99%	75.16%	0.87%	74.53%	75.78%
SVC_linear	Chi2	40	TF-IDF	3	77.33%	0.78%	76.78%	77.89%	0.266	79.98%	1.74%	78.73%	81.23%	74.89%	1.20%	74.04%	75.75%
SVC_linear	Chi2	50	TF-IDF	3	77.33%	0.68%	76.84%	77.82%	0.229	80.18%	1.62%	79.02%	81.34%	74.71%	1.09%	73.93%	75.49%
SVC_linear	F-value	35	TF-IDF	3	77.33%	0.62%	76.88%	77.77%	0.206	80.98%	1.35%	80.02%	81.95%	74.01%	0.92%	73.35%	74.67%
NuSVC	F-value	35	TF-IDF	2	77.32%	0.52%	76.95%	77.69%	0.170	80.32%	1.53%	79.23%	81.41%	74.57%	1.10%	73.79%	75.35%
SVC_linear	Chi2	85	TF-IDF	2	77.32%	0.58%	76.91%	77.74%	0.188	78.34%	1.32%	77.40%	79.28%	76.36%	1.14%	75.54%	77.18%
SVC_linear	Chi2	75	TF-IDF	2	77.32%	0.50%	76.96%	77.67%	0.158	78.50%	1.30%	77.57%	79.43%	76.20%	1.12%	75.40%	77.00%
SVC_linear	F-value	30	TF-IDF	3	77.32%	0.61%	76.88%	77.75%	0.189	80.71%	1.40%	79.71%	81.72%	74.21%	0.85%	73.60%	74.82%
SVC_linear	F-value	20	TF-IDF	2	77.31%	0.62%	76.87%	77.76%	0.190	78.61%	1.58%	77.48%	79.74%	76.10%	1.21%	75.24%	76.96%
NuSVC	Chi2	40	TF-IDF	3	77.30%	0.71%	76.79%	77.81%	0.204	80.75%	1.25%	79.86%	81.64%	74.15%	1.07%	73.38%	74.92%
NuSVC	Chi2	35	TF-IDF	3	77.30%	0.65%	76.83%	77.76%	0.181	80.66%	1.61%	79.51%	81.81%	74.23%	0.98%	73.53%	74.93%
SVC_linear	Chi2	70	TF-IDF	3	77.30%	0.79%	76.73%	77.86%	0.223	79.71%	1.64%	78.54%	80.89%	75.04%	0.99%	74.33%	75.75%
SVC_linear	F-value	25	TF-IDF	3	77.29%	0.77%	76.74%	77.84%	0.211	80.27%	1.75%	79.01%	81.52%	74.54%	0.75%	74.01%	75.08%
NuSVC	Chi2	100	TF-IDF	3	77.29%	0.87%	76.66%	77.91%	0.243	79.43%	1.77%	78.16%	80.70%	75.29%	1.25%	74.40%	76.18%
NuSVC	F-value	100	TF-IDF	3	77.29%	0.87%	76.66%	77.91%	0.243	79.43%	1.77%	78.16%	80.70%	75.29%	1.25%	74.40%	76.18%
SVC_linear	Chi2	80	TF-IDF	2	77.29%	0.56%	76.89%	77.68%	0.143	78.45%	1.43%	77.42%	79.47%	76.19%	1.15%	75.37%	77.01%
SVC_linear	Chi2	15	TF-IDF	2	77.28%	0.82%	76.69%	77.87%	0.220	78.23%	1.48%	77.17%	79.29%	76.38%	1.43%	75.36%	77.40%
SVC_linear	Chi2	95	TF-IDF	2	77.28%	0.75%	76.74%	77.82%	0.196	78.18%	1.54%	77.08%	79.28%	76.43%	1.13%	75.62%	77.24%
SVC_linear	F-value	15	TF-IDF	2	77.28%	0.66%	76.81%	77.74%	0.162	78.41%	1.63%	77.24%	79.58%	76.21%	1.16%	75.38%	77.04%
SVC_linear	Chi2	45	TF-IDF	3	77.28%	0.76%	76.73%	77.82%	0.195	80.02%	1.66%	78.83%	81.21%	74.75%	1.12%	73.94%	75.55%
SVC_linear	Chi2	70	TF-IDF	2	77.27%	0.53%	76.89%	77.65%	0.124	78.46%	1.29%	77.54%	79.39%	76.14%	1.23%	75.26%	77.02%
SVC_linear	Chi2	40	TF-IDF	2	77.27%	0.53%	76.89%	77.65%	0.120	78.57%	1.33%	77.62%	79.52%	76.03%	1.09%	75.25%	76.81%
SVC_linear	Chi2	95	TF-IDF	3	77.26%	0.64%	76.80%	77.72%	0.146	79.32%	1.51%	78.24%	80.40%	75.33%	0.87%	74.71%	75.95%
NuSVC	Chi2	70	TF-IDF	2	77.26%	0.59%	76.84%	77.68%	0.129	78.88%	1.41%	77.87%	79.88%	75.74%	1.23%	74.86%	76.62%
SVC_linear	F-value	95	TF-IDF	2	77.26%	0.70%	76.76%	77.76%	0.162	78.13%	1.45%	77.09%	79.16%	76.44%	1.02%	75.71%	77.17%

SVC_linear	F-value	20	TF-IDF	3	77.26%	0.77%	76.71%	77.81%	0.183	79.89%	1.75%	78.64%	81.14%	74.82%	0.90%	74.18%	75.46%
NuSVC	F-value	70	TF-IDF	2	77.25%	0.60%	76.82%	77.67%	0.120	79.23%	1.52%	78.15%	80.32%	75.39%	1.31%	74.45%	76.33%
SVC_linear	Chi2	25	TF-IDF	3	77.25%	0.75%	76.71%	77.78%	0.164	79.70%	1.59%	78.56%	80.84%	74.97%	1.07%	74.20%	75.73%
SVC_linear	Chi2	90	TF-IDF	2	77.24%	0.51%	76.88%	77.61%	0.099	78.23%	1.47%	77.18%	79.29%	76.31%	0.96%	75.62%	76.99%
SVC_linear	F-value	10	TF-IDF	3	77.24%	0.85%	76.63%	77.85%	0.194	79.34%	1.84%	78.02%	80.66%	75.28%	0.84%	74.68%	75.88%
NuSVC	Chi2	25	TF-IDF	3	77.24%	0.57%	76.83%	77.64%	0.106	80.68%	1.34%	79.72%	81.64%	74.09%	0.86%	73.48%	74.71%
NuSVC	F-value	30	TF-IDF	2	77.23%	0.52%	76.86%	77.60%	0.094	79.95%	1.54%	78.84%	81.05%	74.73%	1.10%	73.94%	75.52%
SVC_linear	Chi2	35	TF-IDF	2	77.23%	0.63%	76.78%	77.68%	0.117	78.54%	1.49%	77.47%	79.60%	76.00%	1.16%	75.17%	76.83%
SVC_linear	Chi2	45	TF-IDF	2	77.23%	0.60%	76.80%	77.66%	0.108	78.55%	1.41%	77.54%	79.56%	75.98%	1.15%	75.15%	76.80%
SVC_linear	Chi2	65	TF-IDF	2	77.23%	0.70%	76.73%	77.73%	0.133	78.52%	1.57%	77.39%	79.64%	76.01%	1.31%	75.08%	76.95%
SVC_linear	Chi2	30	TF-IDF	3	77.21%	0.82%	76.63%	77.80%	0.156	79.82%	1.65%	78.64%	81.00%	74.79%	1.07%	74.03%	75.56%
NuSVC	F-value	10	TF-IDF	3	77.21%	0.59%	76.79%	77.63%	0.092	80.25%	1.45%	79.22%	81.28%	74.41%	0.83%	73.81%	75.00%
NuSVC	F-value	85	TF-IDF	2	77.21%	0.45%	76.88%	77.53%	0.067	78.43%	1.46%	77.39%	79.47%	76.06%	1.22%	75.19%	76.93%
SVC_linear	Chi2	30	TF-IDF	2	77.20%	0.69%	76.71%	77.69%	0.112	78.43%	1.52%	77.34%	79.52%	76.04%	1.18%	75.19%	76.88%
NuSVC	F-value	75	TF-IDF	2	77.18%	0.59%	76.76%	77.60%	0.078	78.80%	1.54%	77.70%	79.90%	75.67%	1.31%	74.73%	76.60%
SVC_linear	F-value	95	TF-IDF	3	77.18%	0.62%	76.74%	77.62%	0.084	79.23%	1.55%	78.12%	80.34%	75.26%	0.76%	74.71%	75.80%
NuSVC	Chi2	60	TF-IDF	2	77.18%	0.55%	76.79%	77.57%	0.069	78.86%	1.57%	77.73%	79.98%	75.60%	1.03%	74.86%	76.34%
SVC_linear	Chi2	50	TF-IDF	2	77.18%	0.61%	76.74%	77.61%	0.079	78.57%	1.52%	77.48%	79.66%	75.86%	1.15%	75.04%	76.68%
NuSVC	Chi2	80	TF-IDF	2	77.17%	0.58%	76.76%	77.59%	0.073	78.50%	1.54%	77.40%	79.60%	75.93%	1.13%	75.12%	76.73%
NuSVC	Chi2	20	TF-IDF	3	77.17%	0.60%	76.74%	77.60%	0.075	80.46%	1.41%	79.45%	81.47%	74.16%	1.08%	73.39%	74.93%
SVC_linear	Chi2	20	TF-IDF	3	77.17%	0.67%	76.69%	77.64%	0.087	79.38%	1.67%	78.18%	80.57%	75.11%	1.10%	74.33%	75.90%
NuSVC	Chi2	15	TF-IDF	2	77.16%	0.67%	76.67%	77.64%	0.083	78.34%	1.10%	77.55%	79.12%	76.03%	1.36%	75.06%	77.01%
NuSVC	F-value	90	TF-IDF	2	77.16%	0.70%	76.65%	77.66%	0.090	78.25%	1.63%	77.08%	79.42%	76.12%	1.02%	75.40%	76.85%
SVC_linear	Chi2	25	TF-IDF	2	77.16%	0.68%	76.67%	77.64%	0.083	78.39%	1.56%	77.28%	79.51%	75.99%	1.10%	75.20%	76.78%
SVC_linear	Chi2	20	TF-IDF	2	77.15%	0.58%	76.74%	77.56%	0.060	78.34%	1.39%	77.34%	79.33%	76.02%	1.23%	75.15%	76.90%
NuSVC	Chi2	45	BR	3	77.15%	1.17%	76.31%	77.98%	0.209	76.89%	1.81%	75.60%	78.19%	77.43%	1.54%	76.33%	78.54%
NuSVC	Chi2	65	TF-IDF	2	77.15%	0.62%	76.70%	77.59%	0.068	78.70%	1.45%	77.66%	79.73%	75.68%	1.18%	74.84%	76.53%
SVC_linear	Chi2	35	TF-IDF	3	77.14%	0.75%	76.61%	77.68%	0.094	79.79%	1.61%	78.63%	80.94%	74.70%	1.06%	73.94%	75.46%
SVC_linear	F-value	10	TF-IDF	2	77.12%	0.83%	76.53%	77.71%	0.099	77.93%	1.59%	76.79%	79.07%	76.36%	1.14%	75.54%	77.17%
SVC_linear	F-value	5	TF-IDF	3	77.12%	0.87%	76.50%	77.74%	0.109	78.89%	1.75%	77.64%	80.15%	75.46%	1.16%	74.62%	76.29%
SVC_linear	Chi2	55	TF-IDF	2	77.12%	0.64%	76.66%	77.57%	0.058	78.48%	1.54%	77.38%	79.59%	75.83%	1.23%	74.95%	76.72%
SVC_linear	Chi2	100	TF-IDF	2	77.09%	0.92%	76.43%	77.75%	0.105	77.91%	1.65%	76.73%	79.09%	76.31%	1.16%	75.48%	77.14%
SVC_linear	F-value	100	TF-IDF	2	77.09%	0.92%	76.43%	77.75%	0.105	77.91%	1.65%	76.73%	79.09%	76.31%	1.16%	75.48%	77.14%
SVC_linear	F-value	90	TF-IDF	2	77.09%	0.74%	76.56%	77.62%	0.065	78.00%	1.65%	76.82%	79.18%	76.23%	1.12%	75.43%	77.03%
SVC_linear	F-value	5	TF-IDF	2	77.08%	0.74%	76.55%	77.61%	0.062	77.79%	1.50%	76.71%	78.86%	76.42%	1.24%	75.54%	77.31%

NuSVC	F-value	50	BR	3	77.05%	1.12%	76.25%	77.85%	0.132	76.41%	1.63%	75.24%	77.58%	77.73%	1.51%	76.64%	78.81%
NuSVC	Chi2	50	BR	3	77.04%	1.19%	76.19%	77.88%	0.142	76.30%	1.79%	75.02%	77.58%	77.82%	1.51%	76.74%	78.89%
NuSVC	Chi2	40	BR	3	77.03%	1.19%	76.18%	77.89%	0.142	76.54%	1.92%	75.17%	77.91%	77.58%	1.50%	76.50%	78.65%
ETC	F-value	30	TF	3	77.03%	1.10%	76.24%	77.82%	0.117	82.04%	1.83%	80.73%	83.35%	72.62%	0.98%	71.91%	73.32%
NuSVC	F-value	40	BR	3	77.01%	1.16%	76.18%	77.84%	0.120	76.63%	1.80%	75.34%	77.91%	77.43%	1.51%	76.35%	78.51%
SVC_linear	Chi2	5	TF-IDF	3	77.00%	0.85%	76.40%	77.61%	0.055	78.54%	1.94%	77.15%	79.92%	75.57%	1.20%	74.71%	76.43%
SVC_linear	Chi2	5	TF-IDF	2	77.00%	0.92%	76.34%	77.66%	0.067	77.54%	1.83%	76.23%	78.84%	76.51%	1.17%	75.67%	77.35%
ETC	F-value	40	BR	3	77.00%	1.21%	76.13%	77.86%	0.127	85.02%	2.01%	83.58%	86.45%	70.38%	1.20%	69.52%	71.24%
ETC	Chi2	35	TF	3	76.98%	1.21%	76.11%	77.84%	0.116	82.86%	1.88%	81.51%	84.20%	71.89%	1.05%	71.14%	72.64%
NuSVC	F-value	45	BR	3	76.95%	1.48%	75.89%	78.00%	0.163	76.38%	2.96%	74.25%	78.50%	77.62%	1.66%	76.43%	78.81%
NuSVC	Chi2	35	BR	3	76.94%	1.22%	76.07%	77.81%	0.103	76.11%	1.87%	74.77%	77.45%	77.83%	1.46%	76.78%	78.87%
NuSVC	F-value	35	BR	3	76.93%	1.13%	76.12%	77.74%	0.081	76.13%	1.88%	74.78%	77.47%	77.79%	1.41%	76.78%	78.80%
ETC	F-value	35	TF	3	76.86%	1.07%	76.10%	77.63%	0.052	82.75%	1.79%	81.47%	84.03%	71.78%	1.06%	71.02%	72.53%
ETC	Chi2	40	BR	3	76.84%	1.15%	76.02%	77.66%	0.058	85.09%	1.77%	83.83%	86.35%	70.06%	1.11%	69.27%	70.86%
NuSVC	Chi2	55	BR	3	76.84%	1.24%	75.95%	77.72%	0.071	75.71%	1.85%	74.39%	77.04%	78.02%	1.48%	76.97%	79.08%
ETC	F-value	45	TF	3	76.51%	1.79%	75.23%	77.79%	0.071	82.82%	5.56%	78.84%	86.80%	71.38%	2.36%	69.69%	73.07%
ETC	F-value	45	BR	3	76.09%	2.24%	74.48%	77.69%	0.052	83.25%	6.42%	78.66%	87.84%	70.41%	2.46%	68.65%	72.17%

FS: Feature Selection, P: Percentile, SoW: Set-of-Words, NG: n-grams, SD: Standard Deviation, LB-CI: Lower Boundary – Confidence Interval, UB-CI: Upper Boundary – Confidence Interval, PV: p-value, ETC: Extra Trees Classifier

3. ARTIGO 2 – MAN vs. MACHINE: PREDICTING HOSPITAL BED DEMAND FROM AN EMERGENCY DEPARTMENT

Artigo submetido ao International Journal of Medical Informatics; em revisão.

Abstract

Objective: Recent literature reports promising results from using artificial intelligence to support decision making in healthcare. However, there is still limited knowledge on whether intelligent systems can operate as fully-autonomous, user-independent systems. In this study we compare the performance of a computer-based algorithm and humans in predicting hospital bed demand from an Emergency Department (ED) using only textual medical records generated in the first doctor-patient encounter.

Design: We used a testing set of 230 textual medical records with known patient final status (inpatient or discharged) to compare machines and humans in the task of predicting patients' outcomes. The machine role was played by a text-mining algorithm based on a Support Vector Machine Classifier. The human role was performed by four emergency department physicians with two different levels of experience (novice and experienced).

Measurements: Predictions of machine and physicians were compared in terms of accuracy, sensitivity, specificity, and ROC curves.

Results: Our data indicates that the computer system could predict patient admission or discharge outcomes with 80% accuracy, which was similar or better than the performance of novice and experienced physicians.

Conclusions: Results suggest that the artificial intelligence algorithm could operate as an autonomous and independent system to predict hospitalization of ED patients.

