UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

DETECÇÃO DE ESTADO DE ÂNIMO EM FÓRUNS EDUCACIONAIS UTILIZANDO MINERAÇÃO DE TEXTO

DIÓGENES LUIZ OLIVEIRA DE AZEVEDO

RECIFE
2016

UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

DIÓGENES LUIZ OLIVEIRA DE AZEVEDO

DETECÇÃO DE ESTADO DE ÂNIMO EM FÓRUNS EDUCACIONAIS UTILIZANDO MINERAÇÃO DE TEXTO

Monografia apresentada ao Curso de Bacharelado em Ciência da Computação da Universidade Federal Rural de Pernambuco como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. Rafael Ferreira Leite de Mello

MINISTÉRIO DA EDUCAÇÃO E DO DESPORTO
UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO (UFRPE)
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

http://www.bcc.ufrpe.br

FICHA DE APROVAÇÃO DO TRABALHO DE CONCLUSÃO DE CURSO

Trabalho defendido por Diógenes Luiz Oliveira de Azevedo como requisito para conclusão do curso de Bacharelado em Ciência da Computação da Universidade Federal Rural de Pernambuco, intitulado *Detecção de estado de ânimo em fóruns educacionais utilizando mineração de texto*, orientado por Rafael Ferreira Leite de Mello e aprovado pela seguinte banca examinadora:

Rafael Ferreira Leite de Mello
DEINFO/UFRPE

Pablo Azevedo Sampaio
DEINFO/UFRPE

Valmir Macário Filho
DEINFO/UFRPE

AGRADECIMENTOS

Agradeço primeiramente a toda a minha família, em especial a minha mãe e irmã, Cássia e Fernanda, por todos os anos da graduação em que me apoiaram independentemente dos sucessos obtidos ou não.

Ao meu, orientador Rafael Ferreira, pela paciência e conhecimento a mim repassado durante todo o tempo do desenvolvimento deste projeto, iniciação científica e disciplinas que estivemos juntos.

A todos os professores do Departamento de Estatística e Informática da Universidade Federal de Pernambuco, por terem colaborado na minha formação acadêmica.

Aos meus colegas, pelo companheirismo durante toda a minha graduação, que contribuiu diretamente no sucesso do meu percurso no curso.

Aos meus amigos fora da Universidade que, de alguma forma, assim como eu, também comemoram muito esta etapa vencida.

RESUMO

Com a disseminação da Internet e dos cursos de educação à distância, os dados gerados por professores, alunos e tutores vêm crescendo exponencialmente nos últimos anos, em especial os dados textuais, oriundos dos fóruns educacionais dos ambientes virtuais de aprendizagem. Contudo, mesmo com o sucesso dessa modalidade de ensino, a evasão vem se mostrando um grande problema a ser combatido. A identificação de estudante com possibilidade de desistência é uma das formas encontradas para prevenção da evasão. Técnicas computacionais, como mineração de texto, podem ser usadas para lidar com esse problema. Diante deste cenário, este trabalho propõe uma ferramenta que auxilie o professor na detecção do estado de ânimo do aluno, a partir de postagens de fóruns educacionais, prevendo um possível caso de evasão. Os experimentos propostos mostram que a proposta alcançou a uma taxa de acerto de 73,57%.

Palavras-chave: Mineração de texto, análise de sentimento, educação à distância.

ABSTRACT

With the spread of the Internet and distance learning courses, the data generated by teachers and students have grown exponentially in recent years, especially textual data from the educational forums from virtual learning environments. However, even with the success of this method of education, the dropouts students have proven a major problem to be addressed. The identification of a possible unmotivated student is useful to prevent a dropout. Computational techniques such as text mining could be used to deal with this problem. This paper proposes a tool that helps the teacher to detect the student motivation based on theirs posts in educational forums. The proposed experiments show that the proposal reach 73,57% of acuracy.

Keywords: Text mining, sentiment analysis, distance learning.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS, SIGLAS E SÍMBOLOS

AVA	Ambiente Virtual de Aprendizagem
EaD	Educação à Distância
REA	Roda dos Estados Afetivos
UFRPE	Universidade Federal Rural de Pernambuco
TF	*Term Frequency*
IDF	*Inverse Document Frequency*

SUMÁRIO

1. INTRODUÇÃO

Os primeiros registros da prática da educação à distância datam de 1728, quando alunos de um curso em Boston, nos Estados Unidos, usavam correspondências para desenvolver conhecimento. Nos dias de hoje, as interações que ocorriam em 1728 se fazem em tempo real, através de ferramentas pertencentes a ambientes educacionais disponíveis na Internet, também conhecidos como Ambiente Virtual de Aprendizagem (AVA) (DAMASCENO, 2015).

Dentre as ferramentas disponíveis nos AVAs o fórum educacional é a que proporciona maior interatividade entre alunos, professores e tutores, pois ele oportuniza o esclarecimento, consultas, e a socialização dos participantes (BASTOS, 2011). Porém, mesmo com o crescimento da modalidade do ensino à distância e da interação que o fórum proporciona, a evasão tem sido um dos maiores problemas identificados nesta área (DE ALMEIRA BIZARRIA, 2015).

Barroso e Falcão (2004) afirmam que existem pelo menos três variáveis explicativas para o fenômeno da evasão, são elas:

i) Econômica – quando o aluno não tem condições socioeconômicas de prosseguir no curso;
ii) Vocacional – quando não há identificação do aluno com o curso;
iii) Institucional – quando a evasão se dá por reprovações nas disciplinas iniciais ou inadequação aos métodos de estudo, por exemplo.

Uma das correntes de pesquisa que estudam métodos para prever a eminência de uma evasão é a mineração de texto em fóruns educacionais (DE SOUZA AFIUNE, 2012).

A grande quantidade de informação gerada nos fóruns dificulta a tarefa do professor no acompanhamento individual do aluno (DE SOUZA AFIUNE, 2012) que, por sua vez, por meio de uma postagem no fórum de um AVA, pode demonstrar seu estado de ânimo (LONGHI, 2009), servindo como alerta para identificação de uma possível evasão.

Diante do cenário descrito, a hipótese deste trabalho baseia-se na ideia de que aplicando técnicas de mineração de texto, pode-se identificar o estado de ânimo de uma postagem no fórum educacional para evidenciar um aluno desmotivado, auxiliando o professor na descoberta de alunos propensos a abandonar o curso.

Para isso, será necessário o uso da abordagem de mineração de texto, mais especificamente a área de estudo de análise de sentimento (DEVI e RASHEED, 2015). Esta técnica será utilizada como meio para a criação de um sistema onde, a partir das postagens dos alunos como entrada, o estado de entusiasmo do aluno (motivado ou desmotivado) com a disciplina será obtido como saída. Desta maneira, a ferramenta servirá como um auxílio ao professor na tentativa de evitar que o aluno se desmotive ou abandone o curso.

1.1 JUSTIFICATIVA

Segundo os estudos de Maia (2003), em 2003 a educação à distância no Brasil já gerava uma receita por volta dos R$350 milhões, com 34 cursos criados (PEREIRA, 2007). Com a disseminação da Internet e dos próprios ambientes digitais de aprendizagem, em 2014 o Brasil já tinha mais de 900 cursos ofertados à distância (ALVES, 2015). Atualmente o AVA tem uma grande importância para o EaD (Educação à Distância) e ferramentas como fórum, Wiki e blogs são bastante utilizados. Dentre estas, o fórum tem um papel muito importante pelo fato de ser a plataforma onde os usuários podem se socializar, aproximando-os de uma experiência de ensino presencial (BASTOS, 2011).

