



ELVIS KESLEI DE ASSIS

**ANÁLISE DE SENTIMENTOS BASEADA EM ASPECTOS
PARA A AVALIAÇÃO DE SERVIÇOS DE
TELECOMUNICAÇÃO**

**LAVRAS – MG
2019**

ELVIS KESLEI DE ASSIS

**ANÁLISE DE SENTIMENTOS BASEADA EM ASPECTOS PARA A AVALIAÇÃO
DE SERVIÇOS DE TELECOMUNICAÇÃO**

Monografia apresentada à Universidade Federal de Lavras, como parte das exigências do Curso de Sistemas de Informação, para a obtenção do título de Bacharel.

Prof(a). Dr(a). Paula Christina Figueira Cardoso
Orientador(a)

LAVRAS – MG

2019

ELVIS KESLEI DE ASSIS

ANÁLISE DE SENTIMENTOS BASEADA EM ASPECTOS PARA A AVALIAÇÃO
DE SERVIÇOS DE TELECOMUNICAÇÃO

ANALYSIS OF FEELINGS BASED ON ASPECTS FOR THE EVALUATION OF
TELECOMMUNICATION SERVICES

Monografia apresentada à Universidade Federal de Lavras, como parte das exigências do Curso de Sistemas de Informação, para a obtenção do título de Bacharel.

APROVADA em 12 de Junho de 2019.

Dra. Renata Lopes Rosa UFLA

Dr. Jadson Castro Gertrudes UFLA

Dra. Paula Christina Figueira Cardoso

Paula Christina Figueira Cardoso
Prof(a). Dr(a). Paula Christina Figueira Cardoso
Orientador(a)

Prof(a). Paula Christina Figueira Cardoso
DCCUFLA

LAVRAS – MG

2019

RESUMO

A internet atualmente é uma das maiores fontes de dados que existem, e em especial as redes sociais tem se tornado espaços onde há uma grande criação e compartilhamento de conteúdo, como opiniões sobre produtos e ou serviços. Esses dados que são gerados diariamente se explorados de forma correta podem conter informações relevantes. A extração e o tratamento desses dados não é algo trivial, atualmente existem diversos estudos na área de recuperação da informação, Análise de sentimento que buscam solucionar essas questões. O objetivo deste trabalho é identificar na rede social Twitter as opiniões dos usuários sobre as quatro maiores empresas de telecomunicações do Brasil: Vivo, Claro, Tim e Oi. Por meio da Análise de sentimentos baseada em aspectos deseja-se classificar essas opiniões em positivo, negativo ou neutro a fim de detectar as principais insatisfações como também os aspectos positivos sobre tais empresas. Neste trabalho foi abordado a Análise de Sentimento baseado em aspectos, onde se identificou os aspectos mais relevantes através de substantivos frequentes e após isso foi feita a Análise de Sentimento sobre os tweets relacionados aos aspectos.

Palavras-chave: Análise de Sentimento, aspecto, Twitter, tweets, polaridade, sentimento.

LISTA DE ILUSTRAÇÕES

Figura 1 - POS Tagger.....	17
Gráfico 1 - Aspecto <i>internet</i> da empresa Vivo com seus respectivos sentimentos.....	23
Gráfico 2 - Aspecto <i>atendente</i> da empresa Vivo com seus respectivos sentimentos.....	24
Gráfico 3 - Aspecto <i>outros</i> da empresa Vivo com seus respectivos sentimento.....	24

LISTA DE TABELAS

Tabela 1 - Palavras positivas e negativas.....	11
Tabela 2 - Resultado da frequência de cada palavras juntamente com sua TAG.....	17
Tabela 3 - Identificação sentimento utilizando dicionário léxico.....	19
Tabela 4 - Aspectos identificados para cada operadora.....	21
Tabela 5 - Porcentagem de cobertura dos dicionários em relação às bases de dados.....	22

SUMÁRIO

1 INTRODUÇÃO.....	7
1.1 Objetivos.....	8
1.2 Organização Do Trabalho.....	8
2 CONCEITOS RELACIONADOS COM ANÁLISE DE SENTIMENTOS.....	9
3 REVISÃO BIBLIOGRÁFICA.....	13
4 PROPOSTA.....	15
4.1 Coleta De Dados.....	15
4.2 Identificação Dos Aspectos.....	16
4.3 Identificação Do Sentimento.....	18
4.3.1 Pré-Processamento.....	18
4.3.2 Técnica Utilizada Para Identificação Do Sentimento.....	19
5 RESULTADOS.....	21
5.2 Identificação Dos Aspectos.....	21
5.2 Cobertura Do Léxico Em Relação A Base De Dados.....	22
5.3 Comparação Dos Sentimentos Atribuídos Aos Aspectos.....	23
6 CONCLUSÃO.....	26
6.1 Contribuições.....	26
6.2 Trabalhos Futuros.....	26
7 REFERÊNCIAS.....	28

1 INTRODUÇÃO

Hoje em dia, com o avanço da internet, cada vez mais as pessoas têm estado conectadas, e boa parte do tempo que elas passam online é nas redes sociais, compartilhando informações, sentimentos, trocando experiências e expressando opiniões. Com isso, há a geração de uma enorme quantidade de dados que podem conter informações úteis, como a opinião dos usuários sobre determinado serviço ou produto oferecido por alguma empresa. Assim, as redes sociais podem ser usadas como um grande repositório de dados.

A partir dessa quantidade de dados gerados, conseguir extrair informações úteis pode ser muito importante para as empresas, já que as mesmas poderão saber a opinião de seus consumidores através das experiências positivas e negativas que são compartilhadas, ajudando-as a encontrar seus pontos fortes e pontos fracos. Nesse cenário, inferir o sentimento contido em uma mensagem é de vital importância a fim de entender o comportamento dos usuários e para o mercado de análise (Java et al., 2007; Kwak et al., 2010).

