Análise de Sentimento Utilizando Processamento de Linguagem Natural

Pablo Fagundes Wachsmann

Curso de Bacharelado em Ciência da Computação – Universidade do Vale do Itajaí (UNIVALI) – Campus Itajaí
88302-901 – Itajaí – SC– Brasil

pfagundesw@gmail.com

Abstract. *With the massive data generated by users on the Internet, it becomes increasingly necessary to use tools that are capable of processing these documents and extract useful information. The Natural Language Processing (NLP) has the ability to provide methods capable of interpreting and processing texts written in human language. A task used with NLP is the sentiment analysis, which can be performed through several methods and aims to extract the sentiment or opinion expressed by the author of a given text. In this way, in the present work was developed a tool that performs analysis of opinions collected regarding the change of ENEM 2017 registration standards, identifying positive, negative or neutral sentiments. With the analysis of the results, it was possible to reach the intended objective and to conclude that the system is able to classify the opinions better than a person who does not have specific knowledge in linguistics.*

Resumo. *Com a massa de dados gerada por usuários na Internet, se faz cada vez mais necessário o uso de ferramentas que sejam capazes de processar estes documentos e extrair informações úteis. O Processamento de Linguagem Natural (PLN) possui a capacidade de fornecer métodos capazes de interpretar e processar textos escritos em linguagem humana. Uma tarefa utilizada em conjunto com o PLN é a análise de sentimento, que pode ser realizada através de diversos métodos e visa extrair o sentimento ou opinião expressada pelo autor de determinado texto. Desta forma, neste trabalho foi desenvolvido uma ferramenta que realiza análise de opiniões coletadas a respeito da mudança das normas de inscrição do ENEM 2017, identificando sentimentos positivos, negativos e neutros. Com a análise dos resultados, foi possível atingir o objetivo previsto e concluir que o sistema é capaz de classificar as opiniões melhor do que uma pessoa que não possui conhecimentos específicos em linguística.*

1. Introdução

Segundo Sun, Luo e Chen (2016), com o grande crescimento de textos gerados por usuários na Internet, a extração automática de informação útil de tantos documentos, desperta interesse de pesquisadores em diversas áreas, em particular na área de Processamento de Linguagem Natural (PLN). Ao contrário de linguagens artificiais, como linguagens de programação e notações matemáticas, linguagens naturais evoluem

com o passar das gerações e são difíceis de se definir com regras explícitas. O PLN, oferece ferramentas para a manipulação de linguagens naturais, podendo ser simples como contagens de frequência de palavras para comparar diferentes estilos de escrita, como o completo entendimento de enunciados humanos (BIRD; KLEIN; LOPER, 2009).

De acordo com Covington (1994), o PLN define-se como o uso de computadores para entender linguagens humanas. Não no sentido de um computador poder expressar sentimentos ou pensamentos, mas de maneira que um computador possa reconhecer e usar informação expressada em linguagem humana. Ainda segundo Covington (1994), a estrutura de qualquer linguagem se divide em cinco níveis: fonóloga, que é expressada através dos sons; morfológica, interpretada através das formações das palavras; sintática, levando em consideração a estrutura das sentenças; semântica, com foco nos significados das palavras em si; e pragmática, representando o uso da linguagem dentro de um determinado contexto.

Segundo Devika, Sunitha e Ganesh (2016), a Análise de Sentimento é um processo intelectual de extração de sentimentos e emoções de usuários, sendo um dos campos que procedem o PLN. É o processo de detecção de polaridade em textos, bem como determinar se a opinião de um dado texto é positiva, negativa ou neutra.

A Análise de Sentimento (AS) ou Mineração de Opinião (MO), é o estudo computacional de opiniões, atitudes e emoções de acordo com uma entidade, sendo que esta entidade pode ser representada por indivíduos, eventos ou tópicos. As expressões Análise de Sentimento e Mineração de Opinião possuem o mesmo significado, entretanto alguns pesquisadores apontam uma pequena diferença entre elas. A MO extrai e analisa opiniões pessoais sobre uma entidade, enquanto a AS identifica um sentimento expressado em um texto e então o analisa. Portanto, o objetivo da AS é encontrar opiniões, identificar o sentimento que essas opiniões expressam e então classificar sua polaridade (MEDHAT; HASSAN; KORASHY, 2014).

