

CENTRO UNIVERSITÁRIO ALVES FARIA - UNIALFA
PROGRAMA DE PÓS-GRADUAÇÃO EM ADMINISTRAÇÃO
MESTRADO PROFISSIONAL EM ADMININSTRACÃO

ANDRÉ DE OLIVEIRA FIALHO

**ANÁLISE DA EFICIÊNCIA DE DOIS MODELOS DE PROCESSAMENTO DE
LINGUAGEM NATURAL (PLN) QUE CLASSIFICAM O SENTIMENTO DO
CONSUMIDOR DE UMA INSTITUIÇÃO FINANCEIRA**

Goiânia
2021

ANDRÉ DE OLIVEIRA FIALHO

**ANÁLISE DA EFICIÊNCIA DE DOIS MODELOS DE PROCESSAMENTO DE
LINGUAGEM NATURAL (PLN) QUE CLASSIFICAM O SENTIMENTO DO
CONSUMIDOR DE UMA INSTITUIÇÃO FINANCEIRA**

Dissertação apresentada ao Programa de Pós-Graduação em Administração do Centro Universitário Alves Faria, como requisito parcial para a obtenção do título de Mestre em Administração.

Goiânia, 19 de março de 2021

Aprovada em: ____/____/____.

BANCA EXAMINADORA

Prof. Dr. Bento Alves da Costa Filho - UNIALFA (Orientador)

Prof. Dr. Paulo Cesar Bontempo - UNIALFA (Leitor)

Prof. Dr. Ricardo Limongi - UFG (Leitor)

Catálogo na fonte: Biblioteca UNIALFA**F438a****Fialho, André de Oliveira**

Análise da eficiência de dois modelos de processamento de linguagem natural (PLN) que classificam o sentimento do consumidor de uma instituição financeira / André de Oliveira Fialho. – 2021.

73 f. : il.

Orientador: Prof. Dr. Bento Alves da Costa Filho.

Dissertação (mestrado) – Centro Universitário Alves Faria (UNIALFA) - Mestrado em Administração – Goiânia, 2021.

1. Emoções. 2. Consumidores. 3. Análise de sentimentos. I.

Fialho, André de Oliveira. II. UNIALFA – Centro Universitário Alves Faria. III. Título.

CDU: 004.78+159.942.4

ANDRÉ DE OLIVEIRA FIALHO

**ANÁLISE DA EFICIÊNCIA DE DOIS MODELOS DE PROCESSAMENTO DE
LINGUAGEM NATURAL (PLN) QUE CLASSIFICAM O SENTIMENTO DO
CONSUMIDOR DE UMA INSTITUIÇÃO FINANCEIRA**

Dissertação apresentada ao Programa de Pós-Graduação em Administração do Centro Universitário Alves Faria, como requisito parcial para a obtenção do título de Mestre em Administração.

Orientador: Prof. Dr. Bento Alves da Costa Filho

**Goiânia – GO
PPGADM - UNIALFA
Março / 2021**

SUMÁRIO

Resumo	10
Abstract	10
1. Introdução	11
1.1 A abordagem do Tema escolhido	11
1.2 Pergunta – Problema	13
1.3 Objetivos da Pesquisa	14
1.3.1 Objetivo Geral	14
1.3.2 Objetivos Específicos.....	14
1.3.2.1 Investigar se a escolha de características de sentimentos definida no modelo está aderente ao resultado pelo qual o este foi criado	14
1.3.2.2 Comparar as acurácias obtidas dos dois modelos analíticos PLN	15
1.3.2.3 Comparar os índices de concordância entre a avaliação humana e o modelo analítico	15
1.3.2.4 Analisar Escalas de classificação de sentimento do consumidor dos modelos analíticos	15
1.4 Justificativa	16
2. Fundamentação Teórica	18
2.1 A evolução da avaliação do sentimento humano em textos de grande escala	18
2.2 A análise de sentimento no contexto das mídias sociais.....	19
2.3 Os desafios de analisar o sentimento humano nas mídias sociais on-line	20
2.4 Inter-Observador para validar um modelo analítico	23
2.5 Estudos Anteriores	24
3. Metodologia	26
3.1 Quanto a abordagem	26
3.3 Quanto aos objetivos.....	28
3.4 Word2Vec	28
3.5 BERT (Bidirectional Encoder Representations from Transformers).....	31
3.6 Alfa de Cronbach	33
3.7 Cohen Kappa, Fleiss kappa – k e Coeficiente de Kappa ponderado – kw	34
4. Resultados e Discussões	36

4.1 Alfa de Cronbach em Word2Vec e BERT.....	36
4.2 Cohen <i>Kappa</i> – <i>k</i>	37
4.2.1 Resultados de Cohen Kappa - <i>k</i> em Word2VEC	37
4.2.2 Resultados de Cohen Kappa - <i>k</i> em BERT	40
4.3 Fleiss Kappa.....	43
4.3.1 Resultados de Fleiss kappa em Word2Vec.....	44
4.3.2 Resultados de Fleiss kappa em BERT	45
4.4 Kappa Ponderado – <i>Kw</i>	46
4.4.1 Resultados de Kappa Ponderado – <i>kw</i> em Word2Vec.....	47
4.4.2 Resultados de Kappa Ponderado – <i>kw</i> em BERT	50
4.5 Investigações das causas de discordâncias	54
4.5.1 Análise da concordância utilizando apenas as características individuais em Word2Vec	55
4.5.1.1 Exemplos de posts avaliados	58
4.5.2 Análise da concordância utilizando apenas as características individuais em BERT	60
4.5.2.1 Exemplos de posts avaliados	64
5. Conclusão.....	66
6. Referências Bibliográficas.....	69