Extrair informações úteis que ajudem identificar pontos positivos e negativos a partir dessa enorme quantidade de dados não é uma tarefa trivial, pois é necessário analisar gírias, ícones que expressam emoções e detectar quando um usuário está sendo irônico ou sarcástico. A opinião do usuário pode ser extraída a partir da análise de sentimentos das frases coletadas.

A análise de sentimentos tem como objetivo através do processamento de linguagem natural classificar os textos como positivo, negativo ou neutro. A classificação do sentimento é comumente categorizada em duas abordagens básicas: aprendizado de máquina e baseado em léxico (Taboada et al. 2011).

Conhecendo o sentimento do usuário sobre um determinado produto ou serviço, pode ser feita uma sugestão de tópicos relacionados à opinião do usuário por meio de um sistema de recomendação. Por exemplo, sabendo da insatisfação do usuário sobre determinado serviço ou produto pode-se fazer uma recomendação a empresa relacionada com os alvos da insatisfação.

1.1 Objetivos

Existem várias redes sociais, tais como facebook, twitter, instagram. Neste trabalho foram coletadas frases do twitter relacionadas à empresas de telecomunicações no Brasil. Por meio de análise de sentimentos, objetivou-se contabilizar reclamações, insatisfações e outros relatos negativos e positivos dos consumidores em relação às quatro maiores empresas de telecomunicação e com isso gerar um sistema de monitoramento. Por meio de um sistema de monitoramento será possível através de gráficos e relatórios fazer uma comparação sobre a qualidade do serviços prestado por essas empresas como também identificar se as reclamações estão sendo atendidas.

1.2 Organização Do Trabalho

Esta trabalho está organizado da seguinte maneira:

- **Capítulo 2 - Conceitos Relacionados com Análise de Sentimentos:** este capítulo traz os principais conceitos utilizados neste trabalho que são necessários para o entendimento das atividades realizadas.
- **Capítulo 3 - Referencial Teórico:** descreve- toda a fundamentação teórica necessária para o desenvolvimento e entendimento deste trabalho, como também os trabalhos relacionados ao tema.
- **Capítulo 4 - Proposta:** este capítulo expõe os métodos e ferramentas utilizados para o desenvolvimento desta monografia.
- **Capítulo 5 - Resultados:** aqui serão apresentados os resultados obtidos a partir da análise de sentimento feito em cima das opiniões dos usuários sobre as empresas de telecomunicações.
- **Capítulo 6 - Conclusão:** neste capítulo serão apresentados as principais conclusões deste trabalho e sugestões de trabalhos futuros.

2 CONCEITOS RELACIONADOS COM ANÁLISE DE SENTIMENTOS

Neste capítulo, abordam-se alguns conceitos sobre Análise de Sentimentos (AS) que serão utilizados neste trabalho.

Na área de AS, o sentimento pode ser entendido como ato ou efeito de sentir, geralmente expresso na forma de texto (FERREIRA, 1993).

A análise de sentimento pode ser definida também como mineração de opinião, nessa área de estudos são analisadas as opiniões expressadas por pessoas em relação a um determinado assunto (LIU, 2012). Uma opinião expressa por uma pessoa pode representar uma emoção, um sentimento, uma atitude e avaliação sobre um produto, um serviço ou uma empresa.

Em um estudo feito por Pang e Lee (2008), o termo “análise de sentimentos” é descrito com o mesmo significado de mineração de opiniões, que é uma área da mineração de textos com foco, não na classificação de tópicos, mas na classificação de acordo com os sentimentos, ideias e opiniões das pessoas a respeito de um assunto.

Segundo Mukherjee (2012), análise de sentimentos consiste em uma atividade que envolve Processamento de Linguagem Natural (PLN) e Extração de Informações, e que visa obter o sentimento das pessoas que são expressados em comentários positivos ou negativos, perguntas e pedidos em documentos escritos através da análise de uma grande quantidade destes elementos. Em outras palavras, a análise de sentimentos tem por finalidade determinar a opinião de uma pessoa diante de um tópico específico ou diante da totalidade de um documento que aborda uma temática específica.

Os principais conceitos relacionados com AS são polaridade, força do sentimento, sentimento, subjetividade, objetividade, granularidade do processamento, abordagens utilizadas..

Polaridade: Representa o grau de positividade e negatividade de um texto. Normalmente, a polaridade é a saída para os métodos de AS. Alguns métodos tratam a polaridade como um resultado discreto binário (positivo ou negativo) ou ternário (positivo, negativo ou neutro).[Benevenuto, F., Ribeiro, F., & Araújo, M. (2015)] . Por exemplo, a frase, “*A qualidade da minha internet está muito ruim*”, pode ser considerado como um sentimento negativo, “*Hoje estou muito feliz, minha internet está ótima*” é uma frase com polaridade

positiva, e um sentimento neutro pode ser como na sentença “Hoje estou o dia todo na internet”.

Força do sentimento: Representa a intensidade de um sentimento ou da polaridade sendo também uma forma de saída de alguns métodos. Normalmente a força do sentimento é dada como um valor flutuante entre (-1 e 1) ou até entre $-\infty$ e $+\infty$, muitas vezes tornando necessário o uso de um threshold para identificar a neutralidade de uma sentença. Há trabalhos que por exemplo medem a força de sentimentos nos títulos das notícias como o Magnetic News [Reis et al., 2014] [Reis et al., 2015b], capaz de separar eficientemente para o usuário notícias boas de notícias ruins. [Benevenuto, F., Ribeiro, F., & Araújo, M. (2015)]

Sentimento/Emoção: Indica um sentimento específico presente em uma mensagem (ex.: raiva, surpresa, felicidade, etc.). Alguns métodos apresentam abordagens capazes de identificar qual sentimento em específico uma sentença representa. Por exemplo a abordagem léxica Emolex [Mohammad and Turney, 2013], na qual é baseada na avaliação de milhares de sentenças em inglês para 9 sentimentos diferentes: *joy, sadness, anger, fear, trust, disgust, surprise, anticipation, positive, negative*. [Benevenuto, F., Ribeiro, F., & Araújo, M. (2015)]

Subjetividade vs. Objetividade: Uma sentença objetiva possui normalmente um fato ou uma informação, enquanto sentenças subjetivas expressam sentimentos pessoais e opiniões.