2. Processamento de Linguagem Natural

O PLN, visa o desenvolvimento de ferramentas para se executar as tarefas necessárias no pré-processamento de um texto a ser analisado. Segundo Sun, Luo e Chen (2016), etapas de pré-processamento são fundamentais para qualquer análise mais detalhada sobre algum tema específico, para se estruturar o texto a ser analisado e extrair suas características. Tarefas como tokenization (separação de uma sequência de caracteres separados por tokens), desambiguação e part-of-speech tagging (designar categorias para cada palavra identificada), são de fundamental importância para que posteriormente possa ser implementado uma análise de sentimento do texto estudado.

2.1 Part-of-Speech Tagging

Part-of-speech (POS) Tagging é o processo de “etiquetagem”, atribuindo classes ou outra marcação léxica de classe para cada palavra dentro do texto, estas marcações também são aplicadas para pontuações e em muitos casos podem ser ambíguas. O POS representa uma importante tarefa no processo de análise da linguagem natural e na recuperação da informação. A entrada para um algoritmo de POS é uma string de palavras e um conjunto de “etiquetas” especificadas para descrever as classes

(JURAFSKY; MARTIN, 2000). Desta forma, se for estabelecido que os verbos serão representados pela etiqueta VB, artigos por AR, adjetivos por AD e pronomes como PR, se teria uma representação semelhante a demonstrada na Figura 1:

Figura 1. POS da frase "As ruas estavam escuras".

Fonte: adaptado de Jurafsky e Martin (2000).

Jurafsky e Martin (2000) descrevem que esta tarefa é importante, pois existem palavras que podem ser ambíguas. O objetivo do POS é resolver ambiguidades etiquetando as palavras de forma correta dentro do contexto da sentença. Segundo Maletti (2017), uma vez que utilizamos o PLN para preparar os textos para a AS ou MO, e o objetivo final é o de encontrar polaridades, ao se realizar o POS, deve-se ater principalmente as classes que por si só possuem conteúdo, como substantivos, adjetivos, verbos e negações. Deixando de lado classes que expressam funções, como artigos, pronomes, verbos auxiliares, pois palavras que se enquadram nestas classes normalmente servem uma função gramatical e não contém significado léxico. Maletti (2017) alerta para a ambiguidade. A palavra da língua portuguesa "bem", por exemplo, pode fazer parte de pelo menos cinco classes (substantivo, advérbio, pronome, adjetivo e interjeição), observando que a tarefa de POS é muito importante para se tratar estas ambiguidades.

3. Análise de Sentimento

Análise de Sentimento (AS), também chamada de Mineração de Opinião (MO) é o estudo computacional de sentimentos, opiniões, atitudes e emoções expressadas por pessoas a respeito de uma entidade, que pode representar um indivíduo, evento ou tópico. Destes tópicos estudados em AS, destacam-se opiniões de usuários na internet. AS pode ser considerada um processo de classificação e existem três níveis principais. O nível de documento, que visa classificar a opinião de um documento expressando um sentimento positivo ou negativo. Considera todo o documento como base de informação para a tarefa. O nível de sentença, que visa classificar o sentimento expressado em cada sentença da entidade, primeiro identifica se a sentença é objetiva ou subjetiva, e caso seja subjetiva, esta tarefa determina se a sentença expressa um sentimento positivo ou negativo. Finalmente, o nível de aspecto, visa classificar o sentimento que diz respeito aos aspectos específicos de cada entidade. Primeiro, identifica-se a entidade em si e seus aspectos e pode se extrair diferentes opiniões para diferentes aspectos de uma entidade (MEDHAT, HASSAN, KORASHY, 2014).

A maior parte das técnicas de AS podem ser divididas em abordagens de Machine Learning (ML) e Léxicas. Apesar de que as abordagens de ML tenham obtido avanços significativos na área, aplicá-los requer um conjunto de treinamento rotulado. Conseguir organizar um conjunto de dados de treinamento pode requerer tempo e

esforço consideráveis, uma vez que estes dados precisam estar atualizados. A fim de eliminar esta tarefa, são propostos sistemas que geram dados a partir de um conjunto de documentos anotados chamado de corpus (MOREO et al, 2012).