O fórum educacional é uma ferramenta capaz de gerar uma grande quantidade de informação (BASTOS, 2011), pois é um ambiente onde todos os agentes do ensino à distância interagem e se socializam. Um cálculo simples mostra que, se um professor tem uma turma de 30 alunos e cada um deles posta algo no fórum cinco vezes por semana, ao fim do mês o professor terá 600 postagens para ler. A partir da maneira de como o fórum funciona, agindo como um gerador de informação e os intensos investimentos em EaD, ocasionando em mais fóruns em atividade, tem-se como resultado a grande quantidade de informação gerada, em

especial, a informação textual. Uma das maneiras de lidar com tamanha quantidade de informação é utilizar a mineração de texto.

Pang e Lee (2008) afirmam que uma parte importante da mineração de texto é descobrir o que as pessoas estão pensando e sentindo, e que existem vários recursos ricos em opiniões a serem mineradas. O que os autores do texto sentem ou pensam tem sido alvo de grande disputa entre empresas no mundo todo, na busca em saber como seus clientes enxergam o produto ofertado.

No âmbito educacional, o aumento do número de cursos à distância tem gerado uma grande quantidade de informação textual por meio dos fóruns educacionais, o que dificulta o acompanhamento individual do aluno por meio do professor (DE SOUZA AFIUNE, 2012). Sendo assim, a mineração de texto pode ter uma tarefa importante agindo no problema da evasão, pois um dos meios de se combater a evasão é impedir que ela ocorra e, a partir de interações em fóruns educacionais, a análise de sentimento pode evidenciar o estado de ânimo do aluno (AZEVEDO, 2010), identificando um possível caso de desistência do curso.

Diante disto, este trabalho propõe a implementação de uma ferramenta que evidencie o grau de entusiasmo do aluno a partir da sua interação no fórum educacional e auxilie o professor para que o mesmo busque alternativas que motive novamente o aluno, na tentativa de evitar sua evasão.

1.2 OBJETIVOS

Esta seção contém os objetivos gerais e específicos que o trabalho visa atingir.

1.2.1 Geral

Este trabalho tem como objetivo, auxiliar professores e tutores de EaD a identificar alunos desmotivados a partir de postagens em fóruns educacionais.

1.2.2 Específicos

Dos objetivos específicos deste trabalho, podemos destacar os seguintes pontos:

- Implementar uma ferramenta que a partir das interações no fórum educacional, evidencie alunos desmotivados a partir de técnicas de mineração de texto.
- Contribuir com os estudos de análise de sentimento na língua portuguesa, apresentando uma alternativa para o problema da negação.

1.3 DEFINIÇÃO DA METODOLOGIA

A partir da revisão bibliográfica e trabalhos relacionados, adquiriu-se conhecimento suficiente para planejar e executar o roteiro necessário para confeccionar o método de pesquisa a ser adotado neste projeto, podendo ser resumido nos seguintes passos:

1) Definir o arranjo experimental com base nos parâmetros (linguagem de programação) e variáveis (manipulação dos dados);
2) Implementar a ferramenta proposta;
3) Realizar experimentos e analisar os resultados obtidos.

No passo 1, como parâmetros (algo que permanecerá fixo durante todo o processo de implementação), o Python foi definido como linguagem de programação a ser usada na implementação da ferramenta proposta por questões de experiência, simplicidade e afinidade com o ambiente. Além disso, a base de dados utilizada também é um parâmetro do arranjo experimental. Sendo assim, a decisão de usar o *10-fold cross-validation* na etapa de avaliação é identificada como variável, pois outros métodos (como divisão em treinamento e teste) não têm uma saída tão fiel na classificação.

Uma lista de *stopwords* (palavras que acrescentem pouco ou nenhum significado ao texto) é retirada na etapa de pré-processamento dos dados. Esta lista também é considerada uma variável, pois no decorrer dos experimentos, existem palavras que negam uma afirmação (não, nem...), alterando o sentido da postagem, logo, palavras neste contexto deixam ser consideradas *stopwords* para serem tratadas e evitar equívocos no classificador.

O passo 2 está baseado na implementação e utilização do algoritmo *tf-idf* (*term-frequency /inverse document frequency*). O principal objetivo deste algoritmo é identificar termos muito frequentes em mais de uma área (no caso deste trabalho, palavras que aparecem frequentemente nas postagens tanto dos alunos motivados como dos desmotivados), tornando-os de pouco valor discriminatório (AVILA, 2006). Cada palavra tem um valor *tf-idf* motivado e desmotivado associado e assim, uma postagem é vista como um conjunto de palavras e tem um valor resultante do somatório doa valores *tf-idf* desses termos.

Por fim, no último passo, vários testes foram realizados até o trabalho atingir a acurácia atual. Vários métodos de avaliação foram testados usando a validação cruzada. Na etapa de distribuição dos pesos o valor do incremento foi alterado até perceber-se que o dobrar seria o incremento ideal. Como cada palavra tem dois pesos (motivado e desmotivado), uma postagem é o somatório dos pesos das palavras que a formam. Anteriormente cada postagem tinha dois somatórios, um para motivado e outro para desmotivado. O maior somatório servia para polarizar a postagem. Após os testes, percebeu-se que atribuir pesos negativos às palavras desmotivadas e fazer apenas um somatório (caso seja maior que zero, polariza-se como motivado e caso contrário, como desmotivado) melhorou a taxa de acertos do sistema.

1.4 ESTRUTURA DO TRABALHO

Neste trabalho, a introdução, objetivos, justificativas do tema e metodologia adotada foram apresentados no primeiro capítulo. O restante do trabalho divide-se na seguinte maneira: No segundo capítulo, são explanados assuntos reconhecidos como pré-requisito para o total entendimento do trabalho. Estes assuntos contemplam a fundamentação teórica. No terceiro capítulo são apresentados os trabalhos relacionados ao tema escolhido. O quarto capítulo detalha a implementação da ferramenta proposta. No quinto capítulo são apresentados os testes realizados e resultados obtidos, avaliando a acurácia do sistema desenvolvido no trabalho. O sexto capítulo discorre acerca das conclusões e trabalhos futuros.

2. FUNDAMENTAÇÃO TEÓRICA

Este capítulo aborda os temas necessários para a melhor compreensão dos tópicos abordados neste trabalho.

2.1 MINERARÇÃO DE TEXTO

Mineração de texto é uma especialidade da mineração de dados, onde o objeto de estudo são dados textuais (Becker e Tumitan, 2013). A principal tarefa da mineração de texto é fazer com que computadores consigam extrair conhecimento de um documento sem a intervenção de um humano.

Sendo assim, podem-se extrair várias características de um texto em estudo, como o assunto ou tema, relações entre as palavras, agrupamento de textos e até a opinião e sentimento do autor do documento.

2.2 ANÁLISE DE SENTIMENTO

A análise de sentimento é uma subárea da mineração de texto que visa detectar a opinião do autor acerca de algum tema. Abaixo, discorre-se sobre as definições e funcionamento de uma análise de sentimento.