A mineração de opinião feita de forma automatizada é uma alternativa viável em relação aos métodos que são realizadas de forma manual com pesquisas feitas diretamente aos usuários. Na sua grande maioria esse processo visa analisar opiniões que expressam sentimento positivo e negativos, mas em alguns casos dependendo do contexto a ser analisado seja interessante também processar informações neutras (LIU, 2012).

O processamento de uma opinião pode ser feita em diferentes níveis de granularidade, como:

- Nível de documento, onde cada documento expressa uma opinião geral sobre uma entidade.
- Nível de sentença, onde cada sentença (e.g frase, cláusula) expressa uma opinião sobre um ou vários aspectos (e.g atendimento, qualidade internet).
- Níveis de entidade e aspecto, nesse nível considera-se que cada opinião expressa por uma pessoa consiste de um sentimento (positivo, negativo) e um alvo (da opinião). Um documento pode conter opiniões sobre diversos aspectos de diversas entidades.

Algumas técnicas utilizam a análise da objetividade para estimar se compensa realizar a análise de sentimentos como apresentado em [Feldman, 2013]. Portanto, entender se um conjunto de dados possui mais sentenças objetivas ou subjetivas pode influenciar diretamente os resultados. Cabe ressaltar que textos informais (ex.: coletados de redes sociais) tendem a ser mais subjetivos que textos formais (ex.: coletados de notícias).[Benevenuto, F., Ribeiro, F., & Araújo, M. (2015)].

A análise de sentimento pode ser categorizada em duas abordagens básicas: aprendizado de máquina e baseado em léxico.

Segundo (Guimarães, 2017), léxico é o conjunto ou acervo de palavras que um determinado idioma possui. Portanto, a análise léxica estuda as unidades do vocabulário, ou seja, as palavras portadoras de sentido: substantivos, adjetivos, verbos, advérbios entre outras. Outra maneira de se analisar um texto seria a análise sintática, que se encarrega de examinar, classificar e reconhecer as estruturas da sintaxe, isto é, os períodos, as orações e os termos das orações. E por fim, a análise de um texto também pode ser feita através da análise semântica, que verifica o significado de uma ou mais sentenças. Na análise léxica, a intensidade do sentimento pode ser substituída por uma polaridade positiva, negativa ou neutra, como exemplifica a Tabela 1, construída por Pang e Lee (2008).

Tabela 1 - Palavras positivas e negativas

	Palavras
Estudante 1	positivas: brilhante, fenomenal, excelente, fantástico negativas: terrível, horroroso
Estudante 2	positivas: espetacular, legal, excelente negativas: ruim, estúpido, lerdo

Na tabela 1 é possível ver as palavras de sentimento relacionadas ao Estudante 1 e 2 e suas respectivas polaridades. Essas palavras foram identificadas em uma base de dados e se referem as opiniões sobre o estudante 1 e 2.

Segundo a definição de (MEULEMAN; SCHERER, 2013) aprendizagem de máquina, que neste contexto também pode ser tratada como reconhecimento de padrões ou mineração de dados, está relacionada à extração de padrões em grandes conjuntos de dados. Frequentemente, o objetivo é prever com precisão uma classe ou atributo contendo valores reais, com base em outros atributos e um modelo com parâmetros pré-estabelecidos.

Neste capítulo foram apresentados os principais conceitos relacionados a Análise de sentimento que serão utilizados neste trabalho.

3 REVISÃO BIBLIOGRÁFICA

Neste capítulo, são apresentados e discutidos alguns trabalhos relacionados. Diversos trabalhos mostram a importância de avaliar o sentimento das mensagens divulgadas nas redes sociais, blogs, páginas de jornais e etc.

O trabalho desenvolvido por [Oliveira 2013] consiste em analisar comentários da língua portuguesa, utilizando-se do SentiWordNet¹, que é um recurso léxico para a mineração de opinião. Neste trabalho para se utilizar desse sistema os comentários tiveram que ser traduzidos para o inglês, que é língua utilizada pelo SentiWordNet. A tradução foi feita utilizando o serviço de tradução Google Translate. As consultas foram realizadas utilizando diferentes termos sempre procurando diversificar os domínios de negócios como, marcas de produtos, locais, pessoas e etc, para fazer a avaliação da polaridade dos termos presentes nos tweets. A média geral de acertos foi de 46,4%.

Já o estudo realizado por [Martinazzo e 2010] descreve um experimento de um sistema que tem como objetivo identificar uma das seis emoções básicas (alegria, raiva, tristeza, desgosto, medo e surpresa) de notícias de páginas de jornais escritos em português do Brasil, com a utilização de um método baseado em Latent Semantic Analysis (LSA)². Como resultado, ao final do estudo se obteve uma taxa de acerto de 69% na identificação dos sentimentos.

Davidov et al. (2010) treinaram um classificador de sentimentos em tweets. Ao invés de usar especialistas para rotular manualmente as mensagens do conjunto de treinamento, foram utilizadas 50 hashtags e 15 emoticons, estruturas que aparecem naturalmente nos tweets, como os possíveis rótulos. Dessa forma, o trabalho de classificar as mensagens do conjunto de treinamento passou a ser automático, possibilitando gerar um conjunto muito maior em um tempo pequeno.