3.1. Técnica Léxica

Também chamado de análise baseada em dicionário, esta abordagem, apesar de possuir a vantagem de não fazer uso de um conjunto de treinamento, tem pontos negativos. A maioria dos sistemas por exemplo, utiliza websites com glossários de palavras que podem não conter termos técnicos ou expressões coloquiais. Uma vez que é incapaz de considerar o contexto das expressões, esta técnica não costuma obter uma precisão muito alta para cenários de multi-domínios, onde uma palavra pode expressar uma opinião positiva ou negativa de acordo com o seu contexto (MOREO et al, 2012).

Dentro deste contexto, Kim e Hovy (2004) propõe um sistema de análise de sentimento baseado em dicionário. O sistema tem como objetivo encontrar os sentimentos expressados sobre um dado tópico e identificar a pessoa que o expressou. Para evitar problemas com tons de sentimentos foi especificado que o sistema iria identificar sentimentos positivos, negativos ou neutros. Sentenças que não expressam opinião, apenas estabelecem fatos, foram classificadas em um conjunto separado.

Dado um tópico e um conjunto de textos a respeito do tópico, o sistema realiza quatro passos: (i) selecionar as sentenças que contém ambos os tópicos e os detentores de opinião; (ii) delimitar as regiões que se baseiam no detentor da opinião; (iii) calcular a polaridade de todas as palavras que contém sentimento individualmente; e (iv) combinar as seleções e análises anteriores para determinar o sentimento da sentença como um todo (KIM; HOVY, 2004).

Os experimentos foram realizados com a ajuda de três pessoas que classificaram as palavras como sendo positivas ou negativas. O experimento, primeiro comparou os resultados entre as pessoas, a fim de eliminar qualquer disparidade nos testes. E posteriormente comparou os resultados da classificação de palavras do sistema com a classificação humana. Dos dados testados, o sistema classificou 93,07% dos adjetivos e 83,27% dos verbos como sendo positivos ou negativos. O sistema obteve um nível de assertividade de 75,66% com o “humano 1”, 77,88% com o “humano 2”, 81,20% com o “humano 3”.

4. Projeto

O objetivo do presente trabalho, foi o de desenvolver um sistema que realizasse a tarefa de AS de um texto inserido na língua portuguesa. Desta forma, propôs-se o desenvolvimento de um sistema para realizar a análise de sentimento através do uso da abordagem Léxica, identificando sentimentos positivos, negativos e neutros, nos textos inseridos.

A escolha do uso da técnica Léxica justificou-se pelo fato de que abordagens ML exigem o uso de um conjunto de treinamento onde as entradas possuem saídas conhecidas (MEDHAT, HASSAN, KORASHY, 2014), o qual não se possuía. Assim propôs-se a utilização das seguintes técnicas de PLN:

1. Segmentação: utilizada para realizar a segmentação das sentenças, separando as frases em suas palavras específicas, visando identificar as palavras que serão utilizadas na tarefa seguinte de classificação;

2. POS Tagging: utilizada para atribuir classes específicas as palavras. Identificando as principais classes utilizadas na contagem das palavras positivas e negativas da AS.

Para a análise de sentimento, propôs-se a criação de um dicionário de palavras. As palavras especificadas neste dicionário foram, conforme Maletti (2017), adjetivos, pois possuem conteúdo por si só, obtendo assim maior significado léxico. Posteriormente incrementando a quantidade de palavras deste dicionário com sinônimos e antônimos retirados da internet.

Quanto aos dados a serem analisados, propôs-se a criação de um formulário Google para pesquisa com um texto explicativo a respeito das alterações na inscrição do ENEM 2017, onde o mesmo apresentava uma série de pontos que foram alterados na inscrição deste ano, como aumento no valor, mudança nos dias de prova e a possibilidade da solicitação de isenção da taxa de inscrição.

Posteriormente propôs-se a análise das opiniões coletadas por dois alunos do Laboratório de Inteligência Aplicada (LIA) da Univali, e de uma especialista, que classificaram as opiniões expressadas através dos textos coletados com o formulário como: positivas, negativas ou neutras. O “Aluno 1”, faz parte do LIA e é aluno regular do oitavo período do curso de Ciência da Computação da Univali, não possuindo especialização ou curso na área de linguística. O “Aluno 2”, faz parte do LIA e é aluno regular do sexto período do curso de Ciência da Computação da Univali, não possuindo especialização ou curso na área de linguística. A “Especialista”, é formada pelo curso de Publicidade e Propaganda e Especialista em Metodologia da Língua Portuguesa e Literatura pela Uniasselvi - Assevim, possuindo experiência em redação publicitária e literária.