Keywords: intelligent systems; decision making support; emergency department

3.1. INTRODUCTION

Alan Turing famously proposed in 1950 that machines could exhibit intelligent behavior that was equivalent to, or indistinguishable from human behavior (TURING, 1950).

Since then, evolving computer technologies and programming techniques allowed computers to emulate tasks that were normally done by humans (CANTU-ORTIZ, 2014). In fact, there is growing evidence of cases where human performance has already been matched or surpassed by machines (SILVER et al., 2016).

In healthcare operations, the growing use of intelligent systems (e.g. to support clinical decisions) may represent an important transition to a "new healthcare" (ROBSON; BAEK, 2009). Intelligent systems can help dealing with demands for more accurate diagnostics (CHOCKLEY; EMANUEL, 2016), safer and more effective medicines (DIMITRI; LIÓ, 2017), and more effective treatments (ESFANDIARI et al., 2014). Using these systems may lead to improved diagnostic and treatment protocols and procedures, and consequently increased patient wellness (DEO, 2015).

Despite positive reports in the literature (CHOCKLEY; EMANUEL, 2016; DONG; WANG; SI, 2017), the capability of intelligent systems to operate autonomously is still in question (CHOCKLEY; EMANUEL, 2016). Most studies applying the use of computers to exam analysis (DONG; WANG; SI, 2017) or patient diagnosis (KIM et al., 2017) did not compare computer performance with human-generated inputs, other than those used to establish the ground truth. Moreover, studies (BECKER et al., 2017; PORZ et al., 2016; TUFAIL et al., 2017) (mainly in image analysis) that did compare computer and human performances reported varying results. For example, Becker et al. (BECKER et al., 2017) found that their algorithm could detect breast cancer in mammography with similar accuracy to radiologists. On the other hand, Tufail et al. (TUFAIL et al., 2017) found that the three systems in their study were less effective than humans in detecting appropriate outcomes in diabetic retinopathy image assessment.

Our study evaluates the performance of an autonomous intelligent system in the emerging "new healthcare" context (ROBSON; BAEK, 2009). We compare the performance of a computer-based algorithm and humans in predicting hospital bed demand in an emergency department (ED). Lucini et al. (LUCINI et al., 2017) proposed the algorithm based on the analysis of textual medical records generated during the first patient-doctor encounter in an ED. According to the authors, the algorithm enabled staff to complete inward bed capacity planning at least one shift earlier (6 hours) than before, which might lead to significant reduction in overcrowding.

This study contributions are as follows. From an academic perspective, this is to the best of our knowledge the first study to compare the performance of a computer system and humans in predicting ED bed demand based only on textual medical records. From a practical perspective, the results help to validate the effectiveness of a computerized algorithm as an autonomous healthcare management tool in an ED.

3.2. BACKGROUD

3.2.1. Motivation

At least five arguments support the view that intelligent systems may outperform humans in common healthcare tasks. First, in most intelligent systems rules are created by the machine and not by people, allowing non-trivial patterns to be discovered (CHOCKLEY; EMANUEL, 2016). These systems usually combine variables in nonlinear and interactive ways (DEO, 2015), enabling the identification of relations between symptoms that physicians might not have been trained to notice. Second, intelligent systems have less cognitive limitations, and are not affected by the environment (CHOCKLEY; EMANUEL, 2016). It is well known that task complexity affects the efficiency of decision-making (FRANKLIN et al., 2011), i.e. complex tasks require more cognitive effort from analysts. Moreover, common aspects of the healthcare system, such as high workloads, multitasking, interruptions, time pressure, and shift changes may lead to poor decision-making (LAXMISAN et al., 2007). Third, the ever-increasing processing speed of computers make intelligent systems significantly faster to respond (TAJIK, 2016). Certain medical practices (i.e. echocardiographic image analysis) may be time-consuming for humans, but not for computers (CHOCKLEY; EMANUEL, 2016). Fourth, intelligent systems do not present interobserver variation (LEE et al., 2017): two different computers using the same algorithm and trained with the same dataset will give the same answer to a problem. In contrast, different physicians may have different opinions about the same problem, even sharing similar experience or background. Fifth, knowledge libraries of intelligent systems can increase continuously. As time passes, feedback from previous tasks may be used to recalibrate the system, potentially improving future performance (CHOCKLEY; EMANUEL, 2016). Additionally, computers may share knowledge with other computers in a network, without loss of information (TAJIK, 2016).

One critical area in the healthcare system that may benefit from intelligent systems is the emergency department (ED). The ED is a complex environment with high-workload, which is usually both information intensive and starved: decisions may be either made based on anamnesis and complex examinations, or on almost no information (e.g. upon arrival of high acuity patients). EDs are also time sensitive, highly stressful, non-deterministic, interruption-laden, and life-critical environments (FRANKLIN et al., 2011). In these conditions, effective decision making is difficult or even impossible to achieve by humans (FRANKLIN et al., 2011).

However, despite some effort to develop intelligent systems to support ED decision-making (MATOS; RODRIGUES, 2011), there is not much evidence from studies assessing their capability to operate as fully-automated, user-independent systems. We still lack knowledge on whether computer-based systems may be considered suitable to substitute humans in ED decision making, whether decisions are clinical or managerial. This study targets such knowledge gap by investigating the performance of an intelligent system developed to predict hospitalizations and discharges, based on early ED patient text records using the SOAP (Subjective, Objective, Assessment, and Plan) framework.

3.2.2. Related Work

Studies in many areas have explored the benefits and implications of using computer systems rather than humans in complex decision making. In general, decisions are related to games, such as chess or Go (SILVER et al., 2016), or to extracting, identifying or classifying data based on pictures, videos, sounds, text or a combination of them. Examples include speech recognition (XIONG et al., 2016), lip reading (CHUNG et al., 2016), and image classification (HE et al., 2015). The main application in healthcare belongs to the second class of problems, being related to medical imaging diagnostic. Mammography (BECKER et al., 2017), retinal imaging (BURLINA et al., 2017; TUFAIL et al., 2017), and magnetic resonance (PORZ et al., 2014, 2016) are some examples. It is noteworthy that most healthcare studies focus on clinical rather than managerial decisions.

In 2017, a study evaluated the diagnostic accuracy of a deep learning algorithm for the detection of breast cancer based on mammography images (BECKER et al., 2017). Results were compared to the opinion of three radiologists with ten, seven and three years of experience in breast imaging analysis. They concluded that, in terms of accuracy measured by the Area

Under the ROC Curve (AUC) (FAWCETT, 2006), there was no significant difference between radiologists and the algorithm. Radiologists exhibited higher specificity and lower sensitivity compared to the machine. According to authors, advantages of the algorithm included speed and possibility to train and adapt it to different patient populations.

Regarding retinal imaging, a study examined an aged-related macular degeneration categorization task (BURLINA et al., 2017). The authors proposed a multi-class classification system with four aged-related macular degeneration severity levels as classes. Machine results were compared to those obtained from an experienced ophthalmologist: in two classes the machine's accuracy outperformed that of the physician by average 13.6%, while in remaining classes the physician outperformed the algorithm by average 22.98%. The conclusion was that machine and physician performed similarly; however, the latter was much slower. While the physician took several days to complete the task of classifying the 5,664 images in the study, the algorithm took only 68 minutes.

Another study presented a comparison between gradings obtained using three different automated diabetic retinopathy image assessment systems and human graders (TUFAIL et al., 2017). The study also assessed the economic impact of using automated systems as a prior step in the UK National Health Service Diabetic Eye Screening Programme (NHS DESP) protocol (PUBLIC HEALTH ENGLAND, 2017) or as a substitute to human graders in the protocol's initial grading step. Results show that two of the systems achieved acceptable sensitivity compared to human graders, at a level of specificity that makes them cost-effective alternatives to a purely manual grading (TUFAIL et al., 2017).

Two related studies presented results derived from the use of intelligent systems for brain tumor image analysis (PORZ et al., 2014, 2016). They compared a fully-automated system to manual (PORZ et al., 2014), and semi-automatic user-guided (PORZ et al., 2016) segmentation techniques. Results show that automated and manual volume measurements were equivalent for CETV (Contrast Enhancing Tumor Volume) (PORZ et al., 2014). In addition, authors concluded that the automated system might provide similar segmentation results compared to the semi-automatic user-guided technique in a clinical setting (PORZ et al., 2016). Major advantages of the automated system are its independence from human interactions, significant time gains compared to manual segmentation, and capacity to analyze images of tumors in different brain areas (PORZ et al., 2014).

3.3. METHODS

3.3.1. Study Database

The database was made available by the ED of Hospital de Clínicas de Porto Alegre (HCPA), an 842-bed, tertiary care teaching public hospital located in the city of Porto Alegre, Brazil. The data consist of electronic medical records (EMR) from 16,703 patients who visited the ED between January and December 2014. Each EMR contained relevant information on clinical care provided to patients, such as SOAP framework notes, prescriptions and exams. All textual records were written in Brazilian Portuguese. This study was approved by the hospital's Ethical Committee, and authors have complied with the recommendations of the Declaration of Helsinki.

The database was adjusted to be used as input to both machine and human graders. For this reason, two pre-processing steps, similar to the Informative Text Retrieval and the Split to Training and Testing Sets steps presented by Lucini et al (LUCINI et al., 2017), were used. First, duplicated and empty textual records and those where patient final status were explicitly described in the SOAP notes were excluded; e.g. texts such as "HAA" (hospital admission authorized), "patient left", "PD" (patient discharged), "patient not located", and "patient did not answer when called". In addition to the exclusion criteria proposed by Lucini et al (LUCINI et al., 2017), we also excluded incomplete textual records, corresponding to those with missing information on one or more subjects of the SOAP framework. Consequently, the number of EMRs was reduced to 9,030, of which 4,673 were records from discharged patients and 4,357 were records of admitted patients (Table 3.1). Records were structured with two fields of information. The first was a dummy variable indicating patient final status (inpatient or discharged). The second contained the free-form textual SOAP notes entered by HCPA's healthcare providers during the first patient-physician encounter.

Note that the HCPA's ED classifies patient acuity according to the Manchester Triage System (ZACHARIASSE et al., 2017), and is tasked within the local health system with providing care only to patients in the immediate (I), very urgent (VU), and urgent (U) categories. The balanced ratio between inpatient and discharged classes in the dataset reflects the type of care provided by the ED. No sample adjustments were made.

Table 3.1: Selection of reports from the database

Description	Inpatient	Discharged	Total
Number of reports in the original database	8,038 (48%)	8,665 (52%)	16,703
Duplicated reports	-10	-5	-15
Information on final status explicit in reports	-191	-2,009	-2,200
Empty reports	-2,246	-1,067	-3,313
Incomplete SOAP reports	-1,234	-911	-2,145
Number of reports used in the study	4,357 (48%)	4,673 (52%)	9,030

Next, the dataset was split into training and testing sets. The training set was used to develop the machine's algorithm and the testing set was used to compare machine and human graders. To avoid biased results in the human baseline due to cognitive overload, the number of records in the testing set was set to 230; the 8,800 remaining records were considered training set. The ratio between records belonging to classes inpatient and discharged was kept constant in both sets, according to the original sample distribution. As a result, training and testing sets presented 4,553 and 120, and 4,247 and 110 records in classes discharged and inpatient, respectively.

3.3.2. Intelligent System Protocol

The machine role was played by a Support Vector Machine Classifier (SVMC), which was the best predicting algorithm according to the analysis in Lucini et al (LUCINI et al., 2017). Textual information from all records were pre-processed and the algorithm trained. Next, the algorithm was used to predict the final status of 230 patients from the testing set.

The SVMC was first introduced by Vapnik (VAPNIK, 2013), being widely used to solve supervised classification problems from different application domains (CHATTERJEE; FERMOYLE; RAGHAVAN, 2012). The two-class SVMC finds a hyperplane that ensures maximum separation between classes. Separating margins are identified by few data elements named support vectors (VAPNIK, 2013). The best predicting algorithm in Lucini et al (LUCINI et al., 2017) uses a variation of the two-classes SVMC called nu-Support Vector Classifier (nu-

SVC). Proposed by Schölkopf et al (SCHÖLKOPF et al., 2000), nu-SVC uses a parameter to control the number of support vectors and training errors. More details about the classifier are available in (SCHÖLKOPF et al., 2000). In this study, the nu-SVC was implemented using a linear kernel and default parameters of the package scikit-learn (HACKELING, 2014) from Python 3.6 (PYTHON SOFTWARE FOUNDATION, 2017).

As proposed by Lucini et al. (LUCINI et al., 2017), four pre-processing steps were carried out to prepare the textual information to be used as input to the algorithm: normalization, tokenization, feature selection, and conversion to set-of-words. In the normalization step, punctuation marks, numerical characters, and stop words were removed; capital letters were substituted by lowercases; and words were reduced to radical forms (FELDMANN; SANGER, 2007). In the tokenization step, the continuous string of characters of each record was broken down into linguistic units called tokens (FELDMANN; SANGER, 2007), which were delimited by blank spaces in the string (LUCINI et al., 2017). The algorithm uses the combination of unigrams, bigrams and trigrams as features, which are sequences of one, two or three adjacent words from the list of tokens (FELDMANN; SANGER, 2007). In the feature selection step, the F-Value of each feature from the training set was calculated. Features with an F-Value above the 65-th percentile of largest values were selected. The F-Value of a particular feature is calculated by (LUCINI et al., 2017):

$$F\_value(i) = \frac{(X_i^{imp} - X_i)^2 + (X_i^{dis} - X_i)^2}{\frac{1}{n_{imp}-1}\sum_{k=1}^{n_{imp}}(X_{k,i}^{imp} - X_i^{imp})^2 + \frac{1}{n_{dis}-1}\sum_{k=1}^{n_{dis}}(X_{k,i}^{dis} - X_i^{dis})^2} \quad (1)$$

where X_i, X_i^{imp} and X_i^{dis} is the average of the i-th feature in the complete, inpatient, and discharged datasets, respectively; $X_{k,i}^{imp}$ is the i-th feature of the k-th inpatient instance, and $X_{k,i}^{dis}$ is the i-th feature of k-th discharge instance. The last pre-processing step was the conversion to set-of-words representation, which results in a matrix indicating the occurrence of selected features in records. Matrix columns represent each of all selected features, while matrix rows represent each of the records. Matrix cells were filled out using term frequency – inverse document frequency (TF-IDF) indicator results. It may be expressed as (LUCINI et al., 2017):

$$TFIDF(t, d, D) = \frac{f(t,d)}{\max\{f(t,d) : t \in d\}} \times \log \frac{|D|}{|\{d \in D : t \in d\}|} \quad (2)$$

where t denotes the feature, d denotes the record, D is the total number of records in the collection, and $f(t, d)$ is the number of occurrences of feature t in record d.

Once all pre-processing steps were performed, the algorithm was trained and the final status of patients from the testing set was predicted (LUCINI et al., 2017). Records of the testing set could be classified as belonging to one of two classes, inpatient or discharged. It was assumed that if a record was classified as inpatient, the patient stayed in the hospital and a ward bed was required. On the other hand, if a record was classified as discharged, there was no need for hospitalization. As a result, a list indicating the class of the 230 records in the testing set was made available. This list, generated by the best predicting algorithm proposed by Lucini et al. (LUCINI et al., 2017), was used to measure the intelligent system performance in the present study.

3.3.3. Human Baseline

The manual prediction of the testing set was performed by four HCPA's ED physicians. Following Becker et al. (BECKER et al., 2017), recruited physicians had different levels of experience. Two were novice ED physicians, with less than two years of experience working in the HCPA's ED. The other two were experienced, working in the HCPA's ED for more than 10 years.

Manual prediction was carried out in two steps. The first step aimed to verify substantial inter-observer agreement between physicians with similar experience. A trial testing subset of 30 randomized records was created, 14 belonging to the inpatient class and 16 belonging to the discharged class. All four physicians analyzed the records and the Cohen's Kappa coefficient (LANDIS; KOCH, 1977) was used to measure their level of agreement. Kappa values between 0.61 and 0.80, and between 0.81 and 1.00 represent "substantial" and "almost perfect" strengths of agreement respectively (LANDIS; KOCH, 1977). Thus, 0.61 was used as minimum threshold for grouping evaluations from each pair of physicians with same experience.

The pair of novice physicians and the pair of experienced physicians disagreed in 4 and 5 predictions respectively. As a result, Cohen's Kappa coefficient was 0.74 for novice physicians and 0.66 for experienced physicians; both scores indicated "substantial" level of agreement, surpassing the minimum established threshold for grouping evaluations.

In the second step, the remaining 200 records in the testing set were randomly divided into two subsets of 100 records, A and B, observing class frequencies (i.e. 48 belonging to inpatient class and 52 belonging to discharged class). Records in subset A were assigned to one physician in each experience level group, while records in subset B were assigned to the other physician. Eventually, each of the 200 records were analyzed by both experienced and novice physicians.

To collect physicians' input, subsets A and B were assembled in the online survey website SurveyMonkey (SURVEYMONKEY INC., [s.d.]). Physicians received personalized weblinks to access the records they were expected to evaluate. For each record, physicians were instructed to read and answer the following question: "Based exclusively on this SOAP note, what do you think happened to the patient?" Two options were given: (i) "INPATIENT – Patient required a hospital ward bed" or (ii) "DISCHARGED – Patient was discharged after consultation, not requiring a hospital ward bed". Records consisted of four paragraphs, one to each subject of the SOAP framework. Physicians did not receive any feedback about their answers. In addition, it was forbidden for physicians to consult any other records, systems, colleagues, or other resources. Figure 3.1 shows an example of record presented to physicians.