No trabalho realizado por Guimarães, 2017, foi feita uma análise léxica detalhada com 7000 tweets para determinar quais características seriam mais relevantes, como o uso de pontuação, número de caracteres, compartilhamento de mídias, assuntos, entre outros; e quais poderiam ser desconsiderados. Diferentes algoritmos de aprendizagem de máquina foram

¹<http://swn.isti.cnr.it/>

² Latent Semantic Analysis (LSA) é um método matemático/estatístico para identificação de relações entre palavras em textos.

testados em busca do melhor resultado para classificação de usuários por faixa etária. Através da pontuação das frases considerando os advérbios foi possível obter o erro máximo absoluto correspondente a 0.21, inferior aos resultados apresentados por outras ferramentas de análise de sentimentos. Para classificação dos usuários por faixa etária foi utilizada uma Rede Neural Convolutiva Profunda (DCNN) teve o melhor desempenho, atingindo uma precisão de 0.95 nos testes de validação.

O objetivo da extração de aspectos baseados em substantivos frequentes é identificar aspectos explícitos, que são representados por substantivos. O estudo realizado por Moghaddam e Ester (2010), busca identificar aspectos mais frequentes e remover substantivos irrelevantes utilizando co-ocorrências, a remoção foi baseada no número de ocorrências no base de dados, utilizando a remoção de afixos e remoção de palavras comuns. Foi identificado neste trabalho que informações adicionais disponíveis sobre aspectos conhecidos efetivamente melhora a precisão da extração de aspectos. Como resultado este trabalho obteve em média 80% de precisão na identificação dos aspectos e 86% de precisão na recuperação.

Os trabalhos acima abordaram algumas técnicas para Análise de sentimento e também para a identificação dos aspectos. Algumas dessas abordagens serão utilizadas neste trabalho como identificação de aspecto baseado em substantivos frequentes e Análise de Sentimento baseado em léxico.

4 PROPOSTA

As reclamações relacionadas às empresas de telecomunicações são muitas e das mais variadas. Para se ter uma ideia, em 2016 as reclamações protocoladas no Procon somaram 2,46 milhões, sendo que 28,8% foram sobre empresas de telefonia, segundo o Sistema Nacional de Informações de Defesa do Consumidor (Sindec) com dados divulgados pela Secretaria Nacional do Consumidor (Senacon). Diante desse cenário, entender essas insatisfações bem como identificar os pontos positivos é algo de muita relevância tanto para as empresas que prestam esses serviços como para o usuários.

Este trabalho visa identificar tais reclamações de forma automática e também relatos positivos sobre os serviços prestados pelas operadoras de telecomunicações. As empresas em estudo neste trabalho são: Vivo, Claro, Tim e Oi.

Neste capítulo são apresentadas as ferramentas e técnicas utilizadas para a captura dos dados para a construção de um dataset, o pré-processamento sobre os mesmos, a identificação e a análise de sentimentos dos tópicos considerados mais relevantes sobre quais os usuários expressaram sua opinião.

4.1 Coleta De Dados

A base de dados foi formada a partir de posts extraídos da rede social Twitter. A escolha dessa rede social foi devido a sua grande popularidade e possuir uma quantidade enorme de informações. Um tweet é um texto que contém no máximo 280 caracteres.

A obtenção inicial dos tweets para o desenvolvimento do trabalho foi feita entre os dias 07 de Abril de 2019 e 07 de Maio de 2019. Diariamente, foram coletados 2 mil tweets relacionados a cada uma das 4 empresas Vivo, Claro, Tim e OI, a base dados inicial foi composta por mais ou menos 62.000 mil tweets sobre cada empresa.

Para a coleta dos dados, foi utilizada a API³ disponibilizada pelo Twitter. Essa ferramenta possui em seu modo gratuito algumas limitações como recuperar tweets somente de sete dias antes da consulta. Contudo, para o trabalho essa limitação não foi um problema visto que foram feitas consultas todos os dias no período relatado acima. Para a recuperação dos tweets, foi utilizado o nome do perfil oficial de cada empresa na rede social. Na

³ <https://developer.twitter.com/en/docs.html>

realização da consulta são necessários a especificação de alguns parâmetros, neste trabalho foram utilizados os seguintes parâmetros:

- *until* - Especifica a data para a realização da consulta. Ex: *until=2019-04-01*
- *lang* - Especifica qual linguagem será utilizada. Neste trabalho foi utilizado Português do Brasil para a consulta. Ex: *lang=pt*.
- *tweet_mode* - Esse é um parâmetro importante para que ao realizar a coleta dos dados consiga recuperar tweets com o seu limite máximo de 280 caracteres. Ex: *tweet_mode=extended*.

A api retorna um arquivo Json⁴(JavaScript Object Notation), que é um modelo para armazenamento e transmissão de informações no formato texto, muito utilizado em aplicações Web.

4.2 Identificação Dos Aspectos

Este trabalho aborda a Análise de Sentimento baseada em aspectos, onde se busca identificar as opiniões expressas sobre os principais aspectos mencionados em tweets que têm uma relação direta com as principais empresas de telecomunicação no Brasil.

Para a identificação dos aspectos foi utilizada uma abordagem baseada em substantivos frequentes. A seguir serão descritos os procedimentos e ferramentas utilizadas neste processo.

Pré-processamento: Após a coleta de dados, foi feito um pré-processamento visando diminuir a quantidade de dados desnecessários para a identificação dos aspectos. As palavras foram padronizadas para o formato minúsculo. Foram removidos os sinais de pontuação, links e as stopwords, que são palavras sem valor semântico, como exemplo:

não, a, antes, este, isso.