ÍNDICE DE TABELAS

Tabela 1 – Análise de Sentimentos: Resumo de Conceitos Fundamentais	25
Tabela 2 – Estatísticas de confiabilidade	36
Tabela 3 – Matriz quadrada de frequências obtidas entre Ferramenta x Profissional A	38
Tabela 4 – Matriz quadrada de frequências obtidas entre Ferramenta x Profissional B.....	38
Tabela 5 – Matriz quadrada de frequências obtidas entre Profissional A x Profissional B.....	38
Tabela 6 – Matriz quadrada de frequências esperadas entre Ferramenta x Profissional A	40
Tabela 7 – Matriz quadrada de frequências esperadas entre Ferramenta x Profissional B.....	41
Tabela 8 – Matriz quadrada de frequências esperadas entre Ferramenta x Profissional B.....	41
Tabela 9 – Cálculo integral Fleiss Kappa – Word2Vec.....	44
Tabela 10 – Fleiss Kappa por Categorias individuais em Word2Vec	45
Tabela 11 - Cálculo integral Fleiss Kappa – BERT.....	45
Tabela 12 – Fleiss Kappa por Categorias individuais em BERT.....	46
Tabela 13 – Matriz de peso linear de discordâncias – Word2Vec.....	47
Tabela 14 – Matriz quadrada de frequências esperadas entre Ferramenta x Profissional A ...	48
Tabela 15 – Matriz quadrada de frequências esperadas entre Ferramenta x Profissional B....	48
Tabela 16 – Matriz quadrada de frequências esperadas entre Profissional A x Profissional B	49
Tabela 17 – Cálculo do Kappa Ponderado – kw.....	49
Tabela 18 - Matriz de peso linear de discordâncias – BERT.....	51
Tabela 19 – Matriz quadrada de frequências esperadas entre Profissional A x Profissional B	52
Tabela 20 – Matriz quadrada de frequências esperadas entre Profissional A x Profissional B	53
Tabela 21 – Matriz quadrada de frequências esperadas entre Profissional A x Profissional B	53
Tabela 22 – Cálculo do Kappa Ponderado – kw.....	54
Tabela 23 – Contingência Observada entre Ferramenta x Profissional A	55
Tabela 24 – Contingência Observada entre Ferramenta x Profissional B	55
Tabela 25 – Contingência Observada entre Profissional A x Profissional B	55
Tabela 26 – Cálculo do Kappa Ponderado – Características Primárias – kw	58
Tabela 27 – Exemplos de disparidade na classificação de sentimentos	58
Tabela 28 – Exemplos de consenso na classificação de sentimentos	59
Tabela 29 – Contingência Observada entre Ferramenta x Profissional A	60
Tabela 30 – Contingência Observada entre Ferramenta x Profissional B	60
Tabela 31 – Contingência Observada entre Ferramenta x Profissional A	61

Tabela 32 - Cálculo do Kappa Ponderado – Características Primárias – kw.....	63
Tabela 33 – Exemplos de disparidade na classificação de sentimentos	64
Tabela 34 – Exemplos de consenso na classificação de sentimentos	65

ÍNDICE DE FIGURAS

Figura 1 – Fontes de vulnerabilidade no estudo do sentimento captado nos textos das Redes Sociais On-line	23
Figura 2 – Formato dos relatos das interações entregues aos avaliadores	27
Figura 3 – Polaridade das características primárias em Word2Vec	56
Figura 4 – Polaridade das características primárias em BERT	61

ÍNDICE DE GRÁFICOS

Gráfico 1 – Concordâncias nas polaridades Negativo – Neutro – Positivo em Word2Vec	56
Gráfico 2 – Distribuição de Concordâncias das Características de Sentimento e kw em Word2Vec	57
Gráfico 3 – Concordâncias nas polaridades Negativo – Neutro – Positivo em BERT	62
Gráfico 5 – Distribuição de Concordâncias das Características de Sentimento e kw em BERT	62

Resumo

As opiniões coletadas nas plataformas de mídias sociais têm cativado a atenção dos acadêmicos e profissionais de diversas áreas, face à riqueza que a análise dessas conversas proporciona em insights obtidos dos sentimentos dos consumidores e estes exercem uma forte influência nas marcas que consomem. Por outro lado, o mundo corporativo enfrenta o desafio de se adaptar a essa nova dinâmica de consumo e se aproximar do seu público-alvo. A revisão da literatura demonstrou os grandes desafios da evolução dessas ferramentas analíticas quando é medida a acurácia da atividade de monitoramento de sentimentos e classificação do viés de *posts* que podem conter inúmeras emoções dentro de uma única frase. Mas destaca-se também na literatura fundamentações que demonstram ferramentas que estão em constante evolução de aprendizagem capaz obter avaliações preditivas e sua performance está diretamente relacionada à sua seleção, a criação de códigos (programação) e a classificação de dados. Este trabalho teve por objetivo investigar o nível de eficiência de dois frameworks de PNL, *Word2Vec* e *BERT* utilizados por uma instituição financeira para realizar a análise de sentimentos em seis plataformas de mídia social. Esta pesquisa realizou uma comparação de concordância entre dois avaliadores e os modelos PNL utilizados nela, fazendo uso dos indicadores estatísticos de concordância *Cohen Kappa*, *Fleiss kappa – k* e *Coefficiente de Kappa ponderado – kw*. Os resultados demonstraram um baixo índice de concordância entre os modelos analíticos e os avaliadores quando utilizado uma escala longa de características de sentimentos avaliáveis o que pode implicar falhas na obtenção de insights aos gestores dessa instituição.

Palavras-Chaves: Emoções, Consumidores, Análise de Sentimentos, Concordância, Indicadores Estatísticos.

Abstract

The opinions collected on social media platforms have captivated the attention of academics and professionals from different areas, given the richness that the analysis of these conversations provides in insights obtained from consumers' feelings and they have a strong influence on the brands they consume. On the other hand, the corporate world faces the challenge of adapting to this new dynamic of consumption and getting closer to its target audience. The literature review demonstrated the great challenges of the evolution of these analytical tools when measuring the accuracy of the activity of monitoring feelings and classifying the bias of posts that can contain countless emotions within a single sentence. But there are also foundations in the literature that demonstrate tools that are in constant evolution of learning capable of obtaining predictive assessments and their performance is directly related to their selection, the creation of codes (programming) and the classification of data. This work aimed to investigate the efficiency level of two NLP frameworks, *Word2Vec* and *BERT* used by a financial institution to carry out sentiment analysis on six social media platforms. This research carried out a comparison of agreement between two evaluators and the NLP models used in it, using the statistical indicators of agreement *Cohen Kappa*, *Fleiss kappa - k* and *weighted Kappa coefficient - kw*. The results demonstrated a low level of agreement between the analytical models and the evaluators when using a long scale of characteristics of evaluable feelings, which may imply failures in obtaining insights from the managers of this institution.

Keywords: Emotions, Consumers, Sentiment Analysis, Concordance, Statistical Indicators.

1. Introdução

Este capítulo apresenta de forma sintética a importância de se obter a classificação de sentimentos dos consumidores, como uma estratégia gerencial que proporciona um olhar amplo sobre as reputações das marcas, contribuindo para as empresas melhorarem o relacionamento com o cliente, além elevar a qualidade da gestão de crises de imagem. Assim iniciamos a seção 1.1 relatando qual foi a motivação da escolha do tema, trazendo algumas etapas de extração de dados em mídias sociais e a importância da construção de modelos analíticos que classificam sentimentos do consumidor emitidas em mensagens de textos. Nesta sessão também apresentamos a pergunta e os objetivos da pesquisa.