2.2.1 Definições

Mineração de texto é a modalidade da mineração de dados que busca extrair conhecimento, em especial, do texto estudado e vem crescendo principalmente por ser utilizada como mecanismo de descoberta de informação de clientes por grandes empresas no mundo (BECKER e TUMITAN, 2013). Um dos maiores exemplos são a gigante de buscas *Google* e a atual maior rede social do planeta, o *Facebook*, que procura prever até quando seus usuários vão mudar o status do relacionamento.

Uma das áreas da mineração de texto é a análise de sentimento e nessa modalidade em si, busca-se polarizar o sentimento do autor em duas classes, geralmente uma positiva e outra negativa. No caso deste trabalho, as classes buscadas são motivadas e desmotivadas. Sendo assim, resumidamente, pode-se estruturar a problemática da análise de sentimento em três fases: a) identificar a opinião do autor referente a algum assunto; b) polarizar a opinião, isto é, definir a orientação da opinião do autor do texto (geralmente em positiva ou negativa); e c) apresentar os resultados de forma sumarizada (TSYTSARAU e PALPANAS, 2012).

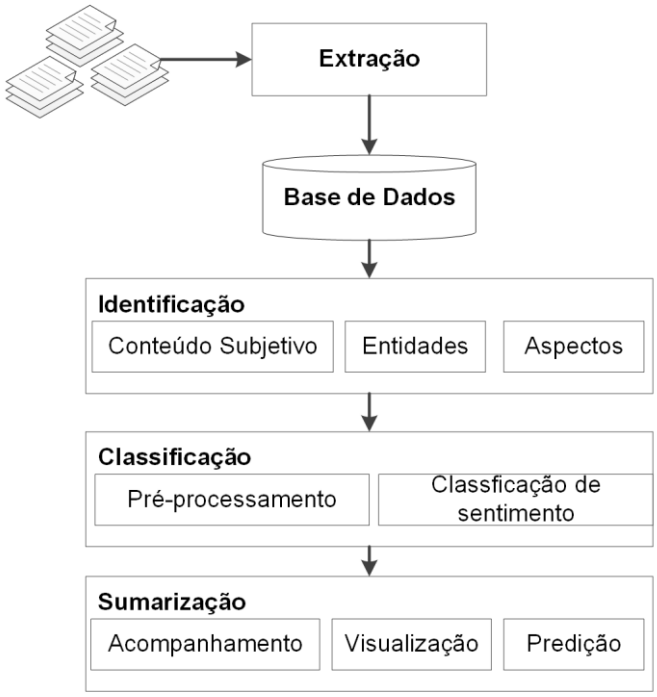
Toda a análise de sentimento é operada basicamente sobre duas entidades: um alvo, que é o assunto em si, podendo-se referir a alguma pessoa, acontecimento, produto, evento, etc. A outra entidade é chamada de sentimento, que representa a opinião propriamente dita do autor para com o alvo (LIU, 2012).

Becker e Tumitan (2013) assumem que existem alguns tipos de opiniões, divididas em regulares, quando o autor de fato expressa seu sentimento sobre o alvo (“Este time joga muito bem”) ou comparativas, quando o autor busca expressar sua opinião com base na relação de similaridade entre um ou mais alvos (“As aulas nesta faculdade são bem melhores que as daquela outra”). Opiniões também podem ser diretas ou indiretas. (“Este carro é muito bom”) é um exemplo de uma opinião direta e (“Depois que troquei os pneus o carro ficou melhor para dirigir”) é uma opinião indireta pois a melhora da dirigibilidade do carro está altamente ligada ao fato da troca dos pneus. E por fim, opiniões ainda podem ser divididas em explícitas, quando o autor demonstra seu sentimento explicitamente (“Estou triste com o término do namoro”) ou implícitas, quando o sentimento para com o alvo não está explícito (“Comprei um celular bem caro que descarrega muito rápido”). Opiniões regulares, diretas e explícitas são mais utilizadas nos estudos de análise de sentimento por serem mais fáceis de se polarizar.

2.2.2 Etapas da Análise de Sentimento

Como já abordado anteriormente, a análise de sentimento pode ser dividida em identificação, classificação e sumarização de resultados.

Figura 1 – Etapas da análise de sentimento.



Fonte: (BECKER e TUMITAN, 2013).

2.2.2.1 Identificação

Nesta etapa, o principal objetivo é identificar o alvo da opinião a ser tratado (AGGARWAL, 2012). Tornou-se corriqueira a busca das empresas em saber a opinião de seus clientes perante sua marca ou seus produtos, mas dependendo de onde forem extraídos os documentos, os alvos podem ser outros (SARAWAGI, 2008). Por exemplo, numa revista é comum o alvo ser uma celebridade ou atleta.

Alguns problemas podem atrapalhar a etapa de identificação. Na frase “O presidente assinou hoje a lei proposta na câmara dos deputados da última semana. Ele deve fazer um pronunciamento ainda hoje acerca disto” é importante saber a data do acontecimento para detectar o presidente em questão. Em “O rubro-negro venceu mais uma partida no fim de semana” o rubro-negro em questão pode ser designado a vários times do futebol brasileiro dependendo de onde o texto foi escrito.

2.2.2.2 Classificação

Em mineração de opinião, a classificação ou polarização pode ser tratado como um problema de saída binária, onde o sentimento geralmente é positivo ou negativo. Uma das tarefas mais importantes nesta etapa é a de pré-processamento dos dados extraídos na fase de identificação, pois nesta etapa são retiradas todas as palavras que não atribuem conhecimento ao texto.

Porém, ainda existem barreiras que na língua portuguesa atrapalham muito a etapa de classificação. Basicamente, detectar ironia e negação são fortes problemas que acabam por confundir o classificador e gerando uma classificação errada. Na frase “O Brasil conseguiu bater seu recorde de desemprego. Nossos políticos merecem os parabéns” há claramente o uso de ironia pelo autor do texto e detectar o real sentimento deste alvo torna-se bastante difícil para o computador. “Este livro passa longe de ser perfeito como dizem” é um exemplo de negação que pode confundir o classificador se não for tratado corretamente. Neste caso, um tratamento mais complexo da frase por meio das análises sintáticas e semânticas pode amenizar o problema.

2.2.2.3 Sumarização

Por fim, a sumarização é apresentação dos resultados da classificação e deve ser feita usando uma base de dados maior possível, não bastando apenas documentos extraídos de um grupo de pessoas (LIU, 2012).

Esta etapa é importante pois sumariza as informações que compradores, por exemplo, acham de um produto, ajudando outros interessados a se decidirem se querem comprar ou não. É bastante comum sites de varejo sumarizarem as informações de seus clientes a cerca de um smartphone, por exemplo, mostrando seus comentários sobre vários aspectos do produto, como duração da bateria, peso e design.

É importante ressaltar que o sentimento sumarizado de grupos de pessoas pode servir também como predição de vencedores de uma eleição, comportamento da bolsa de valores (BOLLEN, 2011), equipes favoritas em casas de apostas

esportivas e até para definir preços de produtos, como sistemas de regressão (ARCHAK, 2007).