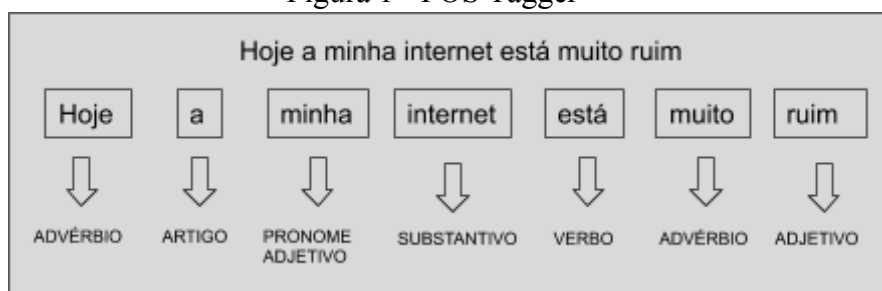
Foi utilizado um toolkit de PLN (Processamento de Linguagem Natural) para algumas dessas tarefas, a ferramenta utilizada foi o NLTK⁵.

⁴ <https://www.json.org/>

⁵ <https://www.nltk.org>

POS tagging: Nessa etapa foi realizado um procedimento etiquetar cada palavra com sua classe gramatical. Para essa etapa foi utilizado outro Toolkit de PLN o NLPNET⁶ para identificar as classes gramaticais da cada palavra na base de dados. A Figura 1 mostra o resultado do processamento de uma frase após utilizar a técnica de POS tagging.

Figura 1 - POS Tagger



Frequência de cada palavra da base: Após o processo descrito anteriormente, foi contabilizado automaticamente a frequência de cada palavra da base de dados. Primeiro foi feita a leitura de todos os arquivos da base, logo em seguida o texto foi separado em tokens de sentenças e posteriormente, em tokens de palavra. Após esse procedimento, contabilizou-se a ocorrência de cada palavra juntamente com sua TAG correspondente a sua classe gramatical. A Tabela 2 mostra o resultado dessa etapa.

Tabela 2 - Resultado da frequência de cada palavras juntamente com sua TAG

Palavra	TAG	Frequência
timbrasil	N	37468
não	ADV	13328
internet	N	6160
tim	N	5728
serviço	N	2399

Identificação dos aspectos: Nessa fase foi levantado quais seriam os possíveis aspectos. A partir de uma análise manual, optou-se por considerar somente as palavras mais frequentes da classe substantivo como candidatas a serem aspectos relevantes. Dessa forma, foram

⁶ <http://nilc.icmc.usp.br/nlpnet/>

identificados 15 possíveis aspectos para cada empresa. Desses 15 aspectos, alguns com pouca frequência e outros que poderiam ser ambíguos, por exemplo a palavra *conta*, que poder ser tanto um verbo ou um substantivo, foram removidos. Assim no final foram identificados sete aspectos para cada operadora.

4.3 Identificação Do Sentimento

Essa etapa consiste em em determinar a polaridade de um tweet de acordo com o seu aspecto em positivo, negativo ou neutro.

4.3.1 Pré-Processamento

Considerando os dados originais, antes de realizar identificação dos aspectos e análise de sentimento, foi realizado um pré-processamento a fim de descartar dados irrelevantes para a classificação. As fases do pré-processamento realizadas nesta etapa foram as seguintes:

- Remoção de pontuação, visto que pontuação não tem nenhuma influência sobre a análise de sentimento.
- Remoção de link, uma vez que esses termos não tem valor semântico.
- Conversão de todas as letras para minúscula e remoção de acentos com o objetivo de padronizar o texto.

Nesta fase não foi feito a remoção de stopwords visto que algumas palavras que são consideradas sem carga semântica se removidas poderiam influenciar na análise, como por exemplo a palavra “*não*”.

Antes de identificar a polaridade de cada tweet, foi necessário identificar se um determinado tweet continha algum dos aspectos que foram identificados na seção 4.2. Para isso, foi feita uma varredura em cada palavra de um tweet e ao encontrar a primeira ocorrência de alguns dos aspectos selecionados, o tweet era considerado ser daquele aspecto. Optou-se por essa abordagem visto que devido a limitação de caracteres do twitter, as pessoas tendem geralmente a falar somente sobre um aspecto. Em alguns casos pode ocorrer de um

mesmo tweet falar sobre mais de um aspecto, mas para o escopo deste trabalho foi assumido que um tweet trata de um único aspecto.

Após identificar o aspecto de cada tweet, foi feita a classificação da polaridade do sentimento de cada tweet. A seguir serão descritos as fases desse processo.

4.3.2 Técnica Utilizada Para Identificação Do Sentimento

A abordagem utilizada para identificação dos sentimentos relacionados a cada tweet foi baseada em dicionário léxico. Um dicionário de palavras contém sentimentos anotados com sua respectiva polaridade, podendo ser positivo (1), negativo (-1) ou neutro (0). Neste trabalho foram utilizados dois léxicos disponíveis para a língua portuguesa: o SentiLex-PT2⁷ e o OpenLexico⁸.

O SentiLex-PT2 é composto por 7.014 lemas e 82.347 formas flexionadas. O OpenLexico, por sua vez, é formado por 32191 palavras. Com o objetivo de se obter uma maior cobertura das palavras de sentimento em relação a base de dados, foi utilizado uma combinação dos dois dicionários.

Para realizar a análise de sentimento, cada tweet foi segmentado em token de palavras com o auxílio do NLTK. Após esse processo, verificou-se a ocorrência de cada palavra nesse novo dicionário que foi gerado a partir da união dos léxicos SentiLex-PT2 e o OpenLexico. Caso a palavra exista no dicionário, é retornado o seu respectivo sentimento; caso a palavra não conste no dicionário, é retornado o valor 0 que corresponde a polaridade neutra. Cada sentimento associado a palavra retornada foi armazenado em uma lista e logo após foi feito o somatório do sentimento de todas as palavras para determinar a polaridade do tweet. A Tabela 3 mostra um exemplo de como é feito o cálculo de polaridade sobre cada tweeter.