1.1 A abordagem do Tema escolhido

Presente em diversas áreas de estudo, como psicologia, medicina, marketing, administração, estatística e computação, os dados estão sendo coletados e acumulados em ritmo exponencial por meio de plataformas de rede de computadores. Assim faz-se necessário a utilização de ferramentas computacionais que permitam a captura de texto em linguagem natural, para um posterior uso da mineração de opinião, que permitirá que sejam extraídos dados relevantes de forma automatizada para geração de insights para eventuais decisões gerenciais.

Segundo Liu (2015), a mineração de opinião, é o estudo computacional das opiniões, sentimentos, atitudes e emoções das pessoas.

Conforme Iglesia e Tsunoda (2020), a mineração de opinião é dirigida principalmente a opiniões coletadas principalmente em mídias sociais, que exprimem sentimentos positivos ou negativos, considerando também expressões que não denotam nenhum sentimento, as denominadas expressões neutras.

Essas teorias e ferramentas são objeto de campo emergente de descoberta de conhecimento em bancos de dados, sendo a análise do sentimento expressada nessas mídias sociais, uma delas, inserida na vasta área de ciências de dados.

Dentre estas ferramentas, destaca-se a mineração de textos, que é a aplicação de algoritmos específicos para extrair padrões de textos que no caso desta pesquisa são as vozes obtidas em 6 plataformas de relacionamento com o consumidor de uma instituição financeira (FaleCom, Facebook, Twitter, Whatsapp, Linkedin e Instagram). A distinção entre a descoberta de conhecimento em bancos de dados e a etapa de mineração de dados (dentro do

processo) é fundamental para definir a preparação, seleção, limpeza de dados e incorporação de conhecimentos (Fayyad et al., 1996).

Assim, Resende et al (2012), esclarece que minerar de textos é um processo que busca descobrir conhecimento útil a partir de coleções textuais que correspondem a padrões, o que pode inclusive facilitar a análise exploratória de mídias sociais. Salustiano (2015), relata que nada é padronizado, e fica sob a responsabilidade de cada empresa, agência ou ferramenta definir a metodologia de trabalho. Na prática o que se vê, são métodos de classificação de sentimentos que têm sido amplamente utilizados para o desenvolvimento de ferramentas sem uma compreensão de sua aplicabilidade no contexto das redes sociais, ou suas vantagens, desvantagens e limitações.

Por isso quando se investiga a análise de sentimentos em plataformas *on-line*, percebe-se um grande embate entre os profissionais de diversas áreas envolvidas: cientista de dados, profissionais de *analytics*, psicoevolucionários, estatísticos (Salustiano, 2016). Porém cada área pode advogar de forma tendenciosa sua perspectiva profissional, sem considerar as vantagens ou limitações que cada um, homem ou máquina, tem na análise dos sentimentos humanos. Profissionais de Psicologia podem, por exemplo, defender que análise humana de sentimentos é mais assertiva que análise automáticas realizadas por algoritmos, o que seria rapidamente refutado por um cientista de dados, que possui em seu bojo um grande volume de dados que inviabilizaria a análise ser realizada por humanos.

Um estudo realizado pela MindMiners (2017) com uma amostra de 1.000 consumidores no Brasil aponta que 54% destes seguem os perfis de suas marcas preferidas nas redes sociais, o que demonstra o quão relevante é possuir ferramentas de monitoramento de sentimentos eficientes e confiáveis.

A mudança de comportamento e os novos hábitos de consumo dela decorrentes faz com que o monitoramento de comunidades e serviços com base em plataformas Web se torne um poderoso instrumento de opinião e poder, já que essas plataformas são estruturas abertas capazes de se expandir de forma ilimitada, integrando atores diversos que compartilham o mesmo código de comunicação (CASTELLS, 1999). Considerando que a experiência de compra de um serviço financeiro reflete algumas características subjetivas e implícitas a este tipo de atividade, não pode ser ignorado que as emoções emanadas das plataformas influenciam de forma perene a maneira pela qual o serviço é consumido ou avaliado (AMARAL et al., 2010). A eficiência desta relação comercial entre os consumidores e a instituição financeira está diretamente relacionada à capacidade de o serviço evocar e provocar reações subjetivas (NETO, 2019).

1.2 Pergunta – Problema

A incerteza em relação à acurácia ou a fiabilidade desses modelos analíticos para percepção de sentimentos é a problemática desta pesquisa, visto que a construção autônoma e não padronizada que a instituição financeira adota para desenvolver estes modelos, podem torná-los inadequados como fonte confiável de insights para tomada de decisões.

Muitos desses modelos analíticos foram propostos para frases completas, não para mensagens curtas em tempo real, que atualmente podem conter apenas um “emoji” ou foto, ou para cada particularidade que pode servir como um entrave no monitoramento (CANHOTO E PADMANABBAN, 2015).

Em relação à acurácia dos modelos e ferramentas analíticas para classificação de sentimentos, encontra-se na literatura comercial dos desenvolvedores das ferramentas, que a acurácia calculada por algoritmos de *machine learning* figura entre 60 e 85% (LEUNG, 2014). Porém não fica claro como a metodologia e os coeficientes de concordância e validação de treinamentos e testes foram executados no modelo construído pela instituição financeira. Essa situação se explica, visto que a instituição não abre a metodologia e etapas de *machine learning* desenvolvida pela área técnica em razão de estratégia de privacidade competitiva diante dos seus concorrentes.

Assim, este estudo investiga a seguinte pergunta da pesquisa: os modelos analíticos de Processamento de Linguagem Natural (PNL) - *Word2Vec* e *BERT*, possuem níveis de acurácia satisfatórios quando comparados com a análise de concordância humana (inter-observadores)? Estes modelos são os dois últimos que estão sendo utilizados pela Instituição Financeira em estudo, com o objetivo classificar as características de sentimentos dos consumidores que interagem com a empresa por meio de mídias sociais oficiais, visando oferecer informações gerenciais das vozes destes consumidores a área estratégica da instituição. Não foi encontrado no referencial teórico em língua portuguesa, uma pesquisa comparativa de modelos analíticos PLN com a classificação de concordância inter-observadores (humanos). Encontramos essa comparação em literatura internacional, porém nenhum delas fazia referência a área financeira.

Esta metodologia é comumente utilizada em ciências médicas, inclusive psicologia, área correlata ao Marketing, e pode ser interpretado como até que ponto a quantidade de concordância observada entre os avaliadores excede o que seria esperado se todos os avaliadores fizessem suas classificações de forma completamente aleatória (Fleiss, 1971, p. 378)

1.3 Objetivos da Pesquisa

Depois de formular a pergunta problema, propõe-se o objetivo geral da pesquisa que se desdobrará em objetivos específicos, listados abaixo.