2.3 O PROBLEMA DA NEGAÇÃO EM MINERAÇÃO DE TEXTO

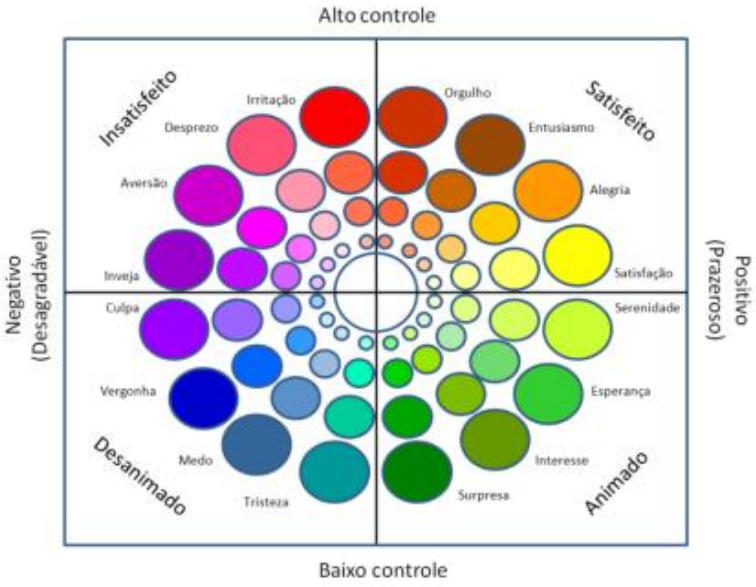
Negar uma sentença qualquer quer dizer inverter o sentido da afirmação. Na frase “Eu não tenho interesse em continuar estudando isso”, o uso do advérbio “não” indica a negação da frase “Eu tenho interesse em continuar estudando isso”. Este é um exemplo simples, mas a negação pode ser feita com outros advérbios ou interjeições. Além disso, pode haver mais de um “não” numa frase o que pode (ou não) continuar invertendo da frase e, no pior dos casos, a negação pode vir disfarçada de ironia (na frase “Minhas notas foram tão boas que reprovei o semestre”, as notas do autor da frase não foram boas, ele usou o artifício da ironia para negar sua sentença).

Em mineração de texto na língua portuguesa, a identificação da negação tem sido um dos grandes problemas na área (BECKER e TUMITAN, 2013) pela pluralidade de formas com que a negação pode se mostrar, devido a riqueza e complexidade do nosso idioma.

2.4 ANÁLISE DE SENTIMENTO EM AMBIENTES EDUCACIONAIS

Na esfera educacional, a análise de sentimento vem começando a ser utilizada no Brasil na descoberta do sentimento dos alunos. Trabalhos como os de Longhi (2009) utilizam a chamada REA (Roda dos Estados Afetivos) definida por Tran (2004) visando a detecção do sentimento dos autores de textos no âmbito da educação.

Figura 2 – Espaço de representação dos estados de ânimo do aluno



Fonte: (TRAN, 2004).

Longhi (2009) afirma que usando a REA, pode-se estabelecer as classes utilizadas na polarização dos alvos. Basicamente, a REA é dividida em quatro quadrantes que representam grupos de sentimentos que identificam um aluno como animado, desanimado, satisfeito e insatisfeito.

• Animado – Indica a demonstração ou não de outros sentimentos ou atitudes presentes em seu quadrante (surpresa, interesse, esperança e serenidade). Representam o aluno que conseguiu alcançar o sucesso e estão vislumbrados com o feito.

• Desanimado – É o estado de ânimo que demonstra ou não os sentimentos de tristeza, medo, vergonha e culpa. Antagonicamente ao quadrante animado, os alunos representados por este quadrante são aqueles indivíduos que pelo insucesso no aprendizado, alimentam o medo e o receio em tentar reverter a situação.

• Satisfeito – Indica a demonstração ou não de orgulho, entusiasmo, alegria e satisfação. Os alunos deste quadrante estão felizes com o que conquistaram até o momento e buscam mais conquistas por meio do entusiasmo.

• Insatisfeito – Indica a presença ou não de sentimentos como irritação, desprezo aversão e inveja. Representam o aluno que não estão contentes com o desempenho até o momento e podem se mobilizar para reverter o quadro.

Assim, os quadrantes da direita (satisfeito e animado) representam a polaridade positiva e os da esquerda (insatisfeito e desanimado) representam a polaridade negativa, identificando a classificação do alvo.

Desta maneira, a análise de sentimento sumarizada na REA mostra-se como uma das alternativas para se detectar um aluno desmotivado, propício à evasão, assim como também identifica alunos motivados dentro do processo de aprendizagem.

2.5 TF-IDF

O *tf-idf*, do inglês *term-frequency/inverse document frequency*, é um algoritmo baseado na função logarítmica onde palavras que aparecem com muita frequência no documento têm seu valor discriminado.

O *tf* é a probabilidade de uma palavra aparecer no documento, calculada por meio da razão entre a frequência que a palavra aparece e o número total de palavras no documento, fórmula 1.

$$tf_{t,d} = \frac{n_t}{\sum_d n_d} \tag{1}$$

Já o *idf*, a frequência inversa, é obtida pelo logaritmo da razão entre o número total de documentos (N) e a quantidade de documentos contendo a palavra t (Nt), fórmula 2. O *tf-idf* é, então, a multiplicação desses dois termos.

$$idf_t = \log \frac{N}{N_t} \tag{2}$$

2.4 PYTHON

Python é uma linguagem de programação lançada por Guido van Rossum sem fins lucrativos no início da dos anos noventa e se tornou bastante conhecida pelo alto nível de abstração e pela abordagem multiparadigma. A linguagem também se destaca por apresentar uma sintaxe clara, por apresentar códigos curtos em comparação com outras linguagens e pelo vasto cartel de bibliotecas nativas, sendo ideal para processamento de textos e dados.

O paradigma do Python escolhido para codificação da ferramenta proposta foi a programação funcional, esta abordagem consiste em tratar a tarefa de programar como definições de funções matemáticas, fazendo com que os dados usados não sejam mutáveis, o que faz desse paradigma uma ótima escolha para a aplicação, pois as alterações feitas nos dados ao longo do código só surtem efeito localmente, apenas na função onde foram chamados, sem alterá-los globalmente quando forem posteriormente utilizados em outras funções.

3. TRABALHOS RELACIONADOS

Os fóruns educacionais são estudados sob diferentes perspectivas, esta seção apresenta aplicações de técnicas de mineração de texto focadas em prevenir a evasão e utilização de análise de sentimento.

O trabalho proposto em (LONGHI, 2009) detalha uma metodologia para realizar análise de sentimento em postagens de fóruns. Ela classifica a postagem em insatisfeito, desanimado, satisfeito e animado. Apesar do trabalho ser promissor, pouco foi detalhada a solução proposta e os experimentos utilizaram apenas 5 postagens.

A proposta de utilização da teoria da semântica lexical computacional é apresentada em (RIGO, 2013). Este trabalho utiliza o SentiLex-PT3, que é um léxico semântico em português, para extrair os sentimentos das postagens. Esse recurso contém 6321 lemas adjetivos e 25406 formas flexionadas que podem ser usadas para determinar a polaridade de um texto.

Em (REIS, 2010) é proposta uma classificação diferente dos trabalhos anteriores. Porém, o trabalho apresenta apenas um estudo, não foi proposta nenhuma ferramenta.

É importante ressaltar que a análise de sentimento pode ser usada para lidar com o problema da evasão. Contudo, os autores dos trabalhos citados não deixam explicita essa relação. Silva (2015) propõe a utilização do quantitativo de postagens em fóruns educacionais para prever a evasão. Apesar de o trabalho ter alcançado um resultado interessante, a aplicação de mineração de texto pode melhorar consideravelmente o sistema proposto.