Tabela 3 - Identificação sentimento utilizando dicionário léxico

Tweet	Polaridade de cada palavra	Soma	Polaridade do tweet
a minha internet está muito ruim	0, 0, 0, 0, 0, 1	-1	Negativo

⁷ <https://b2share.eudat.eu/records/93ab120efdaa4662baec6adee8e7585f>

⁸ <http://ontolp.inf.pucrs.br/Recursos/downloads-OpLexicon.php>

Após feita a análise sobre todos os tweets e seus respectivos aspectos, foi possível identificar de forma sucinta as opiniões dos usuários sobre as empresas de telecomunicações. No próximo capítulo são apresentados os resultados e algumas limitações dessa abordagem.

5 RESULTADOS

Este capítulo tem como objetivo analisar os resultados obtidos através da Análise de Sentimento baseado em aspecto aplicados às quatro empresas de telecomunicação presentes no estudo.

5.2 Identificação Dos Aspectos

Nesta seção será abordado os resultados sobre a identificação dos aspectos.

Após todos os procedimentos descritos na seção 4.2 foram identificados 7 aspectos mais relevantes para cada operadora, a identificação foi feita de forma manual, utilizando a abordagem de substantivos mais frequentes. A tabela 4 apresenta estes aspectos selecionados para cada operadora.

Tabela 4 - Aspectos identificados para cada operadora

Operadora	Aspecto						
Vivo	internet	atendimento	sinal	serviço	atendente	plano	contato
Claro	internet	atendimento	sinal	serviço	contato	plano	ligação
Tim	internet	atendimento	sinal	serviço	atendente	plano	ligação
Oi	internet	atendimento	sinal	serviço	atendente	suporte	plano

Esse procedimento de identificação de aspectos através de substantivos mais frequentes é uma abordagem mais simples que pode não contemplar todos os aspectos relevantes ao domínio do problema. Diante disso como trabalhos futuros pretende-se melhorar este método para tornando-o automatizado a fim de aumentar a precisão com os aspectos selecionados.

5.2 Cobertura Do Léxico Em Relação A Base De Dados

Nesta seção será apresentado a cobertura de cada léxico de forma individual como também a combinação dos dois dicionários em relação a base de dados de tweets de cada empresa em estudo.

A tabela 5 apresenta o resultado da cobertura do dicionário sobre cada operadora.

Tabela 5 - Porcentagem de cobertura dos dicionários em relação às bases de dados

Empresa / Léxico	OpenLexico	SentLex	OpenLexico + SentLex
Vivo	8.8%	8.8%	11.8%
Claro	8.6%	8%	11.7%
Tim	10.5%	9.7%	14.13%
Oi	9%	7.9%	11.8%

Após a união dos dois dicionários pode-se observar uma melhora na cobertura em relação às palavras da base. Essa melhora mesmo sendo pequena tem influência na Análise de sentimento visto que palavras que existam na base de dados mais não conste no dicionário serão consideradas com polaridade neutra.

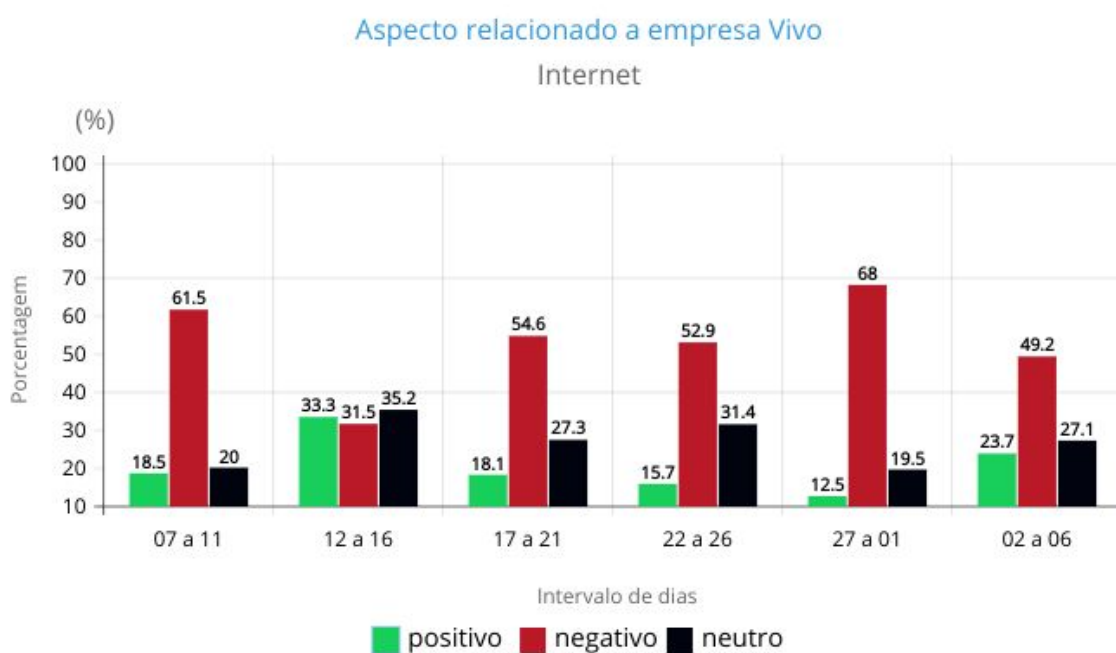
A análise de sentimento baseado em léxico apresenta algumas limitações, como pode ser observado na tabela 3. As palavras de sentimento contidas nesses dicionários muitas vezes não estão contidas na base analisada. Uma alternativa seria montar um base de dados anotada manualmente e utilizar aprendizado de máquina para realizar a atribuição de sentimento a cada tweet. Para este trabalho este método foi satisfatório visto que cada tweet tem uma limitação de 280 caracteres, sendo assim quando há a ocorrência de palavras de sentimento com polaridade positivo, negativo ou neutro em um post esta polaridade tende a ser o sentimento real do tweet.