1.3.1 Objetivo Geral

O objetivo geral desta pesquisa é disponibilizar aos gestores da instituição financeira uma percepção confiável das ferramentas analíticas geradoras de insights. Para isso foi utilizada uma comparação com classificadores humanos além dos testes realizados em *machine learning*. Assim foi comparado os índices de concordância inter-observadores humanos os dois modelos PLN, que avaliam de forma automatizada sentimentos em milhões de frases geradas pelos clientes durante sua interação diária com a instituição financeira, por meio de seis plataformas de mídias sociais já relatadas. A avaliação de sentimentos realizada pelos humanos, pode ser apresentada aos gestores como a referência “padrão-ouro” (*gold standard*) para validação de testes de modelos de monitoramento de sentimentos e que está ligada a área de Marketing.

1.3.2 Objetivos Específicos

A partir do objetivo geral definido, foram elencados para esta pesquisa os seguintes objetivos específicos a seguir.

1.3.2.1 Investigar se a escolha de características de sentimentos definida nos modelos estão aderentes ao resultado pelo qual o este foi criado

Um dos pontos chaves é construir a escala de características de sentimentos com a metodologia adaptada ao contexto que se deseja monitorar (SALUSTIANO, 2015). Para isso, a pesquisa propõe investigar se há aderência da escala de características de sentimentos atualmente utilizada, com a proposta de atendimento aos clientes que consomem serviços bancários desta instituição financeira. Por exemplo, essas características de sentimentos, Negativa, Reclamação, Informação, Elogio, distribuídas na polaridade negativo – neutro – positivo, utilizadas por um dos modelos analíticos criado pela instituição financeira, estão adequadas para compreender e traduzir o sentimento de um consumidor bancário? Estas

características foram escolhidas pela área técnica da instituição financeira em estudo e estão sendo utilizadas pelo modelo para o monitoramento de sentimentos. A falha mais comum apontada na literatura para a criação de metodologias de monitoramento de sentimentos é a identificação dos aspectos relevantes ou que são subjetivos e somente terão algum sentido se forem admitidos o contexto e as nuances do negócio analisado. Essas particularidades ainda não são identificadas pelos algoritmos e ferramentas a algum contexto específico (BRODY e ELHADAD, 2010; COHEN *et al*, 2011; TITOV e MCDONALD, 2008).

1.3.2.2 Comparar as acurácias obtidas dos dois modelos analíticos PLN

Foi realizada comparação de acurácia entre os dois modelos que usam as técnicas Word2vec e BERT, com vistas verificar qual deles obteve o melhor acerto com a avaliação humana, que no caso desta pesquisa pode ser considerado como “regra de ouro”. No item metodologia serão tratados de conceitos e as características técnicas de cada uma dessas ferramentas. Salustiano (2015) defende que a falta de sofisticação e consciência contextual das ferramentas reduz a capacidade para identificar com precisão a polaridade de sentimentos, sugerindo que a área técnica de TI esteja sempre buscando através de comparações de eficiência dos modelos, o melhor ajuste destes. E uma das possibilidades de realizar essa tarefa é utilizando a regra de ouro (*standard gold*) da avaliação humana em bases amostrais estatísticas destes vultosos volumes de dados.

1.3.2.3 Comparar os índices de concordância entre a avaliação humana e os dois modelos analíticos

Percebe-se que um dos pontos que proporcionam as maiores divergências de opiniões na área de análise de sentimentos é a discussão de quem tem a melhor acurácia para avaliar: o ser humano ou máquina. Quem consegue desempenhar melhor o papel de classificador? Seria adequado empregar seres humanos para analisar os resultados e avaliar o sentimento, ou deve-se optar por tecnologias sofisticadas de classificação? (SALUSTIANO, 2016).

1.3.2.4 Analisar Escalas de classificação de sentimento do consumidor dos modelos analíticos

A polaridade representa o grau de positividade (satisfação), neutralidade ou negatividade (insatisfação) de um sentimento expresso por um consumidor perante um

determinado produto ou serviço. Alguns métodos classificam a polaridade de um sentimento em classes discretas do tipo binário (positivo ou negativo) ou ternário (positivo, negativo ou neutro) (ROSSI, 2019). Inseridos neste intervalo de polaridade, podem ser usados outros rótulos de escalas de sentimentos que transitam entre o positivo e o negativo como, por exemplo, interações classificadas como Ofensa, Negativa, Reclamação, Informação e Elogio. Desta forma não se pode descartar que à medida que se aumenta a quantidade de rótulos ou características de sentimentos que transitam nesta escala polarizada, aumenta também a probabilidade de divergência de concordância entre os avaliadores e o modelo analítico. Isto demonstra que se deve sopesar a quantidade de categorias utilizados pela instituição financeira para definir a eficiência dos modelos analítico utilizados para análise de sentimento do consumidor. Algumas dessas categorias e suas combinações de sentimentos podem acabar influenciando no resultado de concordância entre os avaliadores e o modelo analítico.

1.4 Justificativa

Desta forma, esta pesquisa se justifica do ponto de vista gerencial, pois ao avaliar, classificar e depurar as opiniões, traduzindo-as em percepção de sentimentos, as empresas alcançam um importante degrau para identificar essas vozes que conseguem exercer uma forte influência no consumo (SALUSTIANO, 2016). Mas ao mesmo tempo, as ferramentas que catalisam os *insights* de tomadas de decisões devem garantir fiabilidade dos resultados às áreas estratégicas da instituição financeira em estudo.

A comunicação em tempo real nas plataformas presentes na Web 2.0 tem produzido uma sociedade que além de consumir, produz muita informação. A agência de marketing Zephoria⁶ revela que cerca de 300 milhões de fotos são postadas por dia no Facebook, enquanto que a plataforma Internet Live Stats⁷ reporta que 350 mil tweets são postados a cada minuto. Neste contexto, a pesquisa se justifica por avaliar o desenvolvimento de técnicas para analisar este *Big Data*, a fim de explorar potenciais insights para organizações privadas e públicas (SIRQUEIRA, 2018, p. 9). Da perspectiva das empresas, as Redes Sociais On-line ou Mídias Sociais são poderosas fontes de dados sobre o comportamento dos consumidores, além de um novo canal de comunicação e de baixo custo (LOBATO, et al., 2016).

Segundo Silva (2016, p. 5), algumas pesquisas têm sido direcionadas para tratar o problema de análise de sentimentos em redes sociais sob o ponto de vista de um problema de classificação, com pouco consenso sobre qual é o classificador com melhor poder preditivo.