Next, answers to evaluations in the two steps of the manual prediction were unified, forming a list of 230 records and their respective answers by novice and experienced physicians. In case of divergence between answers of physicians in the same experience group (recall that records were evaluated by all physicians in step one), the conservative answer ("Inpatient") was considered. The conservative answer maximized patient safety and was consistent with practice. In this particular hospital, whenever a divergence occurred, the chosen alternative was to keep patients under observation in ward beds. The unified list was used as human baseline.

3.3.4. Statistical Analysis

This study used validation methods similar to those in Becker et al. [11]. The statistical analysis was carried out in R x64 3.3.3 (RIPLEY, 2001). Receiver Operating Characteristic (ROC) (DELONG; DELONG; CLARKE-PEARSON, 1988) analysis was performed for both machine and humans. Accuracies of the intelligent system protocol and human baseline were expressed as the AUC, and compared using the nonparametric test by DeLong et al. (DELONG; DELONG; CLARKE-PEARSON, 1988). Youden's J statistic (YOUDEN, 1950) was used to

calculate the optimal cutoff for sensitivity and specificity. According to (ALPAYDIN, 2010), sensitivity and specificity are the same as true positive rate (i.e., the number of true inpatients divided by the total number of cases classified as inpatients) and true negative rate (i.e., the number of true discharged divided by the total number of cases classified as discharged), respectively. We considered a p-value equal to or smaller than 0.05 as indicative of significant differences.

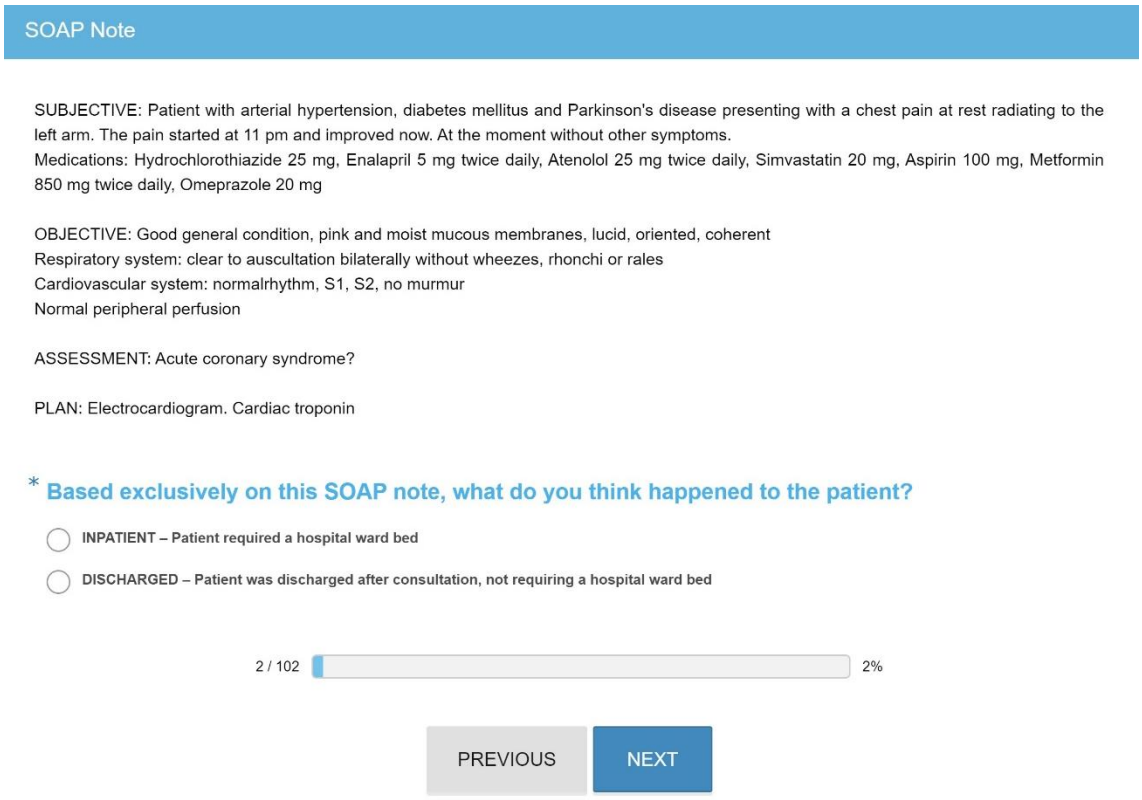

Figure 3.1: Record sample in SurveyMonkey (translated from Portuguese)

3.4. RESULTS

Machine training was completed successfully in 2 minutes and 46 seconds. Processing time per test EMR was 0.00812±0.0009 seconds. In contrast, novice physicians took in average 156.80 seconds per test EMR, while experienced physicians took in average 56.40 seconds per test EMR. All computations were performed on an Intel Core i7-7500U@2.9GHz and 16GB RAM.

ROC curves for machine and novice and experienced physicians are presented in Figure 3.2. Table 3.2 shows accuracy by AUC for the testing set. The best performance was achieved by novice physicians with AUC of 0.82 [95% confidence interval (CI) of 0.77 – 0.87]. The machine performed second best with AUC of 0.80 (95% CI, 0.75 – 0.85). Experienced physicians had the poorest performance, with AUC of 0.76 (95% CI, 0.71 – 0.81). Prediction accuracy by AUC was not significantly different among graders (Table 3.3).

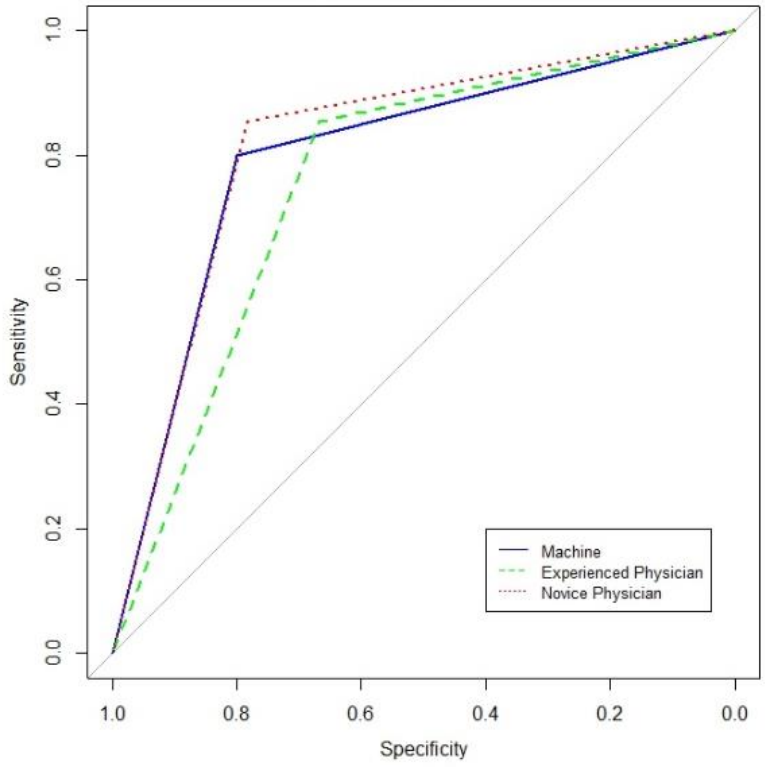

Figure 3.2: ROC curves by group

Table 3.2: Accuracy, sensitivity and specificity with 95% CI considering the testing set

	Machine	**Novice Physicians**	**Experienced Physicians**
Accuracy	0.80 (0.75 – 0.85)	0.82 (0.77 – 0.87)	0.76 (0.71 – 0.81)
Sensitivity	0.80 (0.73 – 0.87)	0.85 (0.78 – 0.92)	0.85 (0.79 – 0.92)
Specificity	0.80 (0.73 – 0.87)	0.78 (0.71 – 0.86)	0.67 (0.58 – 0.75)

Table 3.2 also shows sensitivity and specificity for the testing set at optimal cutoff point. The highest sensitivity (0.85) was achieved by both novice and experienced physicians; however, the latter presented a slightly narrower 95% CI (0.79 – 0.92). Regarding specificity, the highest score was achieved by machine with 0.80 (95% CI, 0.73 – 0.87). Novice physicians had the second highest score with 0.78 (95% CI, 0.71 – 0.86). Experienced physicians had the lowest specificity (0.67, with 95% CI of 0.58 – 0.75).

Table 3.3: DeLong's test *p*-values for two correlated ROC curves

Comparison	***p*-values**
M *vs.* NP	0.5528
M *vs.* EP	0.2395
NP *vs.* EP	0.0517

M: Machine; NP: Novice Physician; EP: Experienced Physician

Table 3.4 summarizes prediction success in the 230 cases. The machine was correct but the physicians were not in 14 cases; the opposite occurred in 19 cases. All graders were correct in 137 cases, but incorrect in 10 cases. There were divergences between physicians in 50 cases.

3.5. DISCUSSION

To our knowledge, this is the first study comparing an algorithm generated through machine learning with ED physicians in predicting patient admissions after their first evaluation. A computer running such algorithm could provide real-time data on bed necessity and aid bed management teams to improve patient flow processes. Our main results show that performances (accuracy) of physicians (novice or experienced) and machine were similar. Highest sensitivity was achieved by both novice and experienced physicians, and higher specificity was achieved by the machine, although confidence intervals had some overlap. In addition, the algorithm was significantly faster when compared to doctors. In average, novice and experienced physicians took 56.40 seconds and 156.80 seconds to analyze each record, respectively. In contrast, the algorithm analyzed the whole testing set in less than two seconds.

Table 3.4: Comparison of mistakes and successes among graders

Situation	Occurrences		
	Discharged Patients	**Admitted Patients**	**Total**
Machine and physicians were wrong	9 (3.91%)	1 (0.44%)	10 (4.35%)
Machine was correct and all physicians were wrong	12 (5.22%)	2 (0.87%)	14 (6.09%)
All physicians were correct and machine was wrong	10 (4.35%)	9 (3.91%)	19 (8.26%)
Divergences between novice and experienced physicians	23 (10.00%)	27 (11.74%)	50 (21.74%)
Machine and physicians were correct	66 (28.70%)	71 (30.86%)	137 (59.56%)

In our study, the computer algorithm was as accurate as the doctors. Note that physicians reviewed records outside working hours in the ED, in a calm setting, without multitasking or time constraints. It is difficult to predict how their performances could be affected if evaluations were carried out in parallel with their ED working activities, but it is reasonable to speculate that they would be worse. On the other hand, machine performance is not affected by the working environment.

Improving patient flow and discharge processes through bed management supporting teams (EPSTEIN, 2014) have resulted in reduction in length of stay, cancelled interventions and increase in planned discharges (BAKER; ESBENSHADE, 2015). The use of a computer algorithm as decision support system could enable hospital staff and health decision makers to better manage hospital inpatient beds, thus potentially reducing costs and inpatient length of stay (MATOS; RODRIGUES, 2011). Previous studies evaluated mathematical programs to model bed assignment (SCHMIDT; GEISLER; SPRECKELSEN, 2013) and prediction models to improve efficiency (SRIKANTH; ARIVAZHAGAN, 2017) with positive results, but none could made the predictions in real time, as it is possible with the herein tested algorithm.

The algorithm presented similar sensitivity and specificity values, while doctors tended to present higher sensibility and lower specificity. That could be due to a clinical conservative bias – when in doubt, it would be safer to admit a patient then to discharge her. However, from a management standpoint a more balanced cut-off point on the ROC curve would be potentially more beneficial, since bed demand prediction would then be more stable

with similar false negative and false positive rates. As Proudlove et al. (PROUDLOVE; GORDON; BOADEN, 2003) have discussed, a key requirement for effective bed management is information; anticipatory planning requires prediction of admissions and discharges. The authors have stated that there is a huge potential to improve information made available to bed management teams, and that few bed management functions have access to reliable data on patients expected to come in and be admitted.

The proportion of cases in which physicians were correct while the machine was wrong, and in which the machine was correct while physicians were wrong was similar (8.26% and 6.09%, respectively). The former group (physicians correct) is composed of nine admissions and ten discharges. On the discharges, it would seem that the algorithm has difficulty in recognizing information on patients being sent to other health services (two cases). On six other cases, there seemed to be extensive information on patient health history on the SOAP note, which was not directly related to the current condition of the patient. On the admissions, it is harder to presume causes for incorrect classification; for human readers, some SOAP notes clearly indicated either emergencies or the need for interventions: e.g. unstable angina, intestinal obstruction, nephrolithiasis necessitating a double J stent, and new onset atrial fibrillation. Since these descriptions were shortly stated on the Assessment item of the note, it is possible that the algorithm did not find enough information in the remaining of the note to justify admission, and could not interpret the medical condition alone.

Regarding the cases where the algorithm was correct and physicians were wrong, most cases were discharges (12 of the 14 cases). This again may show a conservative bias from physicians, tending to classify cases as “admissions” when in doubt. It is hard for human readers to explain why the algorithm was correct; some SOAP notes describe serious diagnostic hypotheses, such as suspicion of acute coronary syndrome or of deep vein thrombosis, while others describe cases which appear to be non-complicated urinary tract infections, gastroenteritis or headaches. Human understanding of AI decisions is known to be limited, and it has been proposed that it would be possible to train AI itself to provide natural language justifications of its decisions (PARK et al., 2016).

A possible refinement to the tested algorithm would be to include lab results into the available data; however, since the authors’ goal was to have the earliest possible prediction on patient admission or discharge, they chose to use only the first SOAP note registered by a physician. Indeed, the combination of four data mining operations including from lab results to

clinical information let to a correct diagnosis in 98% of the cases in a pediatric emergency room, although retrospectively (GRIGULL; LECHNER, 2012).

In closing, the sophisticated computer algorithm here tested could predict patient admission or discharge with 80% accuracy, based solely on the first SOAP note from the ED. The algorithm had comparable performance to both novice and experience ED physicians. We believe there are potential uses of intelligent computer systems in aiding hospital management, and it is possible that these systems could be more easily implemented than those focusing on clinical decisions. The proposed algorithm is one such example; its implementation could provide useful data for bed management teams, improving patient flow processes throughout the hospital.

3.6. REFERENCES

ALPAYDIN, E. **Introduction to machine learning**. 2. ed. Cambridge: The MIT Press, 2010.

BAKER, S. J.; ESBENSHADE, A. Partnering effectively with inpatient leaders for improved emergency department throughput. **Advanced Emergency Nursing Journal**, v. 37, n. 1, p. 65–71, 2015.

BECKER, A. S. et al. Deep learning in mammography: diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer. **Investigative Radiology**, v. 52, n. 7, p. 434–440, 2017.

BURLINA, P. et al. Comparing humans and deep learning performance for grading AMD: a study in using universal deep features and transfer learning for automated AMD analysis. **Computers in Biology and Medicine**, v. 82, p. 80–86, 2017.

CANTU-ORTIZ, F. J. Advancing artificial intelligence research and dissemination through conference series: Benchmark, scientific impact and the MICAI experience. **Expert Systems with Applications**, v. 41, n. 3, p. 781–785, 2014.

CHATTERJEE, A.; FERMOYLE, K.; RAGHAVAN, P. Characterizing sparse preconditioner performance for the support vector machine kernel. **Procedia Computer Science**, v. 1, n. 1, p. 367–375, 2012.

CHOCKLEY, K.; EMANUEL, E. The End of Radiology? Three Threats to the Future Practice of Radiology. **Journal of the American College of Radiology**, v. 13, n. 12, p. 1415–1420,

2016.

CHUNG, J. S. et al. **Lip reading sentences in the wild**. In Asian Conference on Computer Vision. **Anais**...Springer, 2016Disponível em: <http://arxiv.org/abs/1611.05358>

DELONG, E. R.; DELONG, D. M.; CLARKE-PEARSON, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. **Biometrics**, v. 44, n. 3, p. 837–845, 1988.

DEO, R. C. Machine learning in medicine. **Circulation**, v. 132, n. 20, p. 1920–1930, 2015.

DIMITRI, G. M.; LIÓ, P. DrugClust: A machine learning approach for drugs side effects prediction. **Computational Biology and Chemistry**, v. 68, p. 204–210, 2017.

DONG, X.; WANG, C.; SI, W. ECG beat classification via deterministic learning. **Neurocomputing**, v. 240, p. 1–12, 2017.

EPSTEIN, N. E. Multidisciplinary in-hospital teams improve patient outcomes: a review. **Surgical Neurology International**, v. 5, n. Suppl 7, p. S295–S303, 2014.

ESFANDIARI, N. et al. Knowledge discovery in medicine: Current issue and future trend. **Expert Systems with Applications**, v. 41, n. 9, p. 4434–4463, jul. 2014.

FAWCETT, T. An introduction to ROC analysis. **Pattern Recognition Letters**, v. 27, n. 8, p. 861–874, 2006.

FELDMANN, R.; SANGER, J. **The text mining handbook: advanced approaches in analyzing unstructured data**. New York: Cambridge university press, 2007.

FRANKLIN, A. et al. Opportunistic decision making and complexity in emergency care. **Journal of Biomedical Informatics**, v. 44, n. 3, p. 469–476, 2011.

GRIGULL, L.; LECHNER, W. M. Supporting diagnostic decisions using hybrid and complementary data mining applications: a pilot study in the pediatric emergency department. **Pediatric Research**, v. 71, n. 6, p. 725–731, 2012.