Diante disto, este trabalho propõe um sistema de mineração de texto para identificar o ânimo do aluno fazendo o paralelo com o problema da evasão. As principais novidades do trabalho são:

- Utilização de uma abordagem que utiliza técnicas estatísticas com a aplicação um léxico semântico.
- Tratamento de palavras emocionais e de negação
- Utilização de um corpus de postagens maior para avaliação;

- Propõe o paralelo entre análise de sentimento e evasão escolar.

Tabela 1 – Comparação com trabalhos relacionados

Trabalho	Mineração de texto	Educacional	Foco na evasão	Tratamento de negação	Implementação/ uso de ferramenta
(LONGHI, 2009)	Sim	Sim	Não	Não	Sim
(RIGO, 2013)	Sim	Sim	Não	Não	Sim
(REIS, 2010)	Sim	Sim	Não	Não	Não
(SILVA, 2015)	Não	Sim	Sim	Não	Sim
O autor	Sim	Sim	Sim	Sim	Sim

Fonte: O autor.

4. SISTEMA PARA IDENTIFICAÇÃO DE ÂNIMO

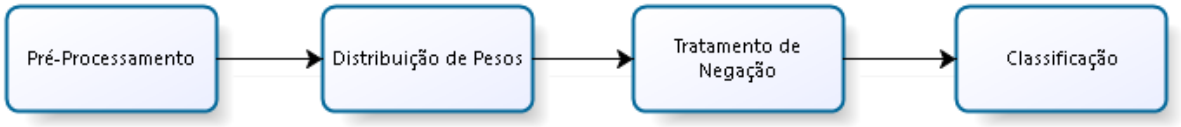
Esta seção aborda os algoritmos utilizados e a metodologia utilizada na classificação das postagens.

O sistema é dividido em quatro etapas: pré-processamento, distribuição de pesos, tratamento de negação e classificação, Figura 3. As próximas subseções tratam cada uma dessas etapas de modo que fique claro como foram codificadas e como funcionam, bem como as configurações de saída de cada etapa, que serve de entrada para as seguintes.

Para melhor ilustrar o funcionamento das etapas, a frase abaixo será manipulada em cada uma das fases, servindo como exemplo da aplicação.

Frase: *as minhas notas estão baixas, assim eu não irei aprovar na disciplina...*

Figura 3 – Sequência de passos de codificação da ferramenta



Fonte: O autor.

4.1 PRÉ-PROCESSAMENTO

Na etapa de pré-processamento foram executados dois serviços. São eles:

- Tokenização: uma prática em mineração de texto que consiste em decompor o documento em palavras usando delimitadores. Estes delimitadores são espaços em branco, quebra de linha, pontuação e caracteres especiais.
- Eliminação de *stopwords*: prática que consiste em remover palavras que agregam pouco ou nenhum significado ao texto (como artigos, preposições e pronomes). Existem várias listas de *stopwords* disponíveis na Internet nos mais variados idiomas e, para este trabalho, foi utilizada uma lista de *stopwords* disponível em https://gist.github.com/alopes/5358189. O *github* é um repositório

> bastante utilizado e conhecido por conter códigos e documentos que podem ser reutilizados por outros programadores. A relevância desse repositório foi considerada suficiente para a escolha da lista de *stopwords* em português utilizada neste trabalho.

Foram implementadas as duas funções acima descritas. Com base na frase-exemplo, as pontuações são removidas pela *tokenização*:

Frase: *as minhas notas estão baixas~~,~~ assim eu não irei aprovar na disciplina~~...~~*

E depois, as palavras presentes na lista de stopwords também são removidas:

Frase: ~~*as*~~ ~~*minhas*~~ *notas* ~~*estão*~~ *baixas assim* ~~*eu*~~ *não* ~~*irei*~~ *aprovar* ~~*na*~~ *disciplina*

Após os dois serviços da etapa de pré-processamento, só restam as palavras que atribuem significado ao texto e a frase-exemplo fica assim:

Frase: *notas baixas assim não aprovar disciplina*

4.2 DISTRIBUIÇÃO DE PESOS

Agora que o documento já está processado, só restam palavras que contribuem e incrementam significado ao documento. Com isso, o documento está pronto para seguir a cadeia de funções a qual é submetido e, como próxima etapa, o foco é a distribuição de peso de cada palavra de acordo com sua relevância para o documento. Para isso, são utilizadas duas técnicas: a implementação do algoritmo *tf-idf* e o incremento de peso baseado no trabalho de Pasqualotti (2008), que assegura que palavras que demonstram emoção são mais significativas quando se quer identificar o estado de ânimo do autor do texto.

O *tf-idf*, visto na seção 2.3, é uma maneira inteligente e eficaz de distribuir pesos paras as palavras que restaram no documento, pois palavras que se repetem muito tanto nas postagens motivadas como desmotivadas terão valores menores.

Em 2008, Pasqualotti agrupou substantivos e adjetivos em 15 conjuntos de palavras que representam emoção. Cada grupo de palavras representa um conjunto de sentimentos: amor, ódio, raiva, entusiasmo e indiferença são exemplos de grupos de palavras distintos. Todas as palavras desses grupos, sejam elas de sentimento positivo ou negativo, são duplicadas, fazendo com que o classificador atribua uma maior importância a elas. Primeiramente, o peso dessas palavras era incrementado em 1 (um) mas depois de alguns testes percebeu-se que a taxa de acerto melhorou quando o peso era duplicado.

Tendo a implementação das duas fases da distribuição dos pesos pronta, a base de dados já pode ser utilizada de forma que se extraia conhecimento dela e se aplique na classificação. Para isso, a base é dividida em duas partes: uma base de dados só com as postagens de alunos motivados e outra parte com as postagens dos alunos desmotivados. Dessa forma, cada palavra terá obrigatoriamente um peso positivo (quando está presente na base dos motivados) e um peso negativo (quando a palavra aparece na base dos alunos desmotivados). Assim, existem três possibilidades na distribuição de pesos:

• Quando uma palavra só existe na base dos motivados – Nesta condição, o peso negativo da palavra será zero.

• Quando uma palavra só existe na base dos desmotivados – Analogamente à situação acima, desta vez o peso positivo será zero.

• Quando uma palavra existe em ambas as bases – Nesse caso, ambos os pesos são armazenados normalmente.

Voltando a frase-exemplo, seguindo os passos acima, cada palavra da frase terá dois pesos definidos pelo *tf-idf*. A tabela 2 tem os pesos de cada uma das palavras. Vale salientar que para melhor visualização, os pesos da tabela em questão são fictícios.

Frase: *notas baixas assim não aprovar disciplina*

Tabela 2 – Distribuição de pesos

Motivado	Peso
notas	0,3
baixas	0,1
assim	0,2
não	0,2
aprovar	0,7
disciplina	0,4

Desmotivado	Peso
notas	0,5
baixas	0,8
assim	0,1
não	0,4
aprovar	0,2
disciplina	0,3

Fonte: O Autor

4.3 TRATAMENTO DA NEGAÇÃO

A negação é um dos maiores problemas em mineração de texto pois ela inverte o sentido da opinião (BECKER e TUMITAN, 2013). Assim, por exemplo, nas frases (“A aula do professor é imersiva”) e (“A aula do professor não é imersiva”) a palavra “imersiva” tem sentidos diferentes o que confunde o classificador e consequentemente diminui a acurácia do sistema.