Quanto a qualidade do experimento, esperou-se atingir um nível de precisão médio próximo ao do trabalho similar de Kim e Hovy (2014), sendo este valor um percentual de 79,5%.

5. Desenvolvimento

O sistema proposto, foi implementado na linguagem de programação Java. E através da criação de um projeto Apache Maven, foi possível automatizar a compilação e utilizar a biblioteca CoGrOO, que é desenvolvida na mesma linguagem escolhida.

O uso da classe Analyzer da biblioteca CoGrOO, permitiu a realização das tarefas de PLN propostas, ou seja, identificar as palavras através da tarefa de segmentação, e realizar a tarefa de POS Tagging, onde é atribuída uma classe gramatical para cada palavra dentro do texto. Os métodos desta classe, permitiram identificar: (i) os lexemas, ou seja, as palavras que compõem o texto; (ii) seus lemas, que são as palavras em sua forma canônica, por exemplo o lema da palavra “acho” é achar, o da palavra “sendo” é “ser” e assim sucessivamente; e (iii) a classe de cada uma dessas palavras, onde é possível dizer a qual classe gramatical cada palavra pertence, e identificar as classes que possuem importância para a análise, como os adjetivos.

O sistema possui apenas uma tela principal, que pode ser observada na Figura 2, onde é possível selecionar o arquivo com as opiniões a serem analisadas e, ao clicar no botão “Carregar”, o sistema realiza a leitura do arquivo com as opiniões e carrega o dicionário de palavras para comparar com os adjetivos encontrados no texto. O botão “Carregar” também realiza a segmentação das frases lidas a partir do arquivo e a classificação das palavras, bem como a análise da polaridade das opiniões encontradas, e apresenta na parte inferior da tela a saída.

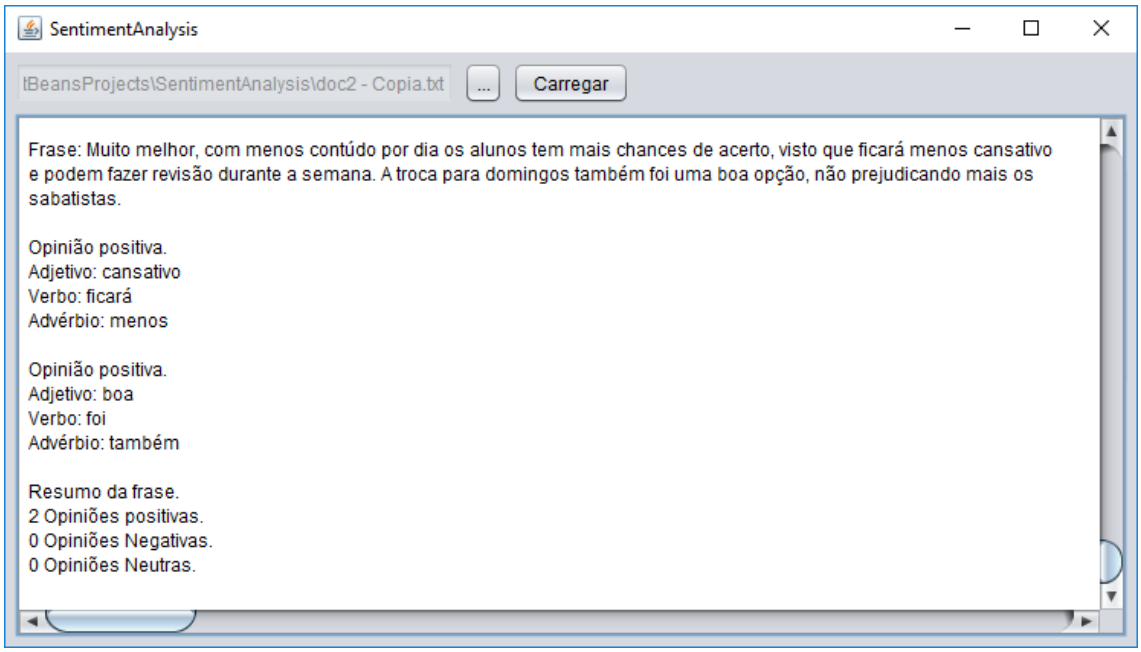

Figura 2. Tela principal.