5.3 Comparação Dos Sentimentos Atribuídos Aos Aspectos

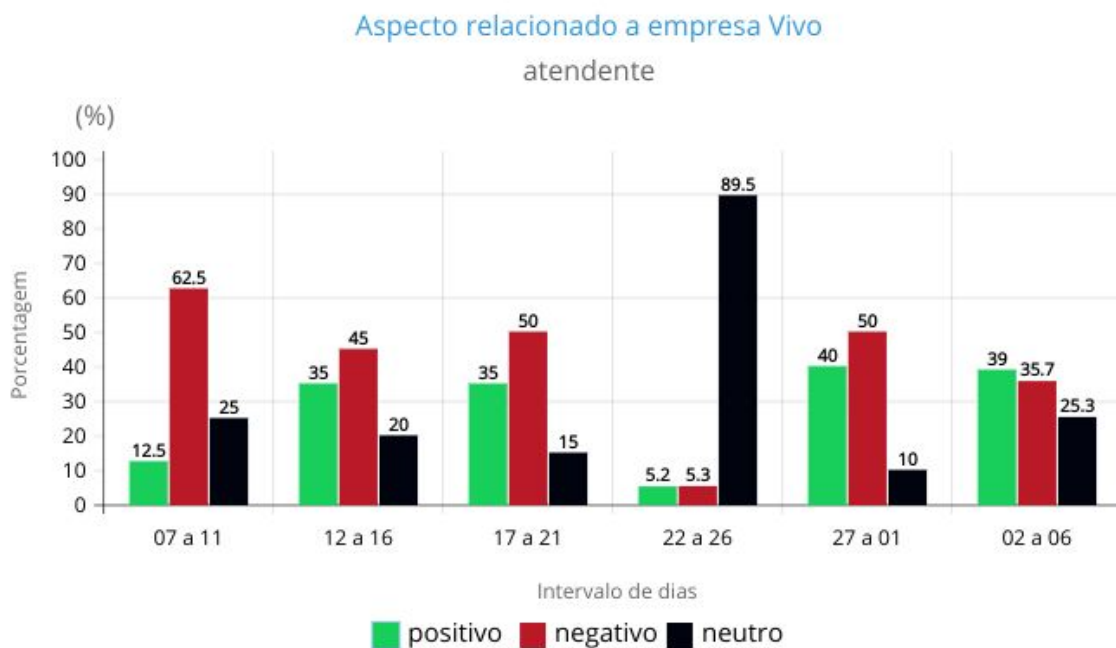
O Objetivo desta seção é mostrar graficamente a relação dos sentimentos positivos, negativos e neutros associados a alguns aspectos em uma timeline de 5 em 5 dias, os intervalos correspondem do dia 07 de Abril de 2019 a 06 de Maio de 2019.

O gráfico 1 representa o aspecto internet relacionado a empresa Vivo com seus respectivos sentimentos. Observa-se há mais negativos do que positivos.

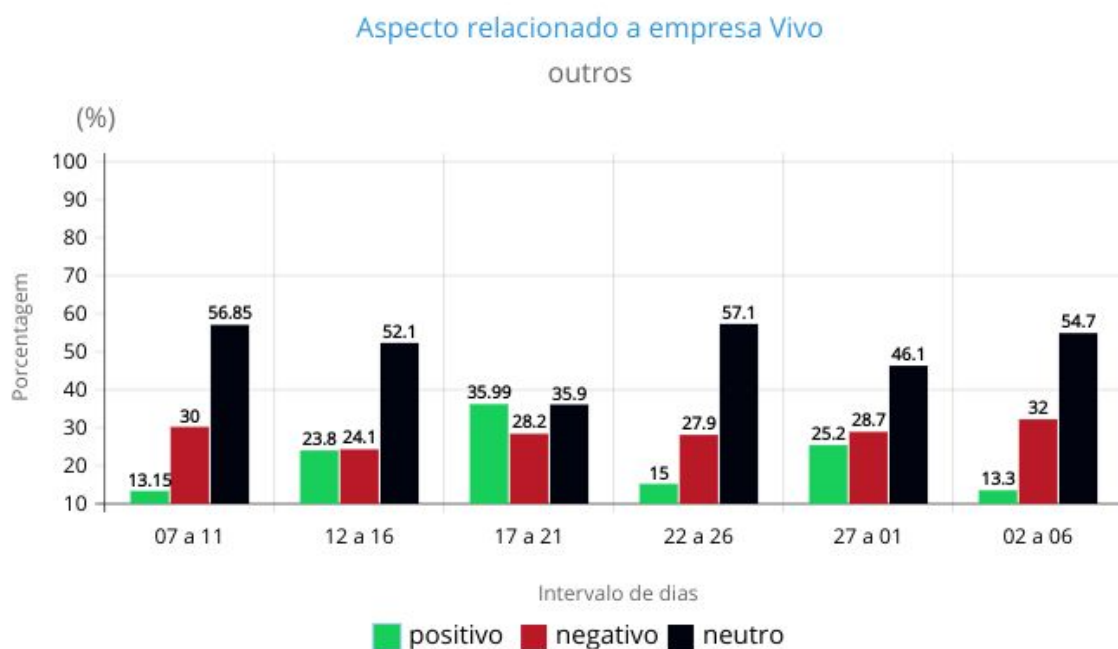
Gráfico 1 - Aspecto *internet* da empresa Vivo com seus respectivos sentimentos



O gráfico 2 mostra mais um exemplo da detecção de polaridade sobre os aspectos. Neste gráfico é possível ver que no intervalo de dias 22 a 26, 89.5% dos tweets tiveram polaridade neutra, isso implica em uma das limitações da utilização do léxico para Análise de Sentimento, pois há situações em que em um mesmo tweet há quantidades de palavras positivas e negativa iguais o pode levar um sentimento neutro.