Assim este trabalho pretende além de investigar a acurácia dos dois modelos analíticos, também compará-los para verificar qual obteve a melhor calibragem na análise de sentimentos dos textos nas mídias sociais da instituição financeira.

Esta pesquisa pode ser justificada também pela baixa quantidade de trabalhos que abordem os modelos analíticos de avaliação de sentimentos em plataformas on-line especificamente de instituições financeiras nacionais, embora isso esteja mudando nos últimos anos, dada a importâncias que a indústria financeira tem dado ao tema. O trabalho ganha relevância na medida em que essas empresas estão inseridas num mercado de alta competitividade, oferecendo produtos muitas vezes comoditizados, e por deterem uma quantidade de clientes na escala de centenas de milhares que estão sempre aptos a influenciar positivamente ou negativamente a imagem das marcas dessas empresas.

2. Fundamentação Teórica

Neste Capítulo, são apresentados os principais conceitos que fundamentam a presente pesquisa. Na Seção 2.1, é abordada a evolução histórica da avaliação de sentimentos até o advento das mídias sociais. Na seção 2.2 são elencados os desafios e limitações encontrados pelas ferramentas analíticas para realizar de forma assertiva a avaliação dos sentimentos humanos em mídias sociais on-line. Na seção 2.3 aborda-se os maiores desafios encontrados na literatura para avaliação dos sentimentos das pessoas nas mídias sociais. Em que pese a pesquisa estar sendo desenvolvida na área de marketing, não poderia também deixar de ser relatada a metodologia operacional dos indicadores inter-observadores na seção 2.4. E encerra-se na seção 2.5 a fundamentação teórica com estudos anteriores que usaram a concordância humana (inter-observador) comparando-a com a avaliação automática de modelos analíticos.

2.1 A evolução da avaliação do sentimento humano em textos de grande escala

O sentimento humano, que deriva de suas emoções, é um dos diversos tópicos sobre o qual a academia de marketing procura estudar face à relevância do seu impacto no comportamento do consumidor (CANHOTO E PADMANABBAN, 2015). Captar o sentimento das pessoas é importante para identificar aquelas que estão mais dispostas a correr riscos, facilitando, portanto, o longo processo de tomada de decisão (FORGAS, 1991).

A importância de se compreender o poder de persuasão que uma opinião alheia pode exercer sobre outras pessoas fez emergir uma importante área de pesquisa conhecida como Análise de Sentimentos. Dessa forma, as pesquisas de marketing (como testes com clientes reais, grupos focais, pesquisas de opinião, etc.) são aplicadas há décadas e também consistem em uma forma de se monitorar ambientes (MONTARDO e CARVALHO, 2012).

Atualmente a análise de sentimentos é uma poderosa ferramenta no mundo dos negócios que pode avaliar a variável de investimentos em marketing e a percepção da aceitação de uma marca ou produto. Uma análise tradicional dos sentimentos coleta e realiza experimentações no campo da psicologia. Mas segundo Cohen et al. (2008) essas experiências mostram limitações para induzir, manipular ou isolar o humor das pessoas de efeitos independentes e dependentes.

Em que pese as áreas de Linguística e Processamento de Linguagem Natural (PLN) já serem temas consolidados, que tiveram início em meados da década de 1940 (TAMIS e JONES, 1994), não são encontradas muitas pesquisas feitas sobre a opinião e sentimentos das pessoas antes do ano 2000 (LIU, 2012). A princípio, a classificação das opiniões e emoções deste público que interage de forma *on-line* foi desenvolvida utilizando a captura e análise dos textos publicados de forma manual.

É natural aceitar que os seres humanos conseguem desenvolver a atividade de identificar sentimentos e opiniões a partir de textos de maneira bem mais satisfatória que as ferramentas automatizadas. Em contrapartida, há muitos obstáculos ao se realizar análises em grandes volumes de textos e dados relevantes “escondidos” em *sites*, *blogs* e redes sociais, o que trazem à tona a necessidade da evolução constante de sistemas e aplicações analíticas que, de maneira automatizada, realizem tal atividade (LIU, 2010). Assim, atividades de pesquisa nas áreas de tratamento de opiniões, sentimentos e emoções em textos encontrados em grandes escalas vêm ganhando importância devido à enorme quantidade encontrada nas Mídias Sociais On-line. Este ambiente acaba tornando-se fértil para que usuários sejam induzidos a dar opiniões sobre temas polêmicos e fortemente debatidos ou até mesmo demonstrar opiniões sobre produtos e serviços (LIU, 2012).

2.2 A análise de sentimento no contexto das mídias sociais

O que são redes ou mídias sociais? Segundo Passmore (2011), é uma estrutura social composta de indivíduos (ou organizações) chamados de “nós”, que são ligados (conectados) por um ou mais tipos de interdependência como amizade, parentesco, interesse comum, trocas financeiras, aversões, relacionamentos sexuais ou relacionamentos de crença, conhecimento ou prestígio.

As pesquisas e métodos de Análise de Redes Sociais (ARS) criaram um novo panorama de conhecimento, avançando em conceitos formais para teorias efetivas e estruturais sobre o funcionamento da sociedade (SILVA e STABILE, 2015, p. 241). Silva e Stabile apontam ainda que mesmo que a ARS não proporcione conclusões e postulados absolutos sobre o funcionamento da sociedade, o conhecimento é desenvolvido de forma nunca visto com o que é gerado por métricas, algoritmos, visualizações e modos de ver as dinâmicas sociais.

Analisar sentimentos em pequenos pedaços de textos em mídias sociais não é algo novo ou exclusivo do mundo online, tendo em vista que monitorar nada mais é do que aplicar

metodologias diversas para acompanhar e compreender o que falam de um determinado tema, marca ou produto tanto no ambiente offline quanto no virtual (MONTARDO e CARVALHO, 2012).

Um experimento realizado por Milgram (1967) em que pessoas aleatórias de duas cidades encaminhassem cartas a Boston, definiu já na década de 1960 a ideia de mundo pequeno. Concluiu também que a teoria de influência social é o efeito de um ou mais indivíduos sobre uma pessoa para realizar determinadas ações que não faria de forma espontânea e isolada, podendo ser uma figura pressuposta (autoridade), imaginária ou antecipada.

A Análise de Sentimentos, ou Mineração de Opinião teve início com processos relacionados com a identificação de emoções e opiniões em textos. E embora tenha ocorrido um aumento significativo de pesquisas nessa área nos últimos 5 anos, devido ao seu potencial de aplicabilidade em diversas áreas da ciência, o interesse por esta área já existe há algum tempo (WILSON et al., 2006; WIEBE et al., 2004; LIU, 2012). Segundo Pang e Lee (2008), o ano de 2001 marcou uma nova era sobre as oportunidades e problemas de investigação que essa área pode nos trazer.