Como na língua portuguesa o problema da negação ainda é um problema em aberto como visto no parágrafo acima, para este trabalho foi implementada uma técnica idealizada pelo autor e orientador deste trabalho que muda o sentido das duas próximas palavras após a ocorrência de um “não” no documento. Foram feitos testes mudando o sentido da próxima até as próximas 10 palavras depois do “não” e o teste que se demonstrou mais eficaz ao classificador foi o que alterou apenas o sentido das duas próximas palavras após a negação.

De acordo com o descrito nesta seção, as duas próximas palavras após o “não” na frase-exemplo terão sentido invertido. Para melhor visualização, as palavras estão em negrito.

Frase: *notas baixas assim não* ***aprovar disciplina***

4.4 CLASSIFICAÇÃO

Na etapa de classificação, cada postagem será vista como um somatório de pesos, ou seja, para cada palavra da postagem, é consultado seu maior peso (peso positivo ou negativo). Nos casos e que uma palavra tem o peso negativo maior, esse peso terá o sinal negativado no somatório, diminuindo seu valor final. Desta maneira, ao final da postagem, se o somatório for maior ou igual a zero, a postagem é dita como uma postagem de um aluno motivado e, caso o somatório tenha valor menor que zero, a postagem é considerada uma postagem desmotivada.

Novamente, observando o comportamento da classificação da frase exemplo:

Frase: *notas baixas assim não* ***aprovar disciplina***

-0,5 + (-0,8) + 0,2 + (-0,4) + (**-0,7**) + (**-0,4**) = -2.6

Desmotivado

Tabela 2 – Distribuição de pesos

Motivado	**Peso**
notas	0,3
baixas	0,1
assim	0,2
não	0,2
aprovar	0,7
disciplina	0,4

Desmotivado	**Peso**
notas	0,5
baixas	0,8
assim	0,1
não	0,4
aprovar	0,2
disciplina	0,3

Fonte: O Autor

Seguindo a política da construção do somatório e observando o valor dos pesos de cada palavra determinado pelo *tf-idf*, tem-se que a frase-exemplo é desmotivada, pois seu somatório obteve valor negativo.

Vale ressaltar a importância do tratamento da negação que inverteu o sentido das palavras “aprovar” e “disciplina”, que têm peso motivado maior mas por aparecerem após a negação, acabaram por decrementar o somatório.

5. RESULTADOS

Este capítulo apresenta a descrição e o detalhamento da base de dados e métricas de avaliação utilizadas bem como todos os resultados obtidos durante o desenvolvimento da ferramenta. Todos testes realizados são discutidos e sumarizados em tabelas, facilitando a visualização do progresso nas etapas da implementação.

5.1 BASE DE DADOS

A base de dados que serviu como entrada para a ferramenta foi construída durante o semestre 2015.2, fruto da disciplina de Tópicos Avançados em Inteligência Artificial onde o foco principal foi a mineração de texto. Na disciplina, todos os alunos tinham como exercício semanal a tarefa de simular postagens de alunos matriculados em matemática discreta num fórum educacional. A cada semana os alunos escreviam dez textos de no mínimo vinte palavras com o intuito de retratar o estado de ânimo do aluno, ou seja, se numa semana um grupo de alunos simulavam um aluno decepcionado com a disciplina (sentimento negativo), na semana seguinte este grupo simularia alunos motivados (sentimento positivo).

Após a etapa de escrita foi realizada uma validação das postagens. O principal objetivo é identificar se as postagens estavam dentro da polaridade proposta. Cada aluno analisou entre 30 e 40 postagens por semana, atribuindo a polaridade positiva, negativa ou inconclusiva. As postagens que foram classificadas como inconclusivas por um dos alunos foi automaticamente excluída do banco de dados. Além disto, quando houve alguma divergência entre os alunos ela foi resolvida pelo professor da disciplina.

O total de postagens obtida foi de 460 dividido nas diferentes categorias. Em relação a análise de sentimento, inicialmente foram escritas 230 postagens positivas e 230 negativas. Após a validação das postagens esse número ficou em 106 e 124 para positivas e negativas respectivamente, como mostrado na tabela 3. Para

utilização no sistema proposto, foram utilizadas apenas as postagens após a validação.

Tabela 3 – Quantidade de postagens antes e depois da validação

	Antes da validação	**Depois da validação**
Motivado	230	106
Desmotivado	230	124

Fonte: O autor.

É importante afirmar que a base de dados também contém postagens de alunos com e sem dúvidas e também respostas a exercícios plagiadas da Internet, pois o propósito da construção da base é que ela sirva para outros trabalhos, que tenham motivações diferentes do combate à evasão.

5.1.1 Metodologia de Avaliação

A taxa percentual de acertos, que é a saída do sistema, é medida pela acurácia do sistema, que é a média das 10 *folds* da validação cruzada. A seguir, as subseções 5.1.1.1 e 5.1.1.2 detalham essas duas métricas.

5.1.1.1 Acurácia

Acurácia é o quão próximo o valor obtido experimentalmente está do valor real (TSYTSARAU e PALPANAS, 2012), ou seja, o quão próximo o experimento aproximou-se da meta. No caso deste trabalho, a acurácia ou taxa de acertos é o valor percentual do quão próximo a classificação esteve perante a realidade, ou seja, quantas instâncias percentualmente o classificador conseguiu acertar.

5.1.1.2 Cross-Validation

O *10-fold cross-validaton* é uma metodologia de avaliação utilizada em Inteligência Artificial (KOHAVI, 1995) onde a base é dividida em 10 partes iguais e,

por 10 iterações, 9 partes são usadas para treino e 1 para teste. Vale salientar que no fim das iterações, cada parte é testada apenas uma vez. Esta técnica permite uma avaliação mais fiel, por eliminar a possibilidade da escolha do treino favorecer o teste.

5.2 COMPARAÇÃO DOS RESULTADOS

O que foi discutido nas seções anteriores descreve o estado final da ferramenta após vários testes e tentativas feitas para melhorar a acurácia final do sistema. Entretanto, para que esse estado de maturidade fosse atingido, foram realizados testes com configurações diferentes às vistas até aqui. Foram avaliados quatro cenários: a utilização apenas do classificador com validação cruzada, o incremento dos pesos das palavras dos grupos emoção (PASQUALOTTI, 2008) em 1, otimização com a técnica da negação citada na seção 2.2.3 e a duplicação dos pesos das palavras de emoção ao invés do incremento em 1, chegando ao estado final deste trabalho. Abaixo, a tabela 4 que mostra as acurácias obtidas de acordo com o avanço da implementação da ferramenta.

Tabela 4 – Resultados obtidos

Estado	Abordagem	Acurácia
1	Classificador	66,16%
2	Classificador + Palavras de emoção (+1)	69,43%
3	Classificador + Palavras de emoção (+1) + Tratamento de negação	71,90%
4	Classificador + Palavras de emoção (x2) + Tratamento de negação	73,57%

Fonte: O autor.

O Estado 1 faz uso da técnica da validação cruzada anteriormente mencionada. Foi implementado o *cross-validation* com 10 *folds* onde cada um tem

uma acurácia associada e a taxa de acerto final é a média das dez acurácias obtidas.