Gráfico 2 - Aspecto *atendente* da empresa Vivo com seus respectivos sentimentos

Para os casos onde não foi possível identificar o aspecto ou o mesmo não estava contido em um tweet, este foi rotulado como *outros*, para esses casos somente foi feita a atribuição de polaridade sobre a sentença. O gráfico 3 representa tweets rotulados como *outros*. Como é possível ver quando não se tem um aspecto identificado a maioria dos tweets tem polaridade neutra, isso mostra que os sentimentos mais relevantes, positivos e negativos, estão distribuídos nos tweets que contém algum dos aspectos identificados.

Gráfico 3 - Aspecto *outros* da empresa Vivo com seus respectivos sentimento

O método utilizado baseado em léxico é um método que apresenta algumas limitações que estão relacionadas a cobertura do dicionário em relação às palavras, nessa abordagem o ideal se tenha palavras de sentimento mais específicas relacionados ao domínio em questão. Mesmo com essas limitações a abordagem de Análise de sentimento baseado em aspecto deixou evidente a potencialidade que as informações contidas nas redes sociais, em particular neste trabalho no Twitter, tem para que as empresas possam avaliar a qualidade a dos serviços e ou produtos de forma automática e assim focar em melhorias elevando a satisfação de seus clientes. como também os usuários poderem estar cientes sobre quais são os os pontos fortes e fracos de cada empresa em relação a seus qualidade de seus serviços ou produtos.

6 CONCLUSÃO

Neste trabalho, foi apresentado uma abordagem de Análise de Sentimento baseado em aspectos com o objetivo de identificar na rede social Twitter as principais insatisfações e também pontos positivos com relação as 4 maiores empresas de telecomunicações do Brasil, Vivo, Claro, Tim e OI. Para isso foram utilizados alguns recursos de PLN (Processamento de Linguagem Natural) e também através da consulta na literatura sobre o tema existente para a língua portuguesa do Brasil, foram aplicados tanto para identificação dos aspectos com para a Análise de sentimento técnicas e recursos já disponíveis, como a identificação de aspectos baseados em substantivos frequentes quando a abordagem léxica para a Análise de Sentimento.

6.1 Contribuições

Como contribuição deste trabalho pode se destacar a abordagem utilizada combinando dois léxicos para a Análise de sentimento e com isso uma melhora na classificação das polaridades sobre as opiniões dos usuários.

Com relação a abordagem utilizada, este trabalho contribui mostrando a relevância que a identificação de Sentimentos baseados em aspectos tem para as empresas, visto que essa grande quantidade de dados que são gerados todos os dias tem um grande valor desde que explorados de forma correta.

6.2 Trabalhos Futuros

Neste trabalho a identificação dos aspectos foi feito baseado em substantivos frequentes é após a identificação dos possíveis aspectos com base em sua frequência na base dados, foi selecionado de forma manual levando em consideração alguns critérios como detalhado na seção 4.2. Assim como trabalho futuro, pretende-se utilizar para extração dos aspectos aprendizagem supervisionado, tornando assim a identificação dos aspectos automatizada e com maior precisão.

Para a Análise de Sentimento foi utilizado o método baseado em léxico que também contém que também tem algumas limitações quanto a precisão, cobertura do dicionário em relação às palavras da base, assim para continuação deste trabalho pretende-se construir uma base rotulada para cada uma das empresas em estudo e com isso utilizar o método de aprendizagem supervisionada para a classificação de polaridade dos tweets, assim melhorar a precisão na identificação do sentimentos sobre a opinião do clientes, entregando melhores resultados seguindo a proposta do trabalho, que é identificar as principais insatisfações e também pontos positivos sobre a opinião dos usuários, e assim com base nessas informações as empresas poderão melhorar seus produtos e ou serviços e também os clientes terão mais ciência sobre os principais problemas relatados sobre essas empresas.

7 REFERÊNCIAS

PANG, B.; LEE, L. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, Now Publishers Inc, Hanover, MA, USA, v. 2, n. 1-2, p. 1–135, Out 2008. Disponível em: .

FERREIRA, A. d. H. Dicionário aurélio eletrônico. [S.l.]: Ed. Nova Fronteira, 1993.

Subhabrata Mukherjee. Sentiment analysis - a literature survey, June 2012. Indian Institute of Technology, Bombay. Roll No: 10305061

Benevenuto, F., Ribeiro, F., & Araújo, M. (2015). Métodos para Análise de Sentimentos em mídias sociais.

Esuli and Sebastiani. Sentwordnet: A publicly available lexical resource for opinion mining. In *In Conference on Language Resources and Evaluation*, 2006.

MARTINAZZO, Barbara; PARAISO, Emerson Cabrera. Identificação de Emoções em Notícias Curtas. In: *CLEI-Conferência Latino-Americana de Informática*. 2010. p. 1-10.

DAVIDOV, Dmitry; TSUR, Oren; RAPPOPORT, Ari. Enhanced sentiment learning using twitter hashtags and smileys. In: *Proceedings of the 23rd international conference on computational linguistics: posters*. Association for Computational Linguistics, 2010. p. 241-249.

TABOADA, Maite et al. Lexicon-based methods for sentiment analysis. *Computational linguistics*, v. 37, n. 2, p. 267-307, 2011.

LIU, B. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, Morgan & Claypool Publishers, v. 5, n. 1, p. 1–167, 2012.

MOGHADDAM, S.; ESTER, M. Opinion digger: an unsupervised opinion miner from

unstructured product reviews. In: CIKM. [S.l.: s.n.], 2010. p. 1825–1828.

SCHERER, Klaus R.; MEULEMAN, Ben. Human emotion experiences can be predicted on theoretical grounds: evidence from verbal labeling. **PloS one**, v. 8, n. 3, p. e58166, 2013.