Conforme a Figura 2, o sistema desenvolvido, identifica cada opinião expressada dentro de uma determinada frase, e a classifica de acordo com as palavras especificadas no dicionário. Para isto o sistema utiliza a seguinte lógica:

1. Percorre o texto inserido buscando um adjetivo, para isto o sistema faz uso da classe gramatical identificada na etapa de PLN com a tarefa de POS Tagging;

2. Ao encontrar um adjetivo, o sistema verifica os limites do texto e volta, no máximo quatro posições de palavras, procurando por verbos que caracterizem uma opinião e advérbios que possam inverter o valor das mesmas. Esta etapa, foi testada voltando cinco, quatro e três posições de palavras dentro do texto. Com cinco palavras, a maioria das opiniões atuais que estão sendo analisadas, se confundem com as anteriores, identificando o dobro de opiniões que o texto realmente possui. Com a análise de três posições de palavras, o sistema não identifica a maioria dos advérbios de negação que invertem a polaridade das opiniões como, não, nem, tampouco;

3. Caso o sistema não tenha encontrado um verbo que caracterize a opinião, ele vai adiante do adjetivo encontrado e, checando o limite do texto, percorre no máximo três posições de palavras buscando por um verbo. Esta etapa também foi testada com quatro posições, mas apresentou o mesmo problema que a etapa 2;

4. Se o texto lido contém apenas uma palavra e esta seja um adjetivo, o sistema identifica esta palavra como uma opinião, a fim de compará-la ao dicionário. Caso o sistema encontre um adjetivo que não possua um verbo próximo, não é considerado uma opinião válida;

5. Por fim, o sistema percorre as opiniões, comparando o adjetivo encontrado, com as palavras estabelecidas no dicionário, e se caso o adjetivo esteja indicado no dicionário como um adjetivo positivo e a opinião não possua um advérbio de negação próximo, é considerada uma opinião positiva e vice-versa.

6. Resultados

Os melhores resultados foram obtidos com a classificação realizada pela especialista, que identificou 51 opiniões distintas, das quais 46 (90,19%) foram identificadas pelo sistema, com um total de 43 opiniões de acerto geral, representando 84,31% de opiniões corretas identificadas e 93,47% de acerto considerando apenas as opiniões identificadas.

As opiniões classificadas pelos Alunos 1 e 2, tiveram um percentual de acerto mais ou menos 25% menor do que as identificadas pela especialista. Das 75 opiniões identificadas pelo Aluno 1, 54 opiniões foram identificadas pelo sistema, representando um total de 72%. Destas, 42 estavam corretas, somando um total de 56% de acerto geral e 77,77% de acerto considerando apenas as opiniões identificadas. O Aluno 2 classificou um total de 58 opiniões, das quais 49 foram identificadas pelo sistema, ou seja, 84,48% das opiniões foram identificadas. Dentre as opiniões identificadas pelo sistema, 34 foram corretas, representando um total de 58,62% de opiniões totais corretas e 69,38% de opiniões corretas dentre as que foram identificadas.

Ao se comparar o resultado obtido pela análise das frases realizadas pelos Alunos, com a análise da Especialista, observou-se um percentual de acerto menor do que o do sistema. Das 51 opiniões identificadas pela Especialista, o Aluno 1 conseguiu identificar 50 (98,03%) e 40 estavam corretas, representando um percentual de 78,43% de acerto geral e 80% de acerto considerando as opiniões identificadas.

O Aluno 2 conseguiu identificar 46 opiniões das 51, representando um total de 90,19% opiniões identificadas. Destas, 34 estavam corretas, o que representa 66,66% de acerto geral e 73,91% de acerto considerando as opiniões identificadas.