Os métodos atuais de detecção de sentimentos de sentenças podem ser divididos em duas classes: aqueles baseados em aprendizado de máquinas e os métodos léxicos (dicionários abrangentes). Assim ao iniciar um projeto de modelos analítico de avaliação de sentimentos, deve-se primar pelo método que avalie de forma calibrada as opiniões nas redes sociais, permitindo não só compreender e explicar diversos fenômenos sociais complexos, mas também prevê-los (BENEVENUTO, RIBEIRO e ARAÚJO, 2015).

2.3 Os desafios de analisar o sentimento humano nas mídias sociais on-line

As mídias sociais trouxeram oportunidades de superar as limitações no campo do estudo de sentimentos de forma tradicional, já que as pessoas interagem em plataformas on-line compartilhando comportamentos, relacionamentos e por fim sentimentos de si mesmos (KIETZMANN et al., 2011). O grande volume de opiniões que circula permite observar o estado emocional e o humor de quem as emite. Ao mesmo tempo oferecem oportunidades para gerar insights nas tomadas de decisões das empresas.

Este ambiente se mostra como a principal fonte de pesquisa de avaliação de opiniões que o público expressa em relação ao grau de satisfação sobre produtos e serviços. Pode

mostrar também o estado emocional de quem opina, oferecendo oportunidades na solução de diversos problemas nos mais variados setores (SALUSTIANO, 2016).

Conforme argumentos de Fayyad et al. (1996), o alto volume de objetos (que pode chegar a 10^9) envolvidos faz com que a análise manual se torne muito difícil. A partir daí entra-se na era da automação da análise de sentimentos polarizados (positivos x negativos). Neste contexto, observa-se o desenvolvimento de ferramentas de monitoramento que apenas capturavam os dados, faltando ainda a classificação das opiniões (HAN, KAMBER & PEI, 2012).

A diferença entre os dois tipos de análise (automatizada e manual) acontece porque vários algoritmos que foram desenvolvidos nos últimos 10 anos, em sua maioria, utilizam uma derivação da classificação Bayesiana, em que pese estar sendo muito utilizada na construção de modelos analíticos preditivos, ainda ignora preceitos básicos, como o trabalho com a dubiedade de opiniões, a subjetividade do texto e a ironia da mensagem. Além disso, estão presos a uma abordagem que requer dados previamente conectados (“tagueados”) a um sistema de alimentação constante aos dicionários lexicais que utilizam uma lista predefinida de palavras associadas com um sentimento específico. O grande revés do modelo Bayesiano é a falta de praticidade em utilizar o mesmo dicionário em contextos diferentes (SHARMA; DEY, 2015).

Contudo, mesmo que a acurácia da atividade de monitoramento de sentimentos se desenvolvendo e melhorando a performance em validações, ainda existem dificuldades para a classificação do viés de *posts*, que podem ter variações adicionais como opinativa/subjetiva ou factual/objetivo (WIEBE et al., 2004). Monitorar a polarização usual, “Positiva” e “Negativa”, nem sempre leva em consideração as variáveis e emoções que podem estar presentes dentro de uma única frase (WILSON et al, 2005), que pode expressar mais de um sentimento e referir-se a mais de um objeto, criando incerteza quanto ao sentimento predominante. Por exemplo, o consumidor pode julgar um produto de forma positiva, mas expressar insatisfação com características específicas (LIU, 2010).

O sentimento sobre um objeto pode não ser expresso diretamente, mas por meio de comparações, o que exige do analista profissional possuir conhecimento para identificar se os termos comparativos usados refletem uma opinião positiva ou negativa sobre o produto (LIU, 2010). Segundo Subasic e Huettnner (2001), tentar classificar somente emoções, sem levar em consideração os sentimentos que podem estar subentendidos em uma única frase, como os que incluem felicidade, tristeza, raiva, ódio, ironia, etc, é o grande desafio e de fundamental

importância para o monitoramento de indivíduos online para elevar a fiabilidade dos modelos analíticos.

Em termos dos aspectos sintáticos e estilísticos de expressar sentimento, Canhoto e Padmanabhan (2015) observaram que os usuários de mídia social tendem a aplicar certos coloquialismos e abreviações com significados múltiplos e em constante evolução. Por exemplo, SQN começou a ser um acrônimo para ‘Só que não’, FDS para ‘fim de semana’. Além disso, os usuários também podem empregar uma matriz crescente de símbolos de texto ou *emojis* para apoiar a comunicação de sentimentos. E para aumentar a complexidade ainda mais, classificar sentimentos e as causas de emoções, apesar de serem ações distintas, estão intimamente ligadas entre si, cabendo uma análise mais profunda para identificar dentro de uma mensagem qual é o sentimento e qual é a emoção (GAO et al, 2013). Há diversas palavras que podem ser associadas com qualquer emoção única e outras que podem ser associadas com mais de uma emoção. Além disso, a complexidade de classificar aumenta quando percebemos que às vezes transmite-se emoção sem a utilização de palavras óbvias (subjetividade).

Uma ferramenta que demonstra a polarização dos vieses pode ser observada na Roda de Emoções de Plutchik (2001), cuja aplicabilidade é a realização de análise lexical de frases curtas e posteriormente classificá-las em dois tipos: positivo e negativo. Porém, esse modelo apresenta resultados que podem ser úteis somente para uma análise superficial quando não forem requeridas informações mais ricas.

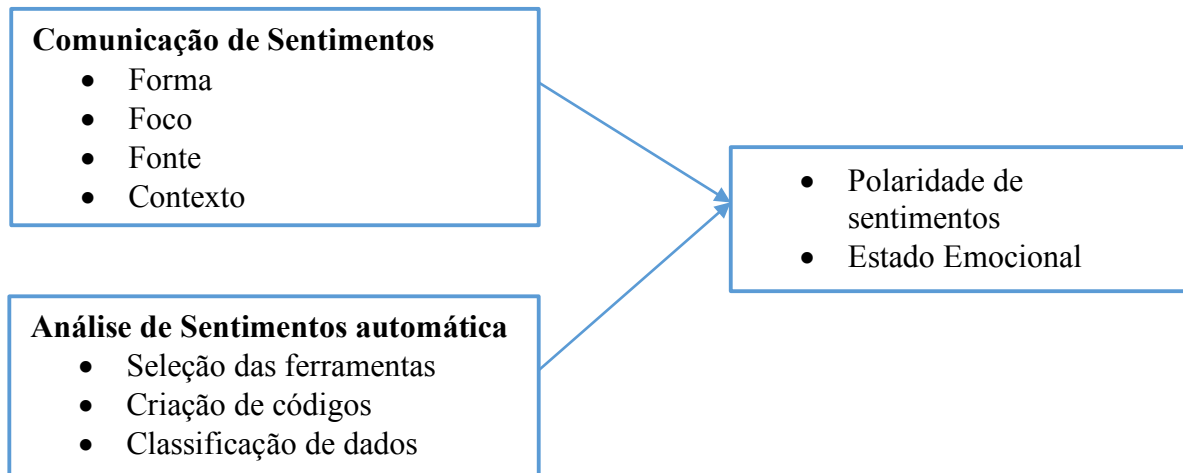
A teoria psicoevolucionária integrativa das emoções, abordada por Plutchik (2001), realiza a classificação das emoções primárias em raiva, medo, tristeza, nojo, surpresa, curiosidade, aceitação e alegria. Porém o estudo do sentimento e o estado emocional nas conversas nas mídias sociais, embora seja algo promissor, pode ser negativamente impactado por questões relacionadas à comunicação de sentimento em conjunto com a análise automatizada de sentimentos.