Com o estudo de Pasqualotti (2008), foram atribuídos pesos maiores às palavras dispostas nos chamados Grupos de Palavras de Emoção, que são substantivos e adjetivos que apontam o estado de ânimo do autor. Essas palavras tiveram seus pesos incrementados em uma unidade, o que elevou a taxa de acerto do Estado 2 de 66,16% para 69,43%.

Tendo em vista os trabalhos relacionados descritos neste trabalho, a acurácia de 70% é vista como boa pela maioria deles e também em demais trabalhos que lidam com mineração de texto em português. Sendo assim, no Estado 3 essa taxa foi batida ao implementar a técnica para tratamento de negação, chegando à 71,90% de instâncias corretamente classificadas.

Por fim, no Estado 4, ao invés do incremento, os pesos das palavras dos grupos de emoção foram duplicados, atingindo a acurácia final do sistema (73,57%). Isto mostra que o incremento estava muito alto, confundindo o classificador.

5.3 RESULTADOS DO TRATAMENTO DA NEGAÇÃO

Na tabela 2, os pesos das palavras “aprovar” e “disciplina” na base dos motivados são 0,7 e 0,4, respectivamente e 0,2 e 0,3 na base dos desmotivados. Na frase-exemplo que ilustrou a classificação desse trabalho (seção 4.4), as palavras que ocorrem após o “não” têm peso maior motivado e, por isto, apenas o sinal de cada peso foi trocado. Um outro teste feito foi além de inverter o sinal, usar o peso desmotivado de cada uma das duas palavras (no lugar de -0,7 e -0,4 usou-se -0,2 e 0,3) mas o resultado da classificação permaneceu o mesmo: 73,57%.

Como descrito ao fim da seção 4.3, a forma de lidar com o problema da negação foi inverter o sinal dos pesos das duas palavras que aparecem imediatamente após o “não”. Foram realizados testes alterando os sinais da primeira até a décima palavra após a ocorrência da negação. A tabela 5 mostra as acurácias obtidas nos testes realizados.

Tabela 5 – Testes do tratamento da negação

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Acurácia	71,30 %	73,57 %	73,01 %	72,71 %	70,23 %	71,17 %	70,54 %	69,86 %	68,03 %	68,14 %

Fonte: O Autor.

5.4 ESTUDO DAS PALAVRAS COM MAIORES PESOS

Devido ao grande número de palavras do documento, a probabilidade de cada palavra ocorrer é extremamente pequena, o que gera valores *tf-idf* também muito pequenos, todos muito menores que 1 (um). Sendo assim, essa seção faz um estudo dos maiores valores *tf-idf* obtidos tanto nas bases dos alunos motivados quanto dos desmotivados. Vale lembrar que por conta do incremento de peso das palavras de emoção, estas tiveram seus valores *tf-idf* duplicados. Logo, para melhor estudo da distribuição de peso, são mostradas tabelas com e sem o incremento.

Como visto acima, as palavras dos Grupos de Palavras de Emoção têm seu peso duplicado. Pode-se notar que todas são substantivos ou adjetivos que buscam evidenciar a emoção do autor. Abaixo segue a tabela 6, que contém estas palavras.

Tabela 6 – Pesos das palavras dos grupos de emoção

Palavra	Peso	Polaridade
animado	0.006032239366084546	Motivado
apaixonado	0.0026581294332714833	Motivado
atenção	0.004499352501104781	Motivado
dedicado	0.0026581294332714833	Motivado
gosto	0.004499352501104781	Motivado
gratificante	0.0026581294332714833	Motivado
impressionante	0.0026581294332714833	Motivado
arrependido	0.0022377007116318737	Desmotivado
louco	0.0022377007116318737	Desmotivado
medo	0.00381416035645632	Desmotivado
preocupado	0.0022377007116318737	Desmotivado
problema	0.0022377007116318737	Desmotivado
terrível	0.005141038692655364	Desmotivado

Fonte: O autor

Quando não há o incremento de peso, as palavras com maiores *tf-idf* são palavras relacionadas ao tema do documento (tabela 7), no caso do dataset, alunos que tentam aprender lógica matemática. Nota-se que as palavras se repetem em ambas as partes da base de dados, tanto nos motivados quanto nos desmotivados. Isto ocorre porque essas palavras se repetem com maior frequência no decorrer de todo o documento, independente da postagem do aluno ser positiva ou não. Nesse caso, no cálculo do somatório de cada postagem leva-se em consideração o maior valor e, caso a palavra ocorra até duas posições depois de um “não”, tem seu sinal trocado, ou seja, sua polaridade alterada, conforme descrito na seção do tratamento de negação.

Tabela 7 – Maiores pesos das palavras ausentes nos grupos de emoção

Palavra	Peso	Polaridade
assunto	0.010291934754526008	Motivado
aula	0.009161495175341226	Motivado
bem	0.011813678242435212	Motivado
entender	0.011097697742893686	Motivado
exercícios	0.011294915339531858	Motivado
lógicas	0.009161495175341226	Motivado
professor	0.010251454943381064	Motivado
assunto	0.00929862018065958	Desmotivado
aulas	0.009300999205031727	Desmotivado
cadeira	0.008236967202155178	Desmotivado
difícil	0.008694576491163798	Desmotivado
entender	0.00927647806044798	Desmotivado
exercícios	0.008749110838409494	Desmotivado
professor	0.009458146380390392	Desmotivado

Fonte: O autor.

Durante os testes realizados, as palavras que se repetiram em ambos os arquivos *tf-idf* foram removidas dos somatórios restando apenas aquelas que só ocorriam em um deles mas a acurácia final do sistema caiu dos 73,57% para 67,08%, o que constata que as palavras que se repetem no quadro acima são importantes para o classificador por representarem o tema e assunto do documento.

5.5 DISCUSSÃO

Como discutido nas seções iniciais, o principal objetivo deste trabalho é oferecer uma alternativa de combate à evasão aos professores e tutores de cursos de EaD. Um outro ponto já visto foi que a quantidade de informação gerada nos fóruns educacionais dificulta o acompanhamento individual dos alunos por parte do professor.

Usando a ferramenta proposta, o professor pode dar mais atenção aos alunos que tiveram suas postagens polarizadas como desmotivadas. Este alerta pode servir como indício que o autor da postagem está com dificuldades no aprendizado e o professor ou tutor pode, por sua vez, poupar seu tempo dando auxílio maior apenas aos alunos com potencial chance de evasão.

Outra forma de se melhorar o ensino do curso é dar atenção às palavras com maior *tf-idf* (tabelas 6 e 7). Desta maneira, o professor pode identificar o sentimento dos seus alunos visualizando as palavras da primeira tabela e evidenciar na tabela 5 palavras-chave relacionadas ao tema do seu curso que possivelmente indicam a causa de problemas no aprendizado. Vimos que na tabela 6 as palavras “exercícios” e “difícil” aparecem com pesos maiores nas postagens desmotivadas e isso pode ser um indício de que os exercícios propostos pelo professor podem estar com dificuldade elevada, desmotivando alguns alunos.

6. CONCLUSÃO

Este trabalho apresentou um sistema para identificação de ânimo de alunos a partir de suas postagens em fóruns educacionais. O trabalho utiliza uma abordagem mista, que combina técnicas estatísticas e semânticas para atingir o objetivo final.