HACKELING, G. **Mastering Machine Learning With scikit-learn**. Birmingham, UK: Packt Publishing Ltd., 2014.

HE, K. et al. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. **Proceedings of the IEEE International Conference on Computer Vision**, p. 1026–1034, 2015.

KIM, E. Y. et al. Diagnosis of major depressive disorder by combining multimodal information from heart rate dynamics and serum proteomics using machine-learning algorithm. **Progress in Neuro-Psychopharmacology and Biological Psychiatry**, v. 76, p. 65–71, 2017.

LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical data. **Biometrics**, v. 33, n. 1, p. 159–174, 1977.

LAXMISAN, A. et al. The multitasking clinician: Decision-making and cognitive demand during and after team handoffs in emergency care. **International Journal of Medical Informatics**, v. 76, n. 11, p. 801–811, 2007.

LEE, L. H. et al. Diagnosing colorectal medullary carcinoma: interobserver variability and clinicopathological implications. **Human Pathology**, v. 62, p. 74–82, 2017.

LUCINI, F. R. et al. Text mining approach to predict hospital admissions using early medical records from the emergency department. **International Journal of Medical Informatics**, v. 100, p. 1–8, 2017.

MATOS, J.; RODRIGUES, P. P. **Modeling decisions for hospital bed management - a review**. HEALTHINF. **Anais**...2011Disponível em: <http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0003135005040507>

PARK, D. H. et al. Attentive Explanations: Justifying Decisions and Pointing to the Evidence. **arXiv preprint arXiv:1612.04757.**, 2016.

PORZ, N. et al. Multi-modal glioblastoma segmentation: man versus machine. **PLoS ONE**, v. 9, n. 5, p. 1–9, 2014.

PORZ, N. et al. Fully automated enhanced tumor compartmentalization: man vs. machine reloaded. **PLoS ONE**, v. 11, n. 11, p. 1–16, 2016.

PROUDLOVE, N. C.; GORDON, K.; BOADEN, R. Can good bed management solve the overcrowding in accident and emergency departments? **Emergency Medicine Journal**, v. 20, n. 2, p. 149–155, 2003.

PUBLIC HEALTH ENGLAND. **Diabetic eye screening: programme overview**. Disponível em: <https://www.gov.uk/guidance/diabetic-eye-screening-programme-overview>. Acesso em: 9 nov. 2017.

PYTHON SOFTWARE FOUNDATION. **The Python Language Reference**, 2017. Disponível em: <https://www.python.org/>

RIPLEY, B. D. The R project in statistical computing. **MSOR Connections. The newsletter of the LTSN Maths, Stats & OR Network**, v. 1, n. 1, p. 23–25, 2001.

ROBSON, B.; BAEK, O. K. **The engines of Hippocrates: From the dawn of medicine to medical and pharmaceutical informatics**. Volume 5 ed. Hoboken: John Wiley & Sons, 2009.

SCHMIDT, R.; GEISLER, S.; SPRECKELSEN, C. Decision support for hospital bed management using adaptable individual length of stay estimations and shared resources. **BMC Medical Informatics and Decision Making**, v. 13, n. 3, p. 1–19, 2013.

SCHÖLKOPF, B. et al. New Support Vector Algorithms. **Neural computation**, v. 12, n. 5, p. 1207–1245, 2000.

SILVER, D. et al. Mastering the game of Go with deep neural networks and tree search. **Nature**, v. 529, n. 7587, p. 484–489, 2016.

SRIKANTH, K.; ARIVAZHAGAN, D. An efficient patient inflow prediction model for hospital resource management. **Indonesian Journal of Electrical Engineering and Computer Science**, v. 7, n. 3, p. 809–817, 2017.

SURVEYMONKEY INC. **Survey Monkey**. Disponível em: <www.surveymonkey.com>. Acesso em: 9 nov. 2017.

TAJIK, A. J. Machine learning for echocardiographic imaging: embarking on another incredible journey. **Journal of the American College of Cardiology**, v. 68, n. 21, p. 2296–2298, 2016.

TUFAIL, A. et al. Automated diabetic retinopathy image assessment software: diagnostic accuracy and cost-effectiveness compared with human graders. **Ophthalmology**, v. 124, n. 3, p. 343–351, 2017.

TURING, A. M. Computing machinery and intelligence. **Mind**, v. 59, n. 236, p. 433–460, 1950.

VAPNIK, V. **The nature of statistical learning theory**. [s.l.] Springer science & business media, 2013.

XIONG, W. et al. Achieving human parity in conversational speech recognition. **arXiv preprint arXiv:1610.05256**, p. 1–12, 2016.

YOUDEN, W. J. Index for rating diagnostic tests. **Cancer**, v. 3, n. 1, p. 32–35, 1950.

ZACHARIASSE, J. M. et al. Validity of the Manchester Triage System in emergency care: a prospective observational study. **PLoS ONE**, v. 12, n. 2, p. 1–14, 2017.

4. ARTIGO 3 – TEXT MINING APPROACH TO EXPLORE DIMENSIONS OF AIRLINE CUSTOMER SATISFACTION USING SPONTANEOUS GENERATED CONTENT

Artigo submetido ao Expert Systems with Applications; em revisão.

Abstract

The airline industry operates in a highly competitive market, in which achieving and maintaining a high level of passenger satisfaction is seen as a key competitive advantage. This study presents a novel framework for measuring customer satisfaction in the airline industry. Using text mining methods we explore Online Customer Reviews (OCRs) to provide guidelines for airlines companies to improve in competitiveness. We analyze a database of more than 55,000 OCRs, covering over 400 airlines and passengers from 170 countries. Using a Latent Dirichlet Allocation (LDA) model we were able to identify 27 dimensions of satisfaction in OCRs; 882 adjectives used to describe those dimensions were also identified and their sentiment scores were calculated. Dimensions and adjectives were used to predict airline recommendation by customers, resulting in an accuracy of 79.95% (95% CI: 79.93–79.97%). The most relevant dimensions for airlines' recommendation prediction were calculated. Our method showed sensitiveness to identify variations in dimensions distribution according to different passenger characteristics and preferences (e.g. nationality, type of traveler, and cabin flown), as well as customer trends.

Keywords: airline industry, competitive advantages, text mining, online customer reviews.

4.1. INTRODUCTION

The airline industry operates in a highly competitive market, in which companies have to deal with various challenges to succeed (CALISIR; BASAK; CALISIR, 2016; DOLNICAR et al., 2011). Examples include oscillating fuel prices, fluctuating demands for service, economic crises, natural disasters, strikes, personnel shortage, restrictive government regulations, and increased security precautions (CALISIR; BASAK; CALISIR, 2016; DOLNICAR et al., 2011). In such a challenging economic environment achieving and

maintaining a high level of passenger satisfaction is seen as a key competitive advantage (CHEN, 2008; LI et al., 2017). In that sense, it is important not only to understand how passengers evaluate airlines' services, but also to identify their most valued dimensions of satisfaction (PARK; ROBERTSON; WU, 2004).

Customer satisfaction may be measured by the gap between perceived quality of the product or service, and pre-purchase quality expectations (CHOW, 2015; FORGAS et al., 2010; GUO; BARNES; JIA, 2017). Customers tend to be satisfied with an airline when service quality attributes deemed most important are met or surpassed (CHOW, 2015); such attributes represent dimensions of satisfaction (GUO; BARNES; JIA, 2017). Several studies state that customer satisfaction plays an important role in motivating customers' behavioral loyalty, which translates into giving positive reviews, returning as customers, or recommending the product or service to others (FORGAS et al., 2010; GUO; BARNES; JIA, 2017; MATTILA, 2004; MORGAN; HUNT, 1994). On the other hand, unsatisfied passengers may reconsider using the same airline in future flights (NAMUKASA, 2013), or start a negative word-of-mouth campaign (which may be electronic) that could cause damage to the company's reputation and image (BLODGETT; LI, 2007).

Measuring customer satisfaction in the airline industry is becoming more frequent and relevant (PARK; ROBERTSON; WU, 2005) ; there are at least five reasons for that (GRIGOROUDIS; SISKOS, 2009). First, customer satisfaction measurement programs could improve the communication with the clientele. Second, companies may examine whether their services fulfill customers' expectations. Moreover, companies may analyze the impact of specific efforts and actions on the clientele. Third, key satisfaction dimensions that should be improved may be identified. Fourth, companies could identify their strengths and weaknesses against competition, based on customer perceptions and judgments. Finally, companies' personnel may be motivated to increase their productivity given that all improvement efforts promoted on services are evaluated by customers themselves.

However, since the relationship between service quality dimensions and customer satisfaction may show a nonlinear pattern, assessing customer satisfaction may become a challenging task (BASFIRINCI; MITRA, 2015). In recent years, different approaches were proposed to identify and measure airline service quality and passenger satisfaction (TSAFARAKIS; KOKOTAS; PANTOUVAKIS, 2017). One research stream uses statistical techniques such as regression (e.g. logistic or ordinal) to model the relationship between quality dimensions and customer satisfaction (ALI; KIM; RYU, 2016; EBOLI; MAZZULLA, 2009;

JOSEPHAT; ISMAIL, 2012); another uses Multi-Criteria Decision Making (MCDM) methods and tools to evaluate airlines' service levels and generate propositions for improvements (LI et al., 2017; LIOU et al., 2011; LIOU; TZENG, 2007).

Conventional qualitative, quantitative or mixed methods (e.g. focus groups, questionnaire surveys, or a combination of both) are generally used by researchers and practitioners as an information source to identify and measure customers' dimensions of satisfaction (GUO; BARNES; JIA, 2017). However, these methods are time consuming and often inaccurate (WAN; GAO, 2015). For example, biased conclusions may result from limited sample sizes or inconsistent measurement items and questions (CHOW, 2015). Not all customers take questionnaires seriously, and many customers fill them out randomly, bringing noise into analyses (WAN; GAO, 2015). In addition, questions are normally set based on previous research (GUO; BARNES; JIA, 2017), compromising the uncovering of subjects not contemplated in questionnaires; e.g. emerging customer preferences (WAN; GAO, 2015).

As an alternative to conventional methods, studies from different fields have shown that user-generated content (UGC) can be used as an information source to understand customer preferences and demands (CHAU; XU, 2012). UGC results from the widespread diffusion of Web 2.0 technologies (GUO; BARNES; JIA, 2017), which enabled customers, including those from the airline industry, to share their experiences and opinions. Thousands of customers may share spontaneous, insightful, and passionate online feedback, thereby creating the "wisdom of crowds" (SUROWIECKI; SILVERMAN, 2007). More important, such feedback is widely available online and easily accessible for free or at low-cost (GUO; BARNES; JIA, 2017). One popular form of UGC is online customer reviews (OCRs) (FELBERMAYR; NANOPOULOS, 2016). Usually, OCRs are provided by customers who had direct (and usually recent) experience with a product or service; they are reported in multiple forms including online ratings (e.g. number of stars) and online reviews (e.g. personal opinion in text format) (FLANAGIN; METZGER, 2013; JIN; JI; GU, 2016; SPARKS; SO; BRADLEY, 2016).

Despite considerable attention devoted to analyzing UGCs in different industries, there is a gap in the literature regarding retrieval and information mining from such source in the airline industry. Previous studies are mostly related to sentiment analysis (LIAU; TAN, 2014; MOSTAFA, 2013; WAN; GAO, 2015) and topic detection (LIAU; TAN, 2014; SREENIVASAN; LEE; GOH, 2012), and have used UGCs from the online social networking service Twitter. On Twitter, users produce online content in the form of micro-messages called "tweets" (LIAU; TAN, 2014). Sreenivasan et al. (2012) investigated how airlines and airline

customers communicate through Twitter analyzing 9,238 tweets related to three airlines, using content analysis. Results suggested that customers use Twitter mainly to give compliments, share information, offer help, and provide personal updates; in opposition, airlines primarily use Twitter for marketing purposes. Based on their sample, the authors concluded that airlines did not appear to be responsive to users' concerns and issues.

Using a different approach, Yao, Yuan, Qian, & Li (2015) analyzed 7,466 OCRs from SkyTrax, a website specialized in reviews of the airline industry. They identified and compared the most representative terms used to refer to 25 different companies, and analyzed similarities between reviews made by users giving different star levels (1-5) to companies. Results showed that users wrote reviews using similar terms regardless of the airline, but that differences in the occurrence of terms in reviews increased with the difference in star levels.

Mostafa (2013) analyzed 2,105 tweets expressing customers' sentiments towards the service provided by 16 airlines. For that, the author analyzed the occurrence of 6,800 seed adjectives with known orientation. Results suggested that most airline services are sub-optimal since sentiment scores were overall negative. However, it is important to note that the analysis did not consider the context in which the adjectives were used.

Another study (LIAU; TAN, 2014) examined 10,895 tweets from customers of a low-cost airline. To study customers' opinions on the airline, two algorithms were tested for topic detection and two algorithms for sentiment analysis. Results showed that customers mainly discussed four topics on Twitter: customer service, ticket promotions, flight cancellations and delays, and post-booking management. Regarding sentiment analysis, customers displayed more positive than negative sentiments across the four main topics. Customers tended to write positive tweets regarding customer service, booking management and ticket promotions, and negative tweets regarding flight cancelations and delays.

More recently, another study (WAN; GAO, 2015) tested seven different classification algorithms for sentiment analysis, one of which was an ensemble algorithm based on majority voting on other six classifiers. Tests used a sample of 12,864 tweets on 16 airlines, which were manually labeled as carrying positive, neutral, and negative sentiments. The algorithms' performances were evaluated for two-class (positive and negative sentiments) and three-class (positive, neutral, and negative sentiments) datasets using a 10-fold cross-validation method. The ensemble algorithm presented the best performance for both two-class and three-class datasets, with an *F*-measure of 84.2% and 91.7%, respectively.

The present study aims to explore the "wisdom of crowds" (SUROWIECKI; SILVERMAN, 2007) contained in thousands of OCRs to provide directions for airlines companies to improve in competitiveness. More specifically, this study aims to: (*i*) identify and extract dimensions of customer satisfaction expressed in OCRs; (*ii*) verify the distribution and importance of those dimensions in OCRs from different groups of airline customers; (*iii*) identify and extract adjectives used to describe perceptions in those dimensions, and calculate the adjectives' sentiment scores; and (*iv*) test and validate the dimensions and adjectives in (*iii*) through regression analysis. By exploring OCRs, we believe that the "voice of the customer" (GASKIN et al., 1993) will be heard, since those records are written in their own language. OCRs are also organized and prioritized in the way customers think about, use, and interact with airline services.

The present study contributes to the state-of-the art on the use of spontaneous generated content to explore dimensions of customer satisfaction in at least three ways. First, to the best of our knowledge, this is the first attempt to use Latent Dirichlet Allocation (LDA) to identify and extract dimensions of customer satisfaction using airline company data. Despite the fact that LDA is considered the principal method to extract latent topics from unstructured texts (ZHAO et al., 2015), there is no evidence of using it to analyze customer satisfaction in the airline industry. Second, we propose to expand the scope of data stratification, normally limited to a specific airline and/or a particular online rating system, to include passenger characteristics and preferences (e.g. nationality, type of traveler, and cabin flown). We also propose a temporal stratification of data to enable the analysis of changes in customer demands over time. Third, this study uses a database with over 55,000 OCRs, referring to more than 400 airlines, and passengers from over 170 countries. Compared to previous studies, our data base is much larger allowing more reliable generalizations [e.g. 2,105 tweets in Mostafa (2013), 7,466 OCRs in Yao et al. (2015), 9,238 tweets in Sreenivasan et al. (2012), 10,895 tweets in Liau & Tan (2014), and 12,864 tweets in Wan & Gao (2015)].

4.2. METHODS

4.2.1. Database

We used data extracted from the website Air Travel Review (ATR), created and currently managed by SkyTrax. ATR is an independent customer forum, which has become the leading review website for airline, airport, and associated air travel passenger reviews (YAO et al., 2015). OCRs in ATR are individually registered, enabling to parse individual reviews and aggregate them for our analytical purposes. For that, we built a web crawler that collected all available OCRs in ATR on August 11th, 2016. In total, 55,775 OCRs were collected and saved in a comma-separated value (.csv) file. All OCRs were comprised of publication date, a free form text review written in English, and a yes or no answer to the question "would you recommend the airline to other passengers?". Table 4.1 presents statistics for length, in terms of number of words, of the free form text reviews.

Table 4.1: Statistics for length of reviews – number of words

Minimum	Q1	Median	Q3	Maximum	Mean
14	76	114	170	988	135

Q1: first quartile, Q3: Third quartile

Some OCRs also exhibited discriminant information, such as nationality, type of passenger, and cabin flown. A general score, ranging from one to ten, was also available in some OCRs. Table 4.2 summarizes frequencies of the available discriminant information. The database refers to 419 airlines evaluated by passengers of 171 nationalities, thus representing a wide spectrum of points of view.