Segundo Canhoto e Padmanabhan (2015), as fontes relacionadas à vulnerabilidade da identificação de sentimentos realizadas de forma manual, compreendem: a forma, o foco, a fonte e o contexto. Já as questões que podem impactar a análise automatizada de sentimentos podem ser: a seleção das ferramentas, a criação de códigos (programação) e a classificação de dados, conforme observado na Figura 1.

Figura 1 – Fontes de vulnerabilidade no estudo do sentimento captado nos textos das Redes Sociais On-line

Questões associadas com:

Impacto Negativo na Acurácia de:



Fonte: Adaptado de Canhoto e Padmanabhan (2015)

2.4 Inter-Observador para validar um modelo analítico

Os modelos analíticos de análise de sentimentos são algoritmos que desempenham o papel de reconhecer os sentimentos captados em mídias sociais de forma automatizada e são criados por programadores de linguagem computacional com o objetivo de recuperar informações através do advento do Processamento de Linguagem Natural (PNL) para analisar textos em um ou mais níveis linguísticos, com o objetivo de emular o processamento humano da língua (LIDDY, 2003).

Estes modelos são desenvolvidos através de criação de códigos lógicos de linguagem de computador que buscam um conjunto de dados de entrada e com base em determinados padrões encontrados, conseguem prever as saídas. Coelho (2020) relata que estes métodos autônomos de algoritmos aprendem interativamente a partir de dados e é o conceito fundamental de Aprendizado de Máquinas ou *Machine Learning*.

Uma das etapas de Aprendizado de Máquinas defendida por Witten (2016) é o treinamento e a avaliação, que consiste da escolha de um algoritmo apropriado e de um método de representação dos dados na forma de um modelo. Os dados são divididos em duas partes: treinamento e teste. Os dados de treinamento são utilizados no desenvolvimento do modelo, e os dados de teste são usados como referência, para testar a acurácia do modelo. Este grupo de dados para a parte do teste não podem ser o mesmo do grupo de dados de treinamento, pois o grupo específico dos dados para testagem são utilizados apenas para avaliar a acurácia e validação do modelo.

Esta pesquisa não possui a intenção de discutir ou propor uma alteração do método de validação automático atualmente utilizado pela área de TI da instituição financeira. A proposta aqui foi tão somente realizar uma abordagem e inferência a fim de refinar a validação da etapa de testagem do modelo analítico. Deste modo foi avaliada apenas a calibragem da acurácia da validação automática do modelo. Para isso foi utilizando uma base de amostragem estatística, que já teve sua avaliação realizada pelas ferramentas automáticas, para as avaliações de concordâncias entre duas pessoas (*inter-observador ou inter-rater agreement*) e a ferramenta.

O método manual ou semi-automatizado depende de no mínimo um profissional para analisar, classificar e tipificar qual é o sentimento da mensagem de texto e possui como pontos positivos, destacados por Salustiano (2015), a maior precisão e a facilidade para trabalhar com subjetividade comparado a análise automatizada.

Por outro lado, a análise de grandes volumes de dados, que é a situação real encontrada em uma grande instituição financeira como a analisada por esta pesquisa, onde transitam centenas de milhares de opiniões em textos por dia, inviabilizaria do ponto de vista de custo e tempo, o advento de uma análise de sentimento na forma manual.

Assim, será utilizado como ferramenta de medição o indicador inter-observador – *kappa* ponderado – para obter o grau de concordância entre os avaliadores humanos e a testagem automatizada do modelo. O resultado indicará uma medida de acurácia do modelo analítico, além da possibilidade de o resultado servir de comparação com a própria medida de acurácia que é obtida de forma automatizada na etapa de teste no arcabouço de *Machine Learning* do modelo.

Os detalhes práticos do método serão abordados na Metodologia.

2.5 Estudos Anteriores

Foi encontrado na revisão de literatura um único estudo que abordou o contexto do tema e os métodos comparativos de análise de sentimento manual e automatizada. Trata-se de Canhoto e Padmanabhan (2015) que realizou a pesquisa utilizando textos retirados do *Twitter* em que o contexto remetia a conversas e relatos de pessoas diversas sobre a bebida Café. Os resultados desta pesquisa mostraram baixos índices de concordância entre a análise manual e a análise automatizada.

Essa pesquisa sugeriu que um dos motivos da dificuldade de ainda não se alcançar uma boa acurácia em modelos analíticos de análise de plataformas on-line, consiste no fato de

os conteúdos de mídia social, principalmente aqueles que fazem críticas a marcas ou serviços, frequentemente empregarem ironia e sarcasmo. É muito difícil para os analistas de TI detectar e classificar o conteúdo sarcástico em geral, devido às suas nuances e multidimensionais natureza (BERGH, LEE, QUILLIAM e HOVE, 2011). A pesquisa de Canhoto e Padmanabhan utilizou também a análise de dois modelos analíticos que foram desenvolvidos por instituições diferentes (uma empresa e uma universidade). A codificação de ambos modelos foi feita com um esquema que refletia a polaridade de emoção (positiva vs. negativa).