Para validar a proposta foi utilizando um banco com 230 postagens divididas nas classes positiva e negativa. Os trabalhos anteriores que fazem análise relacionada a proposta utilizam bases consideravelmente menores para avaliar a proposta. O sistema atingiu uma acurácia de 73,57% ao classificar o aluno como motivado ou desmotivado.

Além de relacionar a análise de sentimento com o problema da evasão, este trabalho contribuiu com a técnica proposta para combater o problema da negação em mineração de texto. De acordo com os testes realizados, a solução implementada foi vista como promissora, podendo ser utilizada em outros trabalhos que lidam com mineração de texto na língua portuguesa.

Por fim, o trabalho apresentou alguns cenários onde o professor pode se beneficiar da identificação de ânimo. Os principais listados são: combater a evasão e identificação dos assuntos que estão gerando mais dúvidas entre os alunos.

6.1 TRABALHOS FUTUROS

Como trabalhos futuros, pode-se usar a teoria dos grafos no problema estudado. Existem problemas em grafos cuja solução já é conhecida e que podem se adaptar aos problemas identificados neste trabalho, como o da negação. Seria interessante aplicar algoritmos de busca em grafos ponderados com o objetivo de identificar a quantidade correta de palavras a terem o sentido modificado após a ocorrência do “não”.

Uma outra técnica que pode melhorar a classificação da ferramenta é o uso da *stemização* na etapa de pré-processamento. Na tabela 7, “aula” e “aulas” têm pesos diferentes e como a *stemização* reduz cada palavra do documento a seu

radical, essas e outras várias palavras seriam representadas por um único radical e consequentemente, por um único valor *tf-idf*.

Por fim, a utilização de algoritmos evolutivos também pode melhorar a classificação, pois após a execução do *tf-idf*, o próprio algoritmo evolutivo se encarregaria a descobrir os melhores para cada palavra presente no documento.

EFERÊNCIAS

AGGARWAL, Charu C.; ZHAI, ChengXiang. ***Mining text data***. Springer Science & Business Media, 2012.

ALVES, Thyanne Michelle Ferreira; MENEZES, Afonso Henrique Novaes; VASCONCELO, Flávia Maria de Brito Pedrosa. *CRESCIMENTO DA EDUCAÇÃO A DISTÂNCIA E SEUS DESAFIOS: UMA REVISÃO BIBLIOGRÁFICA*. **Revista de Educação do Vale do São Francisco-REVASF**, v. 4, n. 6, p. 63-74, 2015.

ARCHAK, Nikolay; GHOSE, Anindya; IPEIROTIS, Panagiotis G. *Show me the money!: deriving the pricing power of product features by mining consumer reviews*. In: **Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining**. ACM, 2007. p. 56-65.

AVILA, Christiano Martino Otero et al. *Desenvolvimento de um Sistema de Recomendação de Artigos Científicos e Avaliação de Métodos de Extração de Palavras-Chave*. 2006.

AZEVEDO, Breno Fabrício Terra; BEHAR, Patricia Alejandra; REATEGUI, Eliseo Berni. *Aplicação da mineração de textos em fóruns de discussão*.**RENOTE**, v. 8, n. 3, 2010.

BARROSO, Marta F.; FALCÃO, Eliane BM. *Evasão universitária: o caso do Instituto de Física da UFRJ*. **ENCONTRO NACIONAL DE PESQUISA EM ENSINO DE FÍSICA**, v. 9, p. 1-14, 2004.

BASTOS, Helvia Pereira Pinto; BERCHT, Magda; WIVES, Leandro Krug. *Presença social e pertencimento em fóruns educacionais: manifestação e percepção de afetividade*. In: **Anais do Simpósio Brasileiro de Informática na Educação**. 2011.

BECKER, Karin; TUMITAN, Diego. *Introdução à mineração de opiniões: Conceitos, aplicações e desafios*. **Simpósio Brasileiro de Banco de Dados**, 2013.

BOLLEN, Johan; MAO, Huina; ZENG, Xiaojun. *Twitter mood predicts the stock market*. **Journal of Computational Science**, v. 2, n. 1, p. 1-8, 2011.

DAMASCENO, Mariana Paiva; MELO, Marlene Catarina de Oliveira Lopes; DE MUYLDER, Cristiana Fernandes. *Educação à Distância em Foco: Um Estudo sobre a Produção Científica Brasileira*. **Revista de Administração Mackenzie**, v. 16, n. 4, 2015.

DE ALMEIDA BIZARRIA, Fabiana Pinto et al. *Papel do tutor no combate à evasão na EAD: percepções de profissionais de uma instituição de ensino superior*. **Educação, Ciência e Cultura**, v. 20, n. 1, p. p. 85-102, 2015.

DE SOUZA AFIUNE, Cally. *Mineração de Dados Educacionais: Predição comportamental em Ambientes de Educação a Distância (EaD)*. 2012.

DEVI, G. Dharani; RASHEED, A. Abdul. *A Survey on Sentiment Analysis and Opinion Mining*. 2015.

KOHAVI, Ron et al. *A study of cross-validation and bootstrap for accuracy estimation and model selection*. In: **Ijcai**. 1995. p. 1137-1145.

LIU, B. *Sentiment Analysis and Opinion Mining: Synthesis Lectures on Human Language Technologies*, vol. 16. **Morgan & Claypool Publishers, San Rafael**, 2012.

LONGHI, Magalí T. et al. *Investigando a subjetividade afetiva na comunicação assíncrona de ambientes virtuais de aprendizagem*. In: **Anais do Simpósio Brasileiro de Informática na Educação**. 2009.

MAIA, Marta de Campos. ***O uso da tecnologia de informação para a educação a distância no ensino superior***. 2003. Tese de Doutorado.

PANG, Bo; LEE, Lillian. *Opinion mining and sentiment analysis*.**Foundations and trends in information retrieval**, v. 2, n. 1-2, p. 1-135, 2008.

PASQUALOTTI, Paulo Roberto. *Reconhecimento de expressões de emoções na interação mediada por computador*. 2008.

PEREIRA, Alice T. Cybis; SCHMITT, Valdenise; DIAS, Maria Regina Álvares C. *AVA-Ambientes Virtuais de Aprendizagem em diferentes contextos*. **Rio de Janeiro: Editora Ciência Moderna Ltda**, 2007.

REIS, Ederclinger M.; VASCONCELOS, F. Herbert Lima; MARTINS, Cibelle A. *Uma análise da interação em fóruns de discussão em um ambiente virtual de aprendizagem*. In: **Anais do Simpósio Brasileiro de Informática na Educação**. 2010.

RIGO, Sandro J. et al. *Abordagem linguística para identificação da dimensão afetiva expressa em textos de Ambientes Virtuais de Aprendizagem–um Léxico da Emoção*. In: **Anais do Simpósio Brasileiro de Informática na Educação**. 2013. p. 738.

SARAWAGI, Sunita. *Information extraction*. **Foundations and trends in databases**, v. 1, n. 3, p. 261-377, 2008.

SILVA, Francisco et al. *Um modelo preditivo para diagnóstico de evasão baseado nas interações de alunos em fóruns de discussão*. In: **Anais do Simpósio Brasileiro de Informática na Educação**. 2015. p. 1187.

TRAN, Véronique. ***The influence of emotions on decision-making processes in management teams=(l'influence des émotions sur les processus de prise de décision dans les équipes de cadres)***. 2004. Tese de Doutorado. University of Geneva.

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