Table 4.2: Discriminant information and frequencies

Variable	Frequency	%
Airline	55,775	100.00
British Airways	1,541	2.76
Spirit Airlines	1,498	2.69
United Airlines	1,361	2.44
American Airlines	1,199	2.15
Other (415 airlines)	50,176	89.96
Year of Review Publication	55,775	100.00
2002 – 2012	11,135	19.96
2013 – 2014	22,492	40.33
2015 – 2016	22,148	39.71
Passenger Nationality	54,094	96.99
United Kingdom	13,435	24.09
United States	12,054	21.61
Australia	6,333	11.35
Canada	4,126	7.40
Other (167 nationalities)	18,146	32.54
Type of Passenger	17,105	30.67
Solo Leisure	5,991	10.74
Couple Leisure	4,861	8.72
Family Leisure	3,554	6.37
Business	2,699	4.84
Cabin Flown	52,756	94.59
Economy Class	40,719	73.01
Business Class	8,667	15.54
Premium Economy Class	2,111	3.78
First Class	1,259	2.26
General Score	51,302	91.98
1	8,944	16.04
2	7,643	13.70
3	4,357	7.81
4	3,330	5.97
5	2,390	4.29
6	3,241	5.81
7	2,537	4.55
8	4,445	7.97
9	7,114	12.75
10	7,301	13.09
Passenger Recommends Airline?	55,775	100.00
Yes	29,491	52.87
No	26,284	47.13

4.2.2. Pre-Processing

According to Tirunillai & Tellis (2014) the analysis of free form text from OCRs is difficult for three main reasons. First, reviews do not have a standard structure, and most texts tend to be casual in their word and grammar usage. Second, only few words are informative regarding the product and its dimensions of quality, calling for cleaning and selection procedures. Third, text must be transformed to a format that enables numerical analysis. The main steps needed to overcome those difficulties and prepare reviews for the statistical analysis are presented next.

The pre-processing steps implemented here are similar to those adopted in previous studies (GUO; BARNES; JIA, 2017; LEE; BRADLOW, 2007; TIRUNILLAI; TELLIS, 2014). First, words containing non-English characters were converted to ASCII code (e.g. "Ljubljana Jože Pučnik Airport" was converted to "Ljubljana Joze Pucnik Airport"). Synonyms were substituted by their most frequent version in reviews. Text illustrating time, periods of time, weights, sizes, prices, dates, flight codes, airline names, airplane models, routes, countries, languages, airport names and codes, and city names and codes were replaced by key words. For example, airport names and codes were replaced by the word "_airport_" (the character "_" at the beginning and end of key words enabled the identification of added words). Table 4.3 presents the key words used and some examples of replaced texts.

Next, reviews were broken into individual sentences delimited by specific punctuation (".", "?" or "!"). The continuous string of characters of each sentence was broken down into linguistic units called tokens, which were delimited by blank spaces in the string (LUCINI et al., 2017). Part-of-speech (POS) tagging was applied to identify the syntax function of each token in sentences. For example, the use of POS tagging in the sentence "Staff was very nice" resulted in "Staff" (noun) "was" (verb) "very" (adverb) "nice" (adjective). Only adjectives and nouns were kept. Compound nouns were adjusted to become single tokens (e.g. "cabin_staff"). All capital letters were replaced by lowercases, stop words were removed (LUCINI et al., 2017), and words were reduced to their base form, excluding inflectional endings (lemmatization) (FELDMANN; SANGER, 2007). Finally, low frequency tokens (i.e. below 2%) were excluded (GUO; BARNES; JIA, 2017). After that, an original review that read:

> *"Two short flights with Adria Airways (Ljubljana - Zurich and Munich -Ljubljana), both on time, aircrafts new, cabin bright and clean. Nice and professional service on both flights. On the morning flight out of Ljubljana I did not take any free newspapers*

and just said that I will sleep. The chief flight attendant before take-off brought me a pillow for a more comfortable sleeps. Nice touch."

Became:

"short flight _airline_ _route_ _route_ on_time aircraft new cabin bright clean nice professional flight morning flight _city_ free newspaper chief cabin_staff pillow comfortable sleep nice touch"

Table 4.3: Key words and examples of replaced texts

Key word	Examples of replaced text	Key word	Examples of replaced text	Key word	Examples of replaced text
time	10 am	_weight_	1 kg	_price_	$ 500
	10:00 am		20,5 g		£ 20
	14:30		2.3 lbs		€ 300
	10.00 pm				12 dollars
					30 euros
_period_time_	10-15 minutes	_airplane_	A380		20 CAD
	1 to 2 days		Boeing 787-800		AUD$660
	3 and a half hours		ERJ-190		
	2.5 months		Dash 8-400	_language_	English
	2 weeks				German
		route	POA-CGH		Portuguese
_flight_code_	AC4601		POA to CGH		
	DL 471		London – New York	_airport_	Heathrow
			POA-GRU-MIA		John F Kennedy
airline	British Airways		POA to MIA via		JFK
	Air Canada		GRU		
	Iberia			_size_	2 cm
		date	on march 27		1.5 m
country	Brazil		on june 4th		2,3 in
	USA		march 27		4 inches
	Australia		on the 13th october		2 ft
			30th december 2015		4 feet
city	London		5th may		6' 2"
	New York		26 april		
	NYC		26 april 2017		
			march 16 2015		
			05/03/2011		
			on june 4th		

The information of the syntax function of each token was kept allowing the selection of tokens to attend the different objectives of the study. For example, to identify and extract dimensions of customer satisfaction only nouns were kept in records. On the other hand, only adjectives were kept to calculate sentiment scores of terms used to describe perceptions of satisfaction dimensions. Finally, nouns and adjectives were used to validate findings through regression analysis.

Once all reviews had been pre-processed, they were converted to set-of-words representation. For that, a matrix was created indicating the occurrence of tokens in reviews. Matrix columns corresponded to tokens and rows corresponded to reviews. Matrix cells were filled using the term frequency – inverse document frequency (TF-IDF) value (LUCINI et al., 2017). All pre-processing steps were carried out using scikit-learn (HACKELING, 2014) and NLTK (BIRD; KLEIN; LOPER, 2009) modules in Python 3.6 (PYTHON SOFTWARE FOUNDATION, 2017).

4.2.3. Identification and extraction of dimensions of customer satisfaction

One objective in this study is to identify and extract potential dimensions influencing customer satisfaction in the airline industry. To achieve that goal we used a popular topic detection method from the field of machine learning and natural language processing called Latent Dirichlet Allocation (LDA). LDA assumes that each document (i.e. review) can be represented as a probabilistic distribution over latent topics (i.e. dimensions of customer satisfaction), and that topic distribution in all documents share a common Dirichlet prior. Each latent topic in the LDA model is represented as a probabilistic distribution over words, and the word distributions in topics also share a common Dirichlet prior. Topics can potentially be shared by all reviews, and every review has its own mixing proportion of topics. Formally, given a corpus D consisting of M documents, with document d having N_d words ($d \in \{1, \dots, M\}$), LDA models D according to the following generative process (BLEI; NG; JORDAN, 2003):

a) Select a multinomial distribution $\emptyset_t$ for topic t ($t \in \{1, \dots, M\}$) from a Dirichlet distribution with parameter β;
b) Select a multinomial distribution θ_d for document $d (d \in \{1, \dots, M\})$ from a Dirichlet distribution with parameter α;

c) For a word w_n ($n \in \{1, \ldots, N_d\}$) in document d,
   a. select a topic z_n from θ_d;
   b. select a word w_n from $\emptyset_{z_n}$.

Words in documents are the only observed variables; others are latent variables ($\emptyset$ and θ) and hyper parameters (α and β) to be estimated. For that, the probability of observed data *D* is computed and maximized as follows (BLEI; NG; JORDAN, 2003):

$$p(D|\alpha,\beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha)\left(\sum_{n=1}^{N_d} p(z_{dn}|\theta_d)p\left(w_{d_n}\middle|z_{dn,\varphi}\right)P(\varphi|\beta)\right)d\theta_d\, d\varphi \quad (1)$$

Due to the coupling between $\emptyset$ and θ in the integrand in eq. (1), exact inference in LDA is intractable; approximate algorithms such as the variational inference (AIROLDI et al., 2008; BLEI; JORDAN, 2003; HOFFMAN; BLEI; BACH, 2010) or the Markov Chain Monte Carlo (GRIFFITHS; STEYVERS, 2004; SHIVASHANKAR et al., 2011; TIRUNILLAI; TELLIS, 2014) are typically used for that. In this study we used the Batch Variational Bayes Inference (BVBI) algorithm (HOFFMAN; BLEI; BACH, 2010), available in the package scikit-learn (HACKELING, 2014). For more detailed descriptions of LDA and BVBI see Blei et al. (2003) and Hoffman et al. (2010), respectively.

Determining the ideal number of topics in a LDA model is a challenging task when the number of dimensions of satisfaction is not known *a priori* (ZHAO et al., 2015; ZHAO; ZOU; CHEN, 2014). An insufficient number of topics could render a model that is too coarse to identify accurate dimensions. On the other hand, an excessive number of topics could result in a model that is too complex, making interpretation and subjective validation difficult (ZHAO et al., 2015). We followed the perplexity-based method (ZHAO et al., 2015) to determinate the ideal number of latent topics in our dataset. This is an iterative approach in which different models with different number of topics are tested and compared using the perplexity measure. Perplexity is a popular measurement in information theory, used to evaluate how well a statistical model describes a dataset (ZHAO et al., 2015); lower perplexity denotes a better probabilistic model. Formally, for a test set of *M* documents the perplexity measure is defined as:

$$perplexity_{D_{test}} = exp\left\{-\frac{\sum_{d=1}^{M} \log p(w_d)}{\sum_{d=1}^{M} N_d}\right\} \quad (2)$$

The dataset was divided in train and test sets to allow k-fold cross-validation considering models ranging from 2 to 100 topics. The dataset was divided in k mutually exclusive subsets of equal size, such that one subset is used for perplexity measurement and $k - 1$ subsets are used for parameter estimation. This process is carried out k times alternating the test subset; performance statistics are calculated from the results (HASTIE; TIBSHIRANI; FRIEDMAN, 2008). In this study, k was set to 5. The ideal number of topics is given by the model with fewer topics among those that overlap the confidence interval of the model with the smallest mean perplexity.

Once the number of topics is identified, two steps are needed to extract and identify dimensions of customer satisfaction expressed in OCRs (TIRUNILLAI; TELLIS, 2014). First, the LDA model is calibrated using the entire dataset and the established number of topics. Then, important attributes of each dimension of satisfaction are identified (i.e. words closely related to each dimension). In that sense, words that have the highest likelihood to belong to a specific topic are considered important. The top ten words of each topic are listed according to the likelihood scores of the fitted model. In the second step each topic is named with a dimension of satisfaction, based on its list of important words. Following (GUO; BARNES; JIA, 2017), the naming of dimensions is conducted by two researchers individually, and results are compared and discussed until they reach consensus. Each researcher subjectively tries to identify logical connections between the top important words (e.g. the top three) for a topic, giving a candidate name. Then the candidate name is further tested subjectively via logical connection with other words in the top-10 list. If a connection is found the topic name is retained; if a word that does not fit the topic name is found, the naming process restarts and keeps going until a name logically connected to all words is devised. Final names assigned to topics are considered as dimensions of customer satisfaction expressed in the OCRs.

4.2.4. Distribution of customer satisfaction dimensions

Once the LDA model is fitted and dimensions of customer satisfaction are identified and extracted, it is possible to evaluate the distribution of those dimensions across different groups of reviews. As a first step, all reviews are analyzed individually in order to calculate the distribution of probabilities of the identified dimensions. Next, reviews are grouped according to five different criteria: airline, publication year, passenger nationality, type of passenger, and

type of cabin flown. Table 4.2 presents the strata used for each criterion. Finally, analysis of confidence intervals is carried out to identify differences and similarities among strata.

4.2.5. Sentiment Analysis

To identify, extract and calculate sentiment scores of adjectives that are normally used by airlines' customers, the pre-processed dataset was adjusted to keep only adjectives in reviews. We propose a method to analyze the sentiments of adjectives based on the scores of a Naïve Bayes Classifier (DUDA; HART; STORK, 2001). In that sense, some reviews were classified as presenting "negative" or "positive" sentiments. We considered "negative" reviews those with a general score of one (the lowest score available, occurring in 8,944 OCRs). On the other hand, reviews with a general score of ten (the highest score available, occurring in 7,301 OCRs) were considered "positive" reviews. All other reviews with general scores ranging from two to nine were not considered for classification. In addition, adjectives that appeared in less than 0.1% of the documents were disregarded from the analysis.

The Naïve Bayes Classifier was calibrated resulting in sentiment scores for each of the adjectives, which were normalized to facilitate interpretation. Normalized scores ranged from –1 to +1, with –1 and +1 denoting a very negative and very positive sentiment, respectively. As a result, the method provided a list of most frequent and important adjectives used by airline customers when describing their sentiment in reviews, as well as their sentiment scores.

4.2.6. Regression analysis

To test and validate the dimensions of satisfaction and adjectives extracted from OCRs, we propose the use of a Logistic Regression Classifier (ZHANG et al., 2016) to predict the recommendation of airlines. For that, we used the entire pre-processed dataset including nouns and adjectives, along with customers' final recommendation. Independent variables in the logistic regression were dimensions in reviews, their importance and sentiment scores. Finally, the dataset was divided in training and testing sets to allow a k-fold cross-validation, with k set to 10. The model's prediction performance was assessed through accuracy measurements. In addition, the most relevant dimensions for the prediction of airline recommendation prediction were determined considering the entire dataset.

4.3. RESULTS

4.3.1. Identification and extraction of dimensions of customer satisfaction

Confidence intervals were calculated for each of the 99 adjusted LDA models (considering models with 2 to 100 topics). The lowest perplexity mean of 72.65 (95% CI: 71.28 – 74.03) was achieved using the model with 85 topics. However, the model with smallest number of topics overlapping the CI of the 85-topic model had 67 topics. We thus considered the ideal number of topics in our dataset to be 67, with a perplexity mean score of 74.12 (95% CI: 72.46 – 75.77). Figure 4.1 displays the mean perplexity as a function of the number of topics in the model.

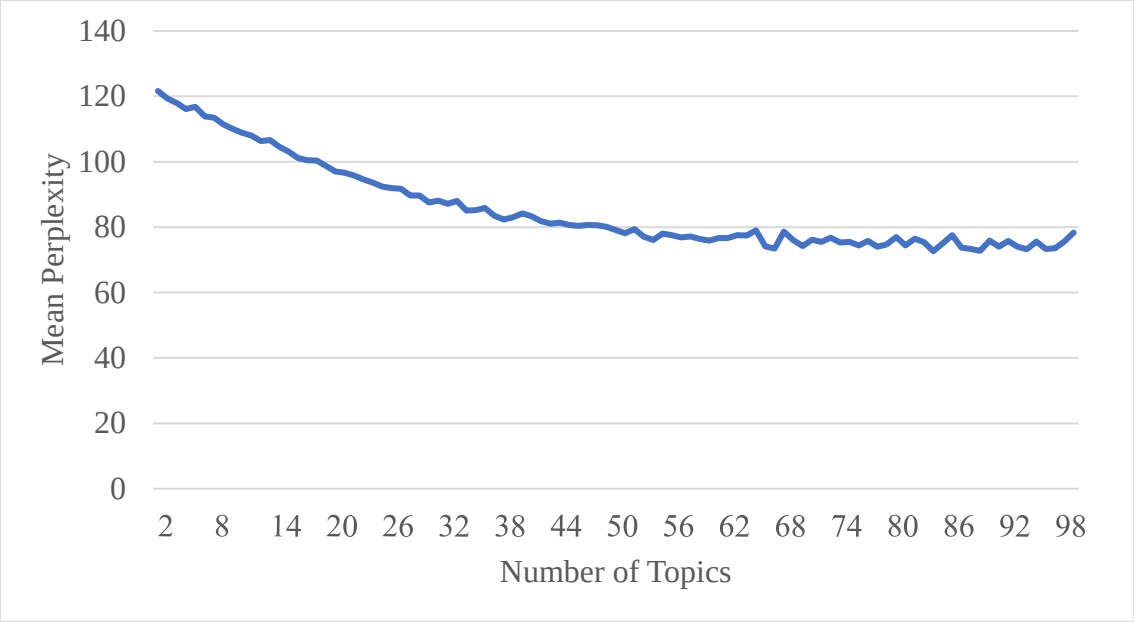

Figure 4.1: Mean perplexity as a function of the number of topics

Once defined the number of topics, the entire dataset was used to calibrate the LDA model. All topics were then named according to the procedure previously described. For example, in Table 4.4 the topic named “checking luggage” is based mainly on the word “luggage” (weight = 0.3964), which appears at the top of the list. Considering that all other words could be related to the act of checking luggage, the topic was named as such. Next, we grouped topics that had a similar content and meaning, and discarded three topics deemed not

relevant (i.e. those with probability 0.05% or less over topics distribution). The final result was 27 dimensions of satisfaction, which are given in Figure 4.2.

Table 4.4: Example of attributes and scores in topic "Checking luggage"

Attributes	Score
luggage	0.3964
__time__	0.1912
__price__	0.1745
person	0.0775
fee	0.0693
flight	0.0448
__period_time__	0.0215
time	0.0190
__route__	0.0045
__city__	0.0010

4.3.2. Distribution and importance of customer satisfaction dimensions

All reviews were analyzed individually and grouped according to five different criteria. 95%-confidence intervals were calculated for each criteria and dimension. Figures 3 to 7 show graphs, with confidence intervals, for the five criteria (airline, publication year, passenger nationality, type of passenger, and type of cabin flown).

Regarding the airline criterion, there are significant differences in the distribution of some dimensions (Figure 4.3). For example, passengers of British Airways write more about food and drink, onboard services, problems, business class, airplane characteristics, couple flights, comfort and airport lounge, when compared to other companies. On the other hand, they write significantly less about connection information, delays, passenger interaction, checking luggage, waiting times, prices at airports, airline prices, gate services, and family flights. Spirit Airlines passengers prefer to write about checking luggage, prices in airports, and airline prices, and write less about onboard service, onboard entertainment, business class, cabin staff,

airplane characteristics, and passengers demands (Spirit Airlines is a low cost company that operates only in the United States). Finally, American Airlines and United Airlines passengers have similar distributions across all dimensions.