Pôde-se observar que as maiores diferenças se concentraram nas opiniões neutras, identificadas pelos Alunos. A análise da especialista, apontou apenas 2 opiniões neutras e o sistema foi capaz de identifica-las corretamente. O Aluno 1 identificou em sua análise, 31 opiniões neutras, e o sistema foi capaz de acertar apenas 10, já o Aluno 2 identificou 17 opiniões como neutras, e o sistema identificou apenas 6 corretamente. Segundo a Especialista, esta diferença deu-se devido a interpretação dos textos analisados. Em sua análise, a Especialista identificou apenas opiniões que continham conteúdo gramatical que caracterizava opiniões, ou seja, verbos e adjetivos que expressavam a opinião do detentor. Já os Alunos, identificaram opiniões realizando inferências e interpretações com base em conhecimento comum. Opiniões como “acho justo” que apareceram 8 vezes nos textos, foram identificadas pela Especialista e pelo sistema, como opiniões positivas, uma vez que “justo” é um adjetivo que expressa

adequação, sensatez. Os Alunos identificaram estas opiniões como neutras, por interpretarem que elas não têm um significado positivo efetivo.

7. Conclusão

O presente artigo refere-se ao Trabalho Técnico-Científico de Conclusão de Curso, do décimo período do curso de Ciência da Computação da Univali e guiou-se a partir de sua Solução Proposta e seus Objetivos. Identificou as técnicas utilizadas pelo PLN para pré-processar textos, segmenta-los nas respectivas palavras desejadas e atribuir classes as palavras através da técnica de POS Tagging. Descreveu os métodos mais utilizados para realizar a tarefa de AS.

Considerando a análise da Especialista como sendo correta, pode-se dizer que o trabalho atendeu seu objetivo referente a qualidade do experimento, obtendo um percentual de 84,31% de acerto geral, ficando acima do percentual esperado inicialmente de 79,5%. Com a análise das opiniões da Especialista, comparada a análise dos Alunos, pode-se concluir também que o sistema é capaz de classificar as opiniões melhor que uma pessoa que não possui conhecimento específico na área de linguística, uma vez que o percentual de acerto geral dos Alunos 1 e 2 foi de 78,43% e 66,66% respectivamente, menor que o do sistema.

Desta forma, a investigação de trabalhos similares teve papel fundamental na elaboração do trabalho, pois a partir dela foi possível identificar os sentimentos que seriam analisados pelo sistema proposto. Assim especificou-se que seriam identificados os sentimentos positivos, negativos e neutros. A importância dos trabalhos similares, deu-se também na escolha do método de implementação, o dicionário Léxico, que foi criado manualmente com um conjunto limitado de palavras e em seguida expandido com o uso de uma base de sinônimos e antônimos. Nesta mesma etapa, também foi possível apontar a qualidade dos trabalhos similares, servindo de apoio para especificar o resultado esperado com a implementação.

O objetivo geral do trabalho foi alcançado, uma vez que o sistema desenvolvido realiza a tarefa de análise de sentimento em textos inseridos na língua portuguesa.

Referências

Sun, S., Luo, C. and Chen, J. “A review of natural language processing techniques for opinion mining systems”, Information Fusion, China, n. 36, p. 10-25, 2016

Bird, S., Klein, E. and Loper, E. “Natural Language Processing With Python”, O’Reilly, 1 edition, 2009.

Covington, M. “Natural Language Processing for Prolog Programmers”, Pearson Education, 1 edition, 1994.

Devika, M., Sunitha, C. and Ganesh, A. “Sentiment Analysis: A Comparative Study On Different Approaches”, Procedia Computer Science, India, n. 87, p. 44-49, 2016.

Medhat, W., Hassan, A. and Korashy, H. “Sentiment analysis algorithms and applications: A Survey”, Ain Shams Engineering Journal, Egito, n. 5, p. 1093-1113, 2014.

Jurafsky, D. and Martin, J. “Speech and Language Processing: An Introduction to Natural Language Processing”, Prentice Hall, 2000.

Maletti, A. “Survey: Finite-state technology in natural language processing”, Theoretical Computer Science, Germany, n. 679, p. 2-16, 2017.

Moreo, A., Romero, M., Castro, J. and Zurita, J. “Lexicon-based Comments-oriented News Sentiment Analyzer system”, Expert Systems with Applications, Spain, n. 39, p. 9166-9180, 2012.

Kim, S. and Hovy, E. “Determining the Sentiment of Opinions”, Proceedings of the COLING conference, Geneva, 2004.