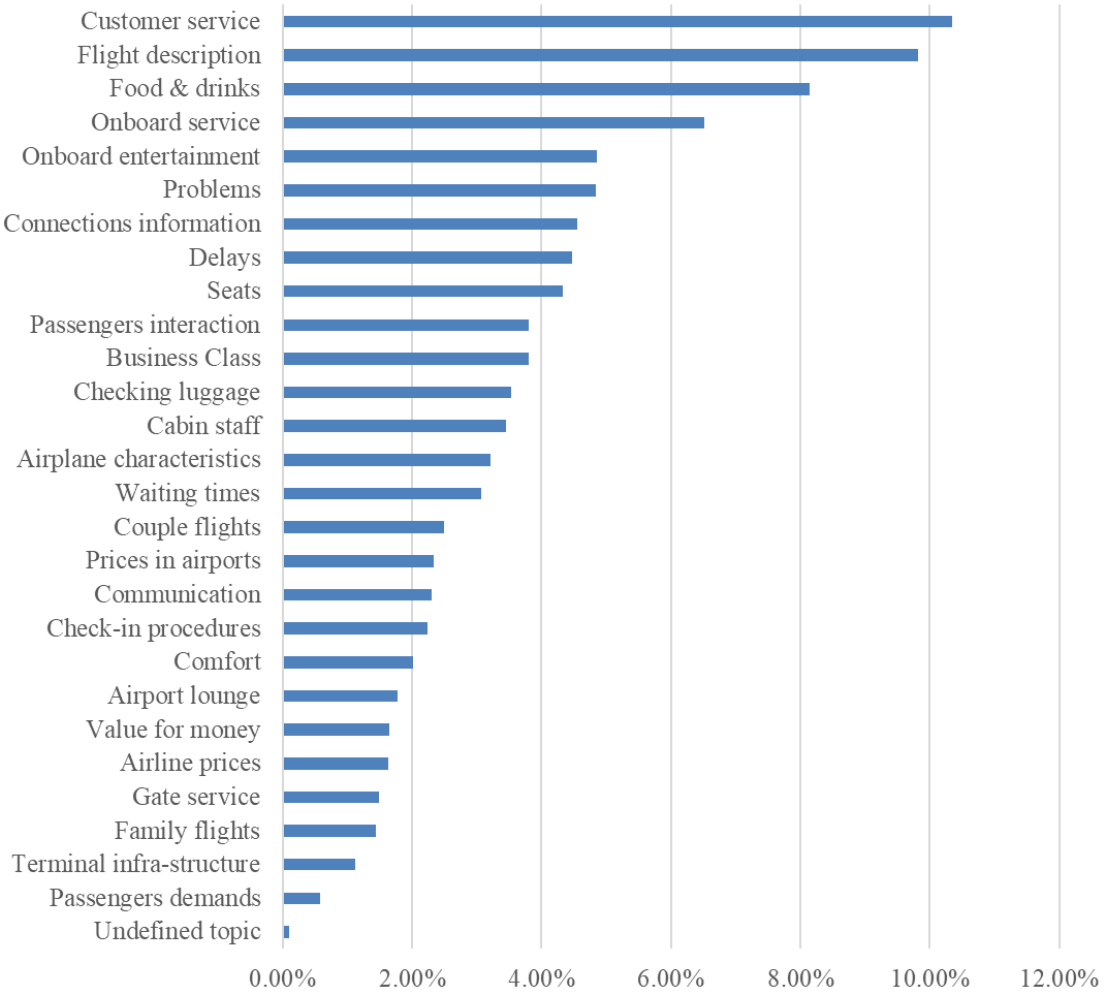

Figure 4.2: Satisfaction dimensions and their distribution over the entire dataset

When analyzing the distribution of dimensions in different periods of time (Figure 4.4), it is possible to notice that many dimensions have similar proportions regardless of the year in which they were written. However, it is possible to verify some trends in the data. For example, passengers wrote less about flight descriptions, food and drinks, and airplane characteristics in recent years. In contrast, they are writing more about problems, passenger interaction, and checking luggage.

The distribution of dimensions considering passenger nationalities (Figure 4.5) point to passengers giving attention to different dimensions, although some similarities among

nationalities could be identified. For example, Americans and Canadians have similar behavior, which is significantly different from the British and Australians when writing about food and drink, onboard service, connection information, passenger interaction, airplane characteristic, and value for money. In contrast, all analyzed nationalities write similarly about seats, communication, check-in procedures, comfort, and terminal infra-structure.

Figure 4.6 shows that most dimensions present similar importance to different types of passenger. However, compared to other types, passengers flying on business trips write more about business class and airport lounge dimensions. Also, passengers traveling with their families write more about family flights, as expected. Solo leisure and business passengers give similar attention to the couple flights dimension, writing significantly less about this dimension when compared to couple and family leisure passengers.

Regarding the type of cabin flown criterion, Figure 4.7 shows that customer service is significantly more important to passengers flying in first class than to other passengers. On the other hand, economy class passengers write less about this dimension. Economy class passenger write significantly more about delays, checking luggage, waiting times, and prices in airports. In addition, economy premium class passengers give significantly more attention to couple flights and comfort dimensions.

4.3.3. Sentiment Analysis

To identify, extract and calculate sentiment scores of adjectives that are normally used by airline customers, the dataset was reduced to contain only reviews with a general score of 1 (representing very negative sentiment) and 10 (representing very positive sentiment). A Naïve Bayes Classifier was fitted to the reduced dataset, yielding a training accuracy of 93.68%. Considering that only adjectives present in at least 0.1% of reviews were kept, the final number of adjectives analyzed was 882. To facilitate interpretations, each adjective's Naïve Bayes Classifier score was rescaled to vary from –1 (very negative sentiment) to +1 (very positive sentiment). Table 4.5 presents the top 20 adjectives for negative and positive sentiments, along with their respective sentiment scores. Note that the top 16 negative adjectives present a sentiment score of –1, corresponding to the most negative sentiment.

Table 4.5: Top 20 adjectives representing negative and positive sentiments

Negative sentiments		Positive sentiments	
Term	**Sentiment score**	**Term**	**Sentiment score**
absurd	–1.00	good	1.00
abysmal	–1.00	excellent	0.93
dismissive	–1.00	great	0.91
disrespectful	–1.00	comfortable	0.91
incompetent	–1.00	friendly	0.88
incorrect	–1.00	nice	0.82
inexperienced	–1.00	new	0.82
not_alternative	–1.00	clean	0.79
not_clear	–1.00	best	0.77
stale	–1.00	helpful	0.76
stranded	–1.00	free	0.74
unapologetic	–1.00	attentive	0.74
uncaring	–1.00	professional	0.73
unfair	–1.00	pleasant	0.73
unorganized	–1.00	efficient	0.73
unsympathetic	–1.00	fantastic	0.67
unacceptable	–0.60	better	0.65
broken	–0.57	easy	0.65
inadequate	–0.55	quick	0.63
unwilling	–0.55	wonderful	0.61

4.3.4. Regression analysis

The dimensions of satisfaction and adjectives extracted from OCRs were tested and validated using a Logistic Regression Classifier to predict the recommendation of airlines. After

data preparation, the dataset was divided in training and testing sets to allow k-fold cross-validation, with k set to 10. The mean accuracy of the testing set was 79.95% (95% CI: 79.93 – 79.97). The most relevant dimensions for the prediction of airline recommendation were calculated and are presented in Table 4.6. The top three dimensions with highest coefficients were “cabin staff” (8.58), “onboard service” (7.77), and “value for money” (6.24); the bottom three dimensions with lowest coefficients were “checking luggage” (0.46), “connections information” (1.39), and “flight description” (2.11).

4.4. DISCUSSION AND CONCLUSIONS

Identifying the dimensions of customer satisfaction is determinant to accurately evaluate how passengers appraise airline services (PARK; ROBERTSON; WU, 2004). Using spontaneous user-generated content (UGC) is believed to lower the biases of artificial responses given by customers to traditional research tools, such as focus groups and questionnaire surveys.

To analyze UCGs from customers of the airline industry, we have employed a highly reliable method to identify the dimensions of their satisfaction – the Latent Dirichlet Allocation (LDA). LDA is a topic detection method from the field of machine learning and natural language processing. It is considered to be one of the main methods to extract latent topics from unstructured texts. There is no evidence, however, of its use to analyze customer satisfaction dimensions in the airline industry.

LDA determines the dimensions of customer satisfaction in a dataset. In order to identify the number of dimensions, which was not known *a priori*, we have used the perplexity-based method in our set of UCGs. Perplexity is a measure in information theory that is used to evaluate how well a statistical model describes a dataset. After identifying the number of dimensions, the LDA model has been calibrated, based on the whole dataset. Important attributes of each dimension were then identified, grouping words related to them. The bias of this grouping process is low, since it is based on spontaneous experience reports from customers (UGCs). We used the Naïve Bayes Classifier to analyze the sentiment scores of adjectives, labeling them as positive or negative. Finally, we have used a Logistic Regression Classifier to predict the recommendation of airlines. The regression analysis allowed us to validate the dimensions of customer satisfaction and adjectives as predictors of recommendation.

Table 4.6: Coefficient of dimensions in the logistic regression decision function

Dimension	Coefficient
Cabin staff	8.58
Onboard service	7.77
Value for money	6.24
Seats	6.20
Couple flights	5.85
Passengers demands	5.51
Airplane characteristics	5.20
Airport lounge	5.14
Customer service	4.24
Food & drinks	4.21
Onboard entertainment	4.17
Communication	4.16
Gate service	3.91
Business Class	3.57
Delays	3.54
Comfort	3.19
Family flights	3.12
Problems	2.90
Terminal infra-structure	2.77
Waiting times	2.48
Check-in procedures	2.46
Prices in airports	2.30
Passengers interaction	2.29
Airline prices	2.16
Flight description	2.11
Connections information	1.39
Checking luggage	0.46

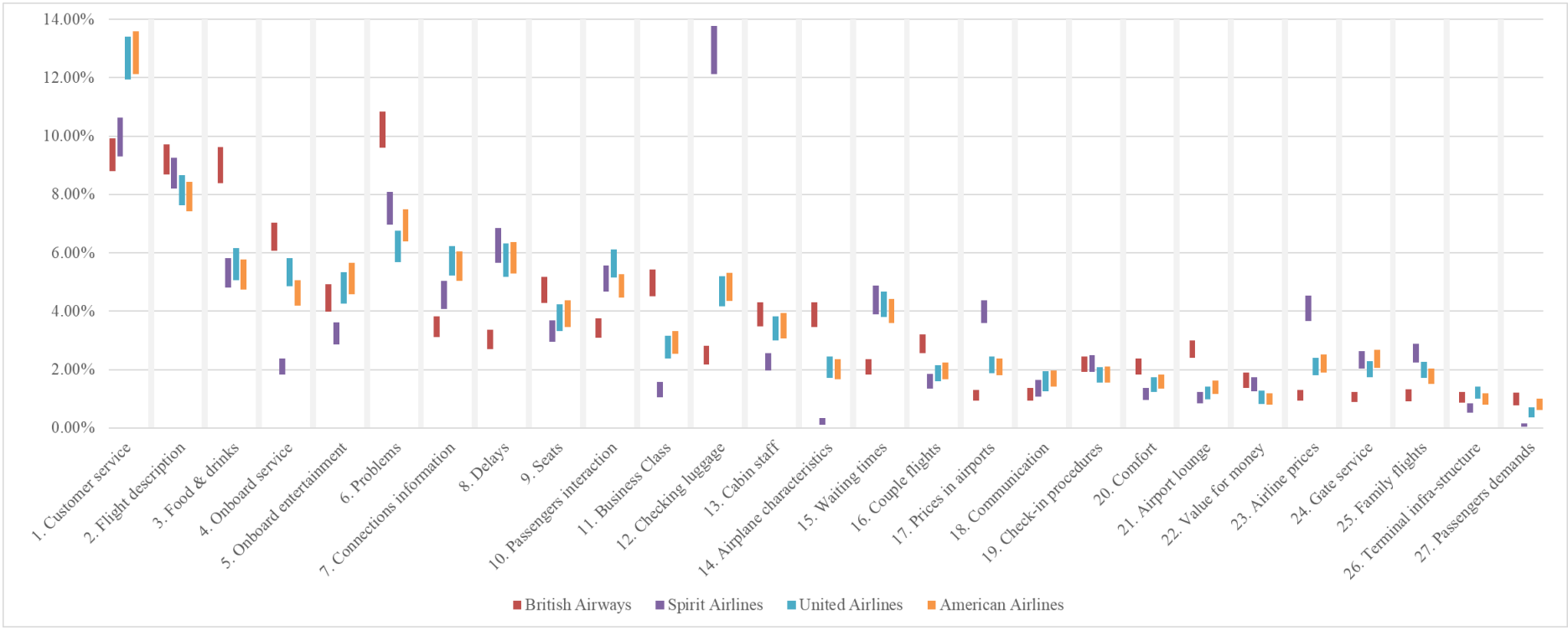

Figure 4.3: Confidence intervals for airline criterion in data stratification

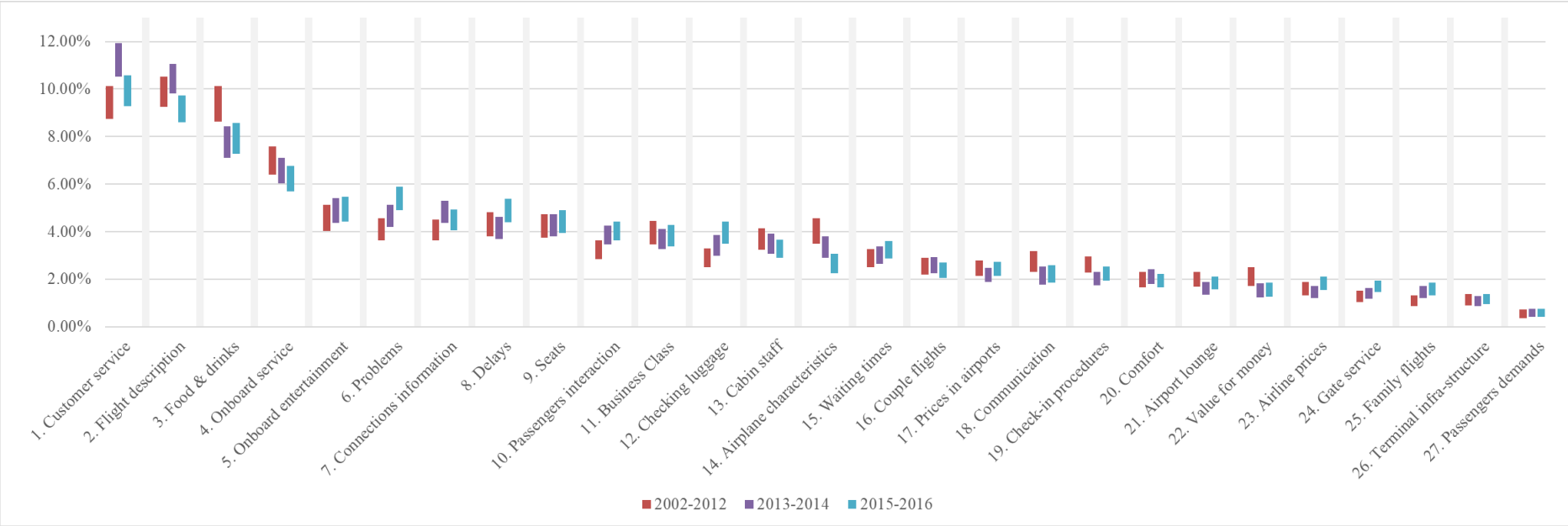

Figure 4.4: Confidence intervals for publication year criterion in data stratification

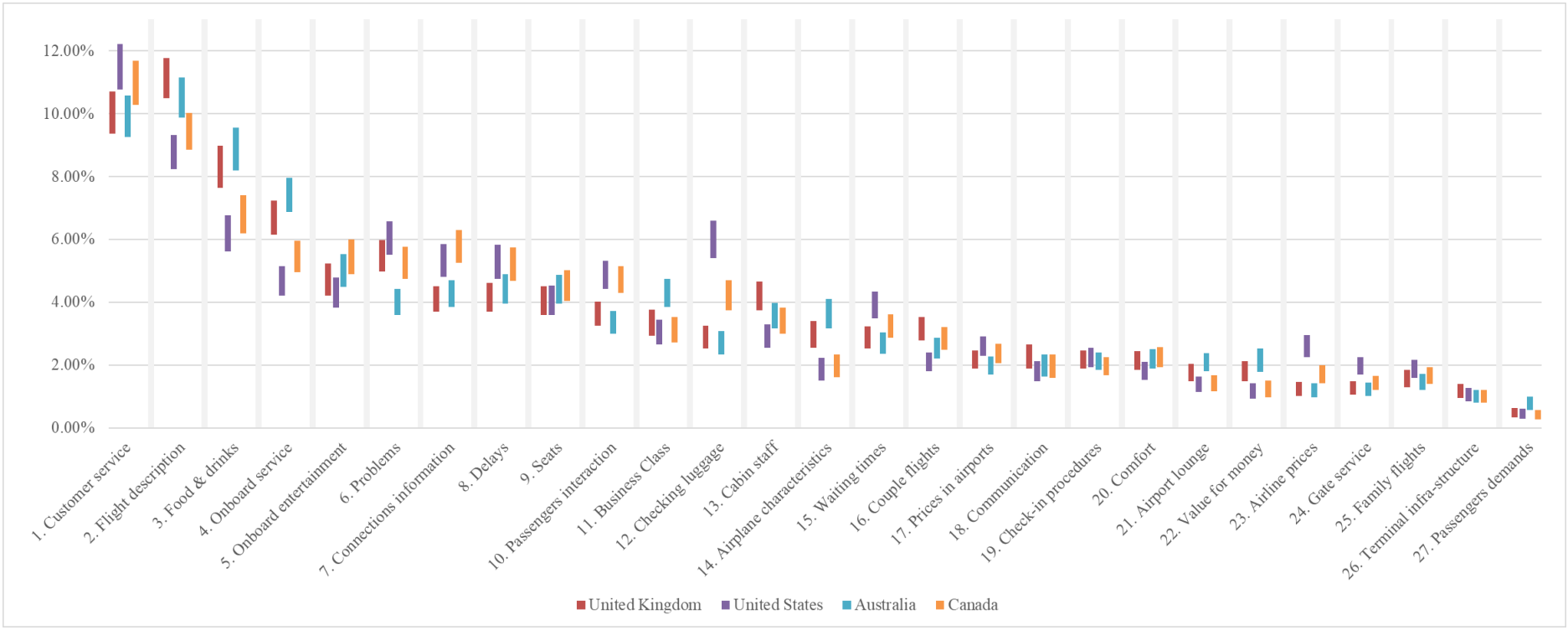

Figure 4.5: Confidence intervals for passenger's nationality criterion in data stratification

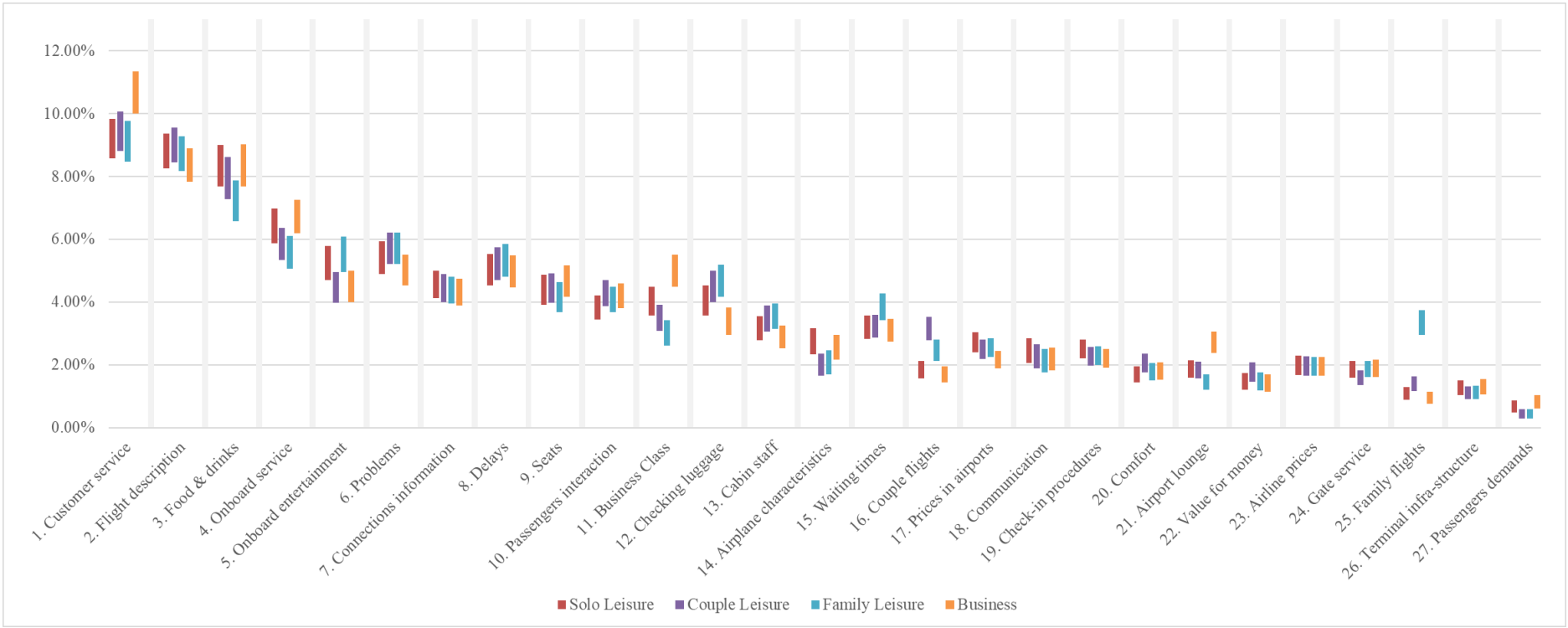

Figure 4.6: Confidence intervals for type of passenger criterion in data stratification

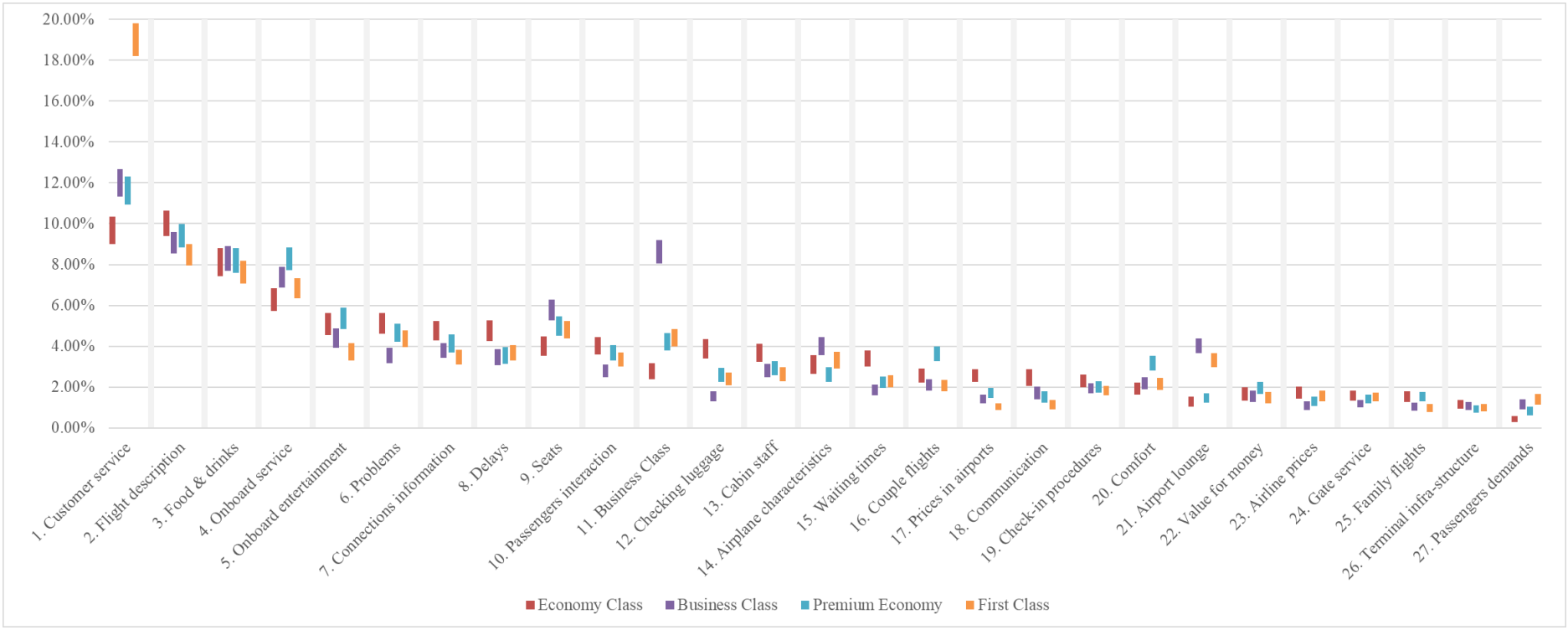

Figure 4.7: Confidence intervals for cabin flown criterion in data stratification

In previous studies (CHOW, 2015), the context of use of adjectives was not evaluated. Our results represent a step forward in analyzing UGCs, since they are segmented by service characteristics and customer profile. Service characteristics (airline, publication year, and type of cabin flown) and customer profiles (type of passenger and passenger nationality) were related to users' levels of satisfaction and recommendation of airlines.

Comparing results by airline, differences in distribution of some dimensions of customer satisfaction were identified. Even though general dimensions of satisfaction were drawn from the dataset, this result suggests that our method is sensitive to detect dissimilarities between different services. Likewise, equivalent services (e.g., from American Airlines and United Airlines) are also identifiable through this method. The sensitiveness of the measurement we made indicates that our method does not only generate general results to be used in academic research, but it also works as a framework for airline service providers to apply in specific companies.

Our method is also sensitive to show customer trends through time, by observing results in different publication years. On one hand, decreasing attention has been given by customers to flight descriptions, food and drinks, and airplane characteristics in recent years. On the other hand, dimensions such as passenger interaction and checking in luggage are crescent focuses of the reviews observed in UGCs. These results indicate that companies are aware of and aligned with customers' needs regarding some of the dimensions of satisfaction. Nevertheless, important issues are being more and more neglected, e.g., luggage restrictions. Even if companies choose not to hear the "wisdom of crowds", e.g., for profit purposes, our results can raise their awareness regarding the relevance of the neglected dimensions of customer satisfaction. Therefore, airline service providers may make informed decisions about their services.

The type of cabin flown has shown to influence the distribution of the dimensions of satisfaction. This result allows airline service providers to focus their efforts on dimensions that are relevant to segmented types of customers (e.g., customer service to passengers that fly in first class, comfort to those who fly premium economy, and checking luggage and waiting time to users flying economy class). Differently, type of passenger (users flying on business trips, solo leisure, and couples and families) was the segmentation variable which impacted the smallest number of dimensions of customer satisfaction.

Passenger nationality has also influenced the distribution of the dimensions of satisfaction. Apart from the evident use of this information to shape services locally, results

have shown that our method was able to identify cultural differences. Even though some dimensions are culture-sensitive, as seen in Figure 4.5, passengers write similarly about seats, communication, check-in procedures, comfort, and terminal infra-structure.

Regarding sentiment scores of adjectives, only “good” hit the highest level possible (+1) according to our measurement method (Naïve Bayes Classifier). This word is not specific of any domain, such as caring behavior or level of experience of cabin crew. Differently, many negative sentiments hit the lowest grade possible (-1) and they refer to specific themes: overall experience (absurd, abysmal, not_alternative, not_clear), uncaring behavior (dismissive, disrespectful, unapologetic, uncaring, unfair), lack of knowledge or experience of service providers (incompetent, incorrect, inexperienced, unorganized, unsympathetic) and general updateness (stale, stranded). The observation of the strength of these negative sentiments offer many inputs to design services to avoid negative experiences.

Finally, positive reviews and recommendations from customers are commonly as results of customer satisfaction (FORGAS et al., 2010; GUO; BARNES; JIA, 2017; MATTILA, 2004; MORGAN; HUNT, 1994). Observing the UGCs in our dataset, the top three dimensions of satisfaction to predict the recommendation of airlines were cabin staff, onboard service, and value for money. Designing services that meet or surpass expectations within these dimensions is key to recommendation (CHOW, 2015).

There are some limitations to this study. They are presented as follows, including a future agenda for research in the field of airline services.

First, the analysis is restricted to UGCs and no relations between these reviews and data generated by conventional methods (e.g. survey questionnaires) are made. Therefore, future studies may relate both types of data to reinforce the belief that analyzing UGC provides more accurate results than other research tools to predict airline recommendation.

Second, we have adopted a single method of topic detection from the field of machine learning and natural language processing (Latent Dirichlet Allocation – LDA). The application of other methods, such as Paragraph Vectors (HASHIMOTO et al., 2016) and Hierarchical Recurrent Neural Network (LU et al., 2017), offers an opportunity for future research.

Third, in our study the measure of satisfaction is the recommendation of airline services. Further measures may be adopted in future research, since we did not compare customers’ pre-purchase expectations and post-purchase evaluations (CHOW, 2015; FORGAS et al., 2010; GUO; BARNES; JIA, 2017), and the recurrence of consumption was also not

assessed (FORGAS et al., 2010; GUO; BARNES; JIA, 2017; MATTILA, 2004; MORGAN; HUNT, 1994).

This research represents a step forward in measuring customer satisfaction. Our method allows extracting dimensions of satisfaction from UGCs (text), providing results similar to factors that would be obtained through factor analysis (quantitative data). It is an innovative way of measuring customer satisfaction that is suitable for future academic and applied studies.

4.5. REFERENCES

AIROLDI, E. M. et al. Mixed Membership Stochastic Blockmodels. **Journal of Machine Learning Research**, v. 9, n. 2008, p. 1981–2014, 2008.

ALI, F.; KIM, W. G.; RYU, K. The effect of physical environment on passenger delight and satisfaction: Moderating effect of national identity. **Tourism Management**, v. 57, p. 213–224, 2016.

BASFIRINCI, C.; MITRA, A. A cross cultural investigation of airlines service quality through integration of Servqual and the Kano model. **Journal of Air Transport Management**, v. 42, p. 239–248, 2015.

BIRD, S.; KLEIN, E.; LOPER, E. **Natural Language Processing with Python**. 1º edition ed. Sebastopol, CA: O'Reilly Media, 2009.

BLEI, D. M.; JORDAN, M. I. **Modeling Annotated Data**. Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. **Anais**...ACM, 2003

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent Dirichlet Allocation. **Journal of Machine Learning Research**, v. 3, p. 993–1022, 2003.

BLODGETT, J. G.; LI, H. Assessing the effects of post-purchase dissatisfaction and complaining behavior on profitability: a Monte Carlo simulation. **Journal of Consumer Satisfaction, Dissatisfaction & Complaining Behavior**, v. 20, p. 1–14, 2007.

CALISIR, N.; BASAK, E.; CALISIR, F. Key drivers of passenger loyalty: A case of Frankfurt-Istanbul flights. **Journal of Air Transport Management**, v. 53, p. 211–217, 2016.

CHAU, M.; XU, J. Business intelligence in blogs: Understanding consumer interactions and

communities. **MIS Quarterly: Management Information Systems**, v. 36, n. 4, p. 1189–1216, 2012.

CHEN, C.-F. Investigating structural relationships between service quality, perceived value, satisfaction, and behavioral intentions for air passengers: Evidence from Taiwan. **Transportation Research Part A**, v. 42, n. 4, p. 709–717, maio 2008.

CHOW, C. K. W. On-time performance, passenger expectations and satisfaction in the Chinese airline industry. **Journal of Air Transport Management**, v. 47, p. 39–47, 2015.

DOLNICAR, S. et al. Key drivers of airline loyalty. **Tourism Management**, v. 32, n. 5, p. 1020–1026, 2011.

DUDA, R.; HART, P.; STORK, D. **Pattern Classification**. 2nd ed ed. New York: Wiley-Interscience, 2001.

EBOLI, L.; MAZZULLA, G. An ordinal logistic regression model for analysing airport passenger satisfaction. **EuroMed Journal of Business**, v. 4, n. 1, p. 40–57, 2009.

FELBERMAYR, A.; NANOPOULOS, A. The Role of Emotions for the Perceived Usefulness in Online Customer Reviews. **Journal of Interactive Marketing**, v. 36, p. 60–76, 2016.

FELDMANN, R.; SANGER, J. **The text mining handbook: advanced approaches in analyzing unstructured data**. New York: Cambridge university press, 2007.

FLANAGIN, A. J.; METZGER, M. J. Trusting expert- versus user-generated ratings online: The role of information volume , valence , and consumer characteristics. **Computers in Human Behavior**, v. 29, n. 4, p. 1626–1634, 2013.

FORGAS, S. et al. Antecedents of airline passenger loyalty: Low-cost versus traditional airlines. **Journal of Air Transport Management**, v. 16, n. 4, p. 229–233, 2010.

GASKIN, S. P. et al. The Voice of the customer. **Marketing Science**, v. 12, n. 1, p. 1–27, 1993.

GRIFFITHS, T. L.; STEYVERS, M. Finding scientific topics. **Proceedings of the National Academy of Sciences of the United States of America**, v. 101 Suppl, p. 5228–35, 2004.

GRIGOROUDIS, E.; SISKOS, Y. **Customer satisfaction evaluation: methods for measuring and implementing service quality**. [s.l.] Springer Science & Business Media, 2009.

GUO, Y.; BARNES, S. J.; JIA, Q. Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. **Tourism Management**, v. 59, p. 467–

483, 2017.

HACKELING, G. **Mastering Machine Learning With scikit-learn**. Birmingham, UK: Packt Publishing Ltd., 2014.

HASHIMOTO, K. et al. Topic detection using paragraph vectors to support active learning in systematic reviews. **Journal of Biomedical Informatics**, v. 62, p. 59–65, 2016.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning: data mining, inference, and prediction**. Second edi ed. [s.l.] Springer, 2008.

HOFFMAN, M. D.; BLEI, D. M.; BACH, F. Online Learning for Latent Dirichlet Allocation. **Advances in Neural Information Processing Systems**, p. 856–864, 2010.

JIN, J.; JI, P.; GU, R. Identifying comparative customer requirements from product online reviews for competitor analysis. **Engineering Applications of Artificial Intelligence**, v. 49, p. 61–73, 2016.

JOSEPHAT, P.; ISMAIL, A. A Logistic Regression Model of Customer Satisfaction of Airline. **International Journal of Human Resource Studies**, v. 2, n. 4, p. 255, 2012.

LEE, T. Y.; BRADLOW, E. T. **Automatic Construction of Conjoint Attributes and Levels from Online Customer Reviews**, 2007.

LI, W. et al. A hybrid approach based on fuzzy AHP and 2-tuple fuzzy linguistic method for evaluation in-flight service quality. **Journal of Air Transport Management**, v. 60, p. 49–64, 2017.

LIAU, B. Y.; TAN, P. P. Gaining customer knowledge in low cost airlines through text mining. **Industrial Management & Data Systems**, v. 114, n. 9, p. 1344–1359, 2014.

LIOU, J. J. H. et al. Using a modified grey relation method for improving airline service quality. **Tourism Management**, v. 32, n. 6, p. 1381–1388, 2011.

LIOU, J. J. H.; TZENG, G.-H. A non-additive model for evaluating airline service quality. **Journal of Air Transport Management**, v. 13, n. 3, p. 131–138, 2007.

LU, H. et al. **Don't Forget the Quantifiable Relationship between Words : Using Recurrent Neural Network for Short Text Topic Discovery**. Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17). **Anais**...2017

LUCINI, F. R. et al. Text mining approach to predict hospital admissions using early medical records from the emergency department. **International Journal of Medical Informatics**, v.

100, p. 1–8, 2017.

MATTILA, A. S. The impact of service failures on customer loyalty. **International Journal of Service Industry Management**, v. 15, n. 2, p. 134–149, 2004.

MORGAN, R. M.; HUNT, S. D. The commitmemt-trust theory of relationship marketing. **Journal of Marketing**, v. 58, n. 3, p. 20–38, 1994.

MOSTAFA, M. M. An emotional polarity analysis of consumers' airline service tweets. **Social Network Analysis and Mining**, v. 3, n. 3, p. 635–649, abr. 2013.

NAMUKASA, J. The influence of airline service quality on passenger satisfaction and loyalty. **The TQM Journal**, v. 25, n. 5, p. 520–532, 2013.

PARK, J.-W.; ROBERTSON, R.; WU, C.-L. The effect of airline service quality on passengers' behavioural intentions: A Korean case study. **Journal of Air Transport Management**, v. 10, n. 6, p. 435–439, nov. 2004.

PARK, J.-W.; ROBERTSON, R.; WU, C.-L. Investigating the Effects of Airline Service Quality on Airline Image and Passengers' Future Behavioural Intentions: Findings from Australian international air passengers. **The Journal of Tourism Studies**, v. 16, n. 1, p. 2–11, 2005.

PYTHON SOFTWARE FOUNDATION. **The Python Language Reference**, 2017. Disponível em: <https://www.python.org/>

SHIVASHANKAR, S. et al. Multi-view methods for protein structure comparison using latent dirichlet allocation. **Bioinformatics**, v. 27, n. 13, p. 61–68, 2011.

SPARKS, B. A.; SO, K. K. F.; BRADLEY, G. L. Responding to negative online reviews: The effects of hotel responses on customer inferences of trust and concern. **Tourism Management**, v. 53, p. 74–85, abr. 2016.

SREENIVASAN, N. D.; LEE, C. S.; GOH, D. H. Tweeting the friendly skies: Investigating information exchange among Twitter users about airlines. **Program**, v. 46, n. 1, p. 21–42, 2012.

SUROWIECKI, J.; SILVERMAN, M. P. The wisdom of crowds. **American Journal of Physics**, v. 75, n. 2, p. 190–192, 2007.

TIRUNILLAI, S.; TELLIS, G. J. Mining Marketing Meaning from Online Chatter: Strategic Brand Analysis of Big Data Using Latent Dirichlet Allocation. **Journal of Marketing Research (JMR)**, v. 51, n. 4, p. 463–479, 2014.

TSAFARAKIS, S.; KOKOTAS, T.; PANTOUVAKIS, A. A multiple criteria approach for airline passenger satisfaction measurement and service quality improvement. **Journal of Air Transport Management**, p. 1–15, 2017.

WAN, Y.; GAO, Q. An Ensemble Sentiment Classification System of Twitter Data for Airline Services Analysis. **Proceedings - 15th IEEE International Conference on Data Mining Workshop, ICDMW 2015**, p. 1318–1325, 2015.

YAO, B. et al. On exploring airline service features from massive online review. **2015 12th International Conference on Service Systems and Service Management, ICSSSM 2015**, 2015.

ZHANG, X. et al. A sequential logistic regression classifier based on mixed effects with applications to longitudinal data. **Computational Statistics & Data Analysis**, v. 94, p. 238–249, 2016.

ZHAO, W. et al. A heuristic approach to determine an appropriate number of topics in topic modeling. **BMC Bioinformatics**, v. 16, n. Suppl 13, p. S8, 2015.

ZHAO, W.; ZOU, W.; CHEN, J. J. Topic modeling for cluster analysis of large biological and medical datasets. **BMC bioinformatics**, v. 15 Suppl 1, n. Suppl 11, p. S11, jan. 2014.

5 CONSIDERAÇÕES FINAIS

Este capítulo apresenta as conclusões da tese, além de sugestões para trabalhos futuros.

5.1 CONCLUSÕES

A presente tese teve por objetivo desenvolver e aplicar técnicas de mineração de textos, de modo a contribuir para a gestão de operações nas áreas médicas e de negócios. O objetivo geral foi atingido em decorrência do encaminhamento de três objetivos específicos.

O primeiro objetivo específico declarado, **identificar e estruturar técnicas de mineração de texto, de modo a elaborar um método para prever internações de pacientes provenientes de emergências hospitalares, tendo como base somente os registros textuais não estruturados escritos por médicos durante o primeiro encontro médico-paciente**, foi encaminhado através do Artigo 1.

O Artigo 1 utilizou um banco de dados que incluía registros textuais do primeiro encontro médico-paciente, antes que os resultados dos exames laboratoriais ou de imagem estivessem disponíveis. Ao todo, 3.024 combinações de ajustes e técnicas de pré-processamento e processamento dos registros textuais foram analisadas. Como resultado, a combinação que utilizou *trigrams*, *term frequency-inverse document frequency, F-value*, percentil de 65 e o algoritmo *Nu-Support Vector Machine – kernel linear* foi identificada como aquela que possui o melhor desempenho, tendo obtido um *F1-score* médio de 77,70% (Desvio Padrão = 0,66%), *Recall* médio de 82,48% (Desvio Padrão = 1,05%), e *Precision* média de 73,47% (Desvio Padrão = 1,14%)

Até onde se tem conhecimento, este foi o primeiro estudo a utilizar a mineração de textos para estimar a demanda por leitos com base nos registros médicos gerados no primeiro encontro médico-paciente em uma EH, mesmo antes que as informações dos exames laboratoriais estivessem disponíveis. Além disso, este é um dos poucos estudos existentes que compararam estratégias de pré-processamento e processamento usando registros médicos de uma EH (MOWERY et al., 2012; ST-MAURICE; KUO, 2012; ZMIRI; SHAHAR; TAIEB-MAIMON, 2012).

Os resultados mostraram os potenciais benefícios da adoção de uma abordagem de mineração de textos em uma EH superlotada, considerando um grande número de pacientes com baixa acuidade e tempo de permanência elevados. O método identificado pode realizar previsões em tempo real sobre a necessidade de leitos e ajudar as equipes de gerenciamento de leitos a melhorar os processos de fluxo do paciente, resultando em uma melhor gestão de operações na área médica.

O segundo objetivo específico declarado, **comparar previsões realizadas pelo método proposto no Artigo 1 com análises médicas realizadas por humanos, de modo a verificar se computadores podem atuar de forma autônoma na tarefa de previsão de internações de pacientes provenientes de emergências hospitalares**, foi encaminhado através do Artigo 2.

O mesmo banco de dados foi utilizado nos Artigos 1 e 2. Entretanto, o Artigo 2 utilizou um subconjunto de testes, para comparar o desempenho de homens e máquinas, composto por 230 registros selecionados aleatoriamente. A representação humana da comparação foi formada por quatro médicos, sendo dois considerados experientes e dois considerados novatos (com mais de dez anos de experiência e até dois anos de experiência em emergências hospitalares, respectivamente). Todos os 230 registros foram avaliados por médicos experientes e novatos. Ao final, as previsões foram reunidas e comparadas com aquelas indicadas pelo método de melhor desempenho, conforme análise do Artigo 1.

Até onde se tem conhecimento, este é o primeiro estudo que compara um método de mineração de textos com médicos que atuam em uma EH na previsão de admissões de pacientes, baseado somente nos registros médicos gerados no primeiro encontro médico-paciente. Os principais resultados mostram que tanto médicos (iniciantes ou experientes), quanto máquina, tiveram desempenhos semelhantes. Entretanto, o algoritmo foi significativamente mais rápido quando comparado aos médicos.

A proporção de casos em que os médicos estavam corretos enquanto a máquina estava errada, e na qual a máquina estava correta enquanto os médicos estavam errados, foi semelhante (8,26% e 6,09%, respectivamente). Entretanto, é difícil para humanos explicarem o porquê das diferenças. Sabe-se que a compreensão humana das decisões de inteligência artificial (IA) é limitada, tendo sido proposto que seria possível treinar a própria IA para fornecer justificativas em linguagem natural de suas decisões (PARK et al., 2016).

Por fim, cabe ressaltar que os médicos analisaram os registros fora do horário de trabalho, em um ambiente calmo, sem realizar multitarefas ou com restrições de tempo. É difícil prever como os desempenhos poderiam ser afetados se as avaliações fossem realizadas em paralelo com suas atividades de trabalho de emergência, mas é razoável especular que elas seriam piores. Acredita-se que há potenciais usos de sistemas inteligentes para auxiliar a gestão hospitalar, e é possível que esses sistemas possam ser implementados com mais facilidade do que aqueles focados em decisões clínicas. O método avaliado no Artigo 2 é um desses exemplos; sua implementação poderia fornecer dados úteis para as equipes de gerenciamento de leitos, melhorando os processos de fluxo de pacientes em todo o hospital.

O terceiro objetivo específico declarado, **identificar e estruturar técnicas de mineração de texto, de modo a elaborar um método para prever a satisfação de clientes de companhias aéreas, tendo como base as avaliações escritas e publicadas por passageiros na internet**, foi encaminhado através do Artigo 3.

O Artigo 3 utilizou um banco de dados extraído da página *Air Travel Review* (ATR), reunindo as opiniões de mais de 55.000 passageiros, de mais 400 companhias aéreas, originários de mais de 170 países. Comparado com estudos anteriores, o banco de dados utilizado é muito maior, o que possibilitou generalizações mais confiáveis (LIAU; TAN, 2014; MOSTAFA, 2013; SREENIVASAN; LEE; GOH, 2012; WAN; GAO, 2015; YAO et al., 2015).

De modo a identificar as dimensões de satisfação presentes nas avaliações online, foi utilizado o método *Latent Dirichlet Allocation* (LDA). Como medida de ajuste do modelo, foi utilizada a medida de perplexidade. O número ideal de tópicos no modelo LDA foi identificado como sendo 67, apresentando uma perplexidade média de 74,12 (Intervalo de Confiança de 95%: 72,46 - 75,77). Entretanto, após o agrupamento de tópicos que apresentaram conteúdo e significado semelhantes, além do descarte de três tópicos considerados não relevantes (aqueles que apresentaram probabilidade de 0,05% ou menos na distribuição de tópicos), o número final de 27 dimensões de satisfação foi identificado. Até onde se tem conhecimento, este foi o primeiro uso de LDA com o intuito de identificar as dimensões de satisfação de clientes no setor aéreo. O método proposto possibilita análises de forma segmentada, conforme características dos passageiros e do serviço prestado. Além disso, também é sensível para mostrar as tendências dos clientes ao longo do tempo.

Em relação à análise dos sentimentos das avaliações *online*, um classificador Naïve Bayes foi ajustado. A acurácia do ajuste de treino foi de 93,68%. Somente os adjetivos presentes em pelo menos 0,1% das avaliações foram mantidos, resultando em 882 adjetivos analisados. Em estudos anteriores (CHOW, 2015), o contexto de uso de adjetivos não foi avaliado. Nesse sentido, o método proposto representa um avanço na análise de avaliações online do segmento de transporte aéreo de passageiros, visto que os sentimentos podem ser identificados por dimensão, considerando diferentes perfis de passageiros, serviços e períodos de tempo. A observação da polaridade desses sentimentos oferece informações para manter atividades que são positivamente reconhecidas pelos passageiros e também para ajustar serviços relacionados a dimensões que estão recebendo avaliações negativas.

Complementarmente, as dimensões de satisfação e adjetivos identificados foram testados e validados usando um Classificador de Regressão Logística na tarefa de previsão de recomendação das companhias aéreas. A acurácia média do conjunto de testes foi de 79,95% (Intervalo de Confiança de 95%: 79,93% - 79,97%). As avaliações positivas e recomendações de clientes são comumente associadas à satisfação de clientes (FORGAS et al., 2010; GUO; BARNES; JIA, 2017; MATTILA, 2004; MORGAN; HUNT, 1994). As três principais dimensões de satisfação necessárias para prever a recomendação das companhias aéreas são: “tripulação”, “serviço de bordo” e a “relação custo-benefício”. Projetar serviços que atendam ou superem as expectativas dentro dessas dimensões é fundamental para a recomendação (CHOW, 2015).

Por fim, esta pesquisa representou um passo adiante na medição da satisfação de clientes de companhias aéreas, sendo importante para a gestão de operações nesta área de negócios. O método proposto permitiu extrair dimensões de satisfação, além dos sentimentos associados a elas, considerando diferentes perfis de passageiros, serviços e períodos de tempo.

5.2 SUGESTÕES PARA TRABALHOS FUTUROS

Pesquisas futuras podem ser desenvolvidas como extensões dos desenvolvimentos aqui propostos. São elas:

a) Analisar outros aspectos da gestão de operações hospitalares que podem ser beneficiados com o uso de mineração de textos (ex.: uso de recursos/capacidades a partir de análise de prontuários, exames, dentre outros).

b) Analisar o uso combinado de técnicas de mineração de texto e mineração de dados na área médica e de negócios.

c) Aprimorar o método proposto no Artigo 1 e validado no Artigo 2, de modo que ele forneça explicações em linguagem natural sobre as suas previsões.

d) Comparar resultados obtidos por mineração de textos e métodos convencionais na análise de dimensões de satisfação de clientes de companhias aéreas.

e) Verificar o desempenho de outros métodos na identificação de dimensões de satisfação e na análise de sentimentos, a partir de avaliações online.

f) Validar o método proposto no Artigo 3 com o uso de outras métricas, além da previsão de recomendação da companhia aérea.

5.3 REFERÊNCIAS

CHOW, C. K. W. On-time performance, passenger expectations and satisfaction in the Chinese airline industry. **Journal of Air Transport Management**, v. 47, p. 39–47, 2015.

FORGAS, S. et al. Antecedents of airline passenger loyalty: Low-cost versus traditional airlines. **Journal of Air Transport Management**, v. 16, n. 4, p. 229–233, 2010.

GUO, Y.; BARNES, S. J.; JIA, Q. Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. **Tourism Management**, v. 59, p. 467–483, 2017.

LIAU, B. Y.; TAN, P. P. Gaining customer knowledge in low cost airlines through text mining. **Industrial Management & Data Systems**, v. 114, n. 9, p. 1344–1359, 2014.

MATTILA, A. S. The impact of service failures on customer loyalty. **International Journal of Service Industry Management**, v. 15, n. 2, p. 134–149, 2004.

MORGAN, R. M.; HUNT, S. D. The commitmemt-trust theory of relationship marketing. **Journal of Marketing**, v. 58, n. 3, p. 20–38, 1994.

MOSTAFA, M. M. An emotional polarity analysis of consumers' airline service tweets. **Social Network Analysis and Mining**, v. 3, n. 3, p. 635–649, abr. 2013.

MOWERY, D. et al. Building an automated SOAP classifier for emergency department reports. **Journal of biomedical informatics**, v. 45, n. 1, p. 71–81, fev. 2012.

SREENIVASAN, N. D.; LEE, C. S.; GOH, D. H. Tweeting the friendly skies: Investigating information exchange among Twitter users about airlines. **Program**, v. 46, n. 1, p. 21–42, 2012.

ST-MAURICE, J.; KUO, M. H. Analyzing Primary Care Data to Characterize Inappropriate Emergency Room Use. p. 990–994, 2012.

WAN, Y.; GAO, Q. An Ensemble Sentiment Classification System of Twitter Data for Airline Services Analysis. **Proceedings - 15th IEEE International Conference on Data Mining Workshop, ICDMW 2015**, p. 1318–1325, 2015.

YAO, B. et al. On exploring airline service features from massive online review. **2015 12th International Conference on Service Systems and Service Management, ICSSSM 2015**, 2015.

ZMIRI, D.; SHAHAR, Y.; TAIEB-MAIMON, M. Classification of patients by severity grades during triage in the emergency department using data mining methods. **Journal of evaluation in clinical practice**, v. 18, n. 2, p. 378–88, abr. 2012.