Abaixo foram sintetizadas literaturas que abordam conceitos e etapas para construção de modelos analíticos de sentimento, incluindo também referencial que demonstra as limitações e os desafios de se desenvolver ferramentas eficientes de análise de sentimento obtidas de eventos de mídias sociais:

Tabela 1 – Análise de Sentimentos: Resumo de Conceitos Fundamentais

Tema	Conteúdo	Autores
Mineração de Dados: Identificação, classificação e modelagens	I) Definição de Dados; Interpretação e avaliação dos resultados; II) Entendimento do negócio; III) A evolução da análise dos dados; IV) Extrair informações realizando a mineração de texto	Fayyad et al. (1996), Chapman et al. (2000); Redhu et al. (2018), Salustiano (2016)
Análise manual Polarizada	Reconhecer expressões com sentimentos em tweets, incluindo palavras / frases formais e de gíria; Avaliar a polaridade de cada expressão de sentimento; Atribuir polaridade a uma expressão de sentimentos sobre o conjunto de tweets.	Han, Kamber & Pei (2012)
Subjetividade na linguagem natural	Os aspectos da linguagem usados para expressar opiniões, avaliações e especulações	Wiebe et al. (2004); Wilson et al. (2005)
A classificação de sentimentos e emoções	O aprendizado conjunto sobre a classificação de sentimentos e emoções visando a melhora do desempenho na análise e classificação de textos livres da web.	Gao et al (2013); Plutchik (2001),
Vulnerabilidades	Ferramentas de análise automatizadas podem limitar a capacidade de identificar polaridade de sentimentos e o estado emocional no conteúdo da mídia social.	Canhoto e Padmanabhan (2015) Liu (2010)

Fonte: Elaborado pelo Autor

3. Metodologia

Neste item são apresentadas as etapas e os procedimentos da pesquisa científica que tem por objetivo responder à questão problema relatada na Introdução. A pesquisa científica leva em consideração um conjunto de procedimentos sistemáticos, que se apoia no raciocínio lógico, usando alguns métodos científicos para encontrar soluções sobre um problema definido previamente pelo pesquisador (TUMELERO, 2019). Conforme defendido por Galliano (1986, p. 26), ao relatar um fato, o conhecimento científico não apenas trata de explicá-lo, mas também busca descobrir suas relações com outros fatos e explicá-los.

A estruturação da pesquisa científica, conforme Garcia (2015, p. 77) é uma das “fases do processo metodológico e está pautada em diferentes etapas, sendo elas, a formulação do problema a ser questionado, a concepção de hipóteses a serem verificadas, a coleta de dados, tabulação dos dados, análise dos dados, discussão dos resultados, conclusões, redação do texto e apresentação do trabalho científico”.

3.1 Quanto a abordagem

Quanto a abordagem essa pesquisa pode ter a primeira etapa dela definida como “Quantitativa” e a segunda etapa definida como “Qualitativa”. Na primeira etapa (quantitativa) foi obtida uma base de amostragem da área de tecnologia da instituição financeira já classificada com as percepções de sentimentos de 10.600 interações de textos escritos pelos clientes nas mídias sociais. Nesta base de amostragem, conforme orientação da área técnica, não havia sido aplicada filtragem de segmentação de atendimento (renda, faturamento, pessoa física, pessoa jurídica e atividade). Essa base de amostragem foi convertida em uma amostra aleatória menor através da função “=INT(ALEATÓRIO()*11200)” que listou de forma aleatória linhas de interações até completar as 400 interações dentre os 10.600 *posts*, sendo 200 para cada modelo analítico e com cada frase ou interação disponibilizada aos avaliadores em uma planilha Excel, conforme figura 1.

No modelo analítico baseado em Word2vec a amostra é do dia 19/03/2020 ao dia 30/04/20, já o modelo baseado em BERT a amostra foi coletada do dia 01/12/20 ao dia 12/01/21. Os métodos e a forma como estes modelos funcionam serão descritos a seguir nas seções 3.4 e 3.5.

Figura 2: Formato dos relatos das interações entregues aos avaliadores

		elogio	reclamacao	negativa	informacao	ofensa
1	text					
2	Não, só isso! obrigada	0	0	0	0	0
3	Boa tarde, nÃ£o estou conseguindo fazer cadastro da minhas chaves no PIX pelo celular ... aparece uma mensagem de erro solicitando contato age	0	0	0	1	0
4	Quero parcelar essas fatura que estao em aberto	0	0	0	1	0
5	Um problema criado pelo BB e VISA. Tenho limite e dinheiro na conta, graças a Deus.	0	0	0	0	0
6	renda atualizada é como da um total de 9mil, fica mais fácil correto?	0	0	0	1	0
7	Bom dia! O meu especial sumiu, pode verificar por favor? Eu encaminhei o e-mail com minha declaração de IR para o e-mail do banco	0	0	0	1	0
8	JÃi preenchi uma avaliação	0	0	0	0	0
9	Nao...	0	0	0	0	0
10	Até que horas posso chegar lá?-	0	0	0	1	0
11	Preciso fazer uma transferência de 1450,00	0	0	0	1	0
12	Ver algo*	0	0	0	1	0
13	Bom dia eu tentei transferir dinheiro de uma conta minha noutro banco para cá mas não consegui!!! Eu fiz pelo PIX do outro banco.	0	0	0	1	0
14	Pra saber como funcionava	0	0	0	1	0
15	consegue me ajuda?	0	0	0	1	0
16	Pronto, combinado	0	0	0	0	0

Fonte: Print da tela do Excel com as interações/posts

A segunda etapa (qualitativa) consistiu em repassar os 400 posts/interações, distinguindo-os dos dois modelos analíticos utilizados pela instituição financeira, a dois juízes avaliadores para classificarem os sentimentos percebidos em cada frase relatada pelos clientes da instituição financeira. Não foi informada qual rede social estavam os textos a serem analisados, e nem a informação de que já haviam sido avaliados e classificados por modelos analíticos, pois essas informações poderiam enviesar o resultado da pesquisa.

Os dois avaliadores são funcionários da instituição analisada, possuem aproximadamente o mesmo tempo de função Gerente de Relacionamento de Unidades Negociais/Comerciais e foram orientados a colocar o número “1” nas células em que eles julgavam ser a coluna de característica escolhida para cada interação/post. E apenas para as interações do modelo BERT, foi orientado aos observadores que para aquelas interações em que não houvesse alguma característica a ser escolhida, que eles não alterassem as células que já estavam com zeros.

A quantidade de avaliadores, que no caso desta pesquisa são três (Modelo Analítico, Avaliador A e Avaliador B), está de acordo com a literatura, que define que a concordância entre juízes (inter-observadores) é o grau em que dois ou mais avaliadores, utilizando a

mesma escala de avaliação, fornecem igual classificação para uma mesma situação observável. (GRAHAM et al, 2012, p. 5).

3.2 Quanto a natureza

Trata-se de uma pesquisa “Aplicada” em que o objetivo é responder uma questão prática que vai medir a acurácia/fiabilidade dos modelos estatísticos desenvolvidos pela instituição financeira, fazendo uso de indicadores estatísticos de confiabilidade e concordância entre os modelos analíticos e observadores humanos com o intuito de identificar a eficiência das ferramentas que já estão sendo utilizadas para tal fim.

3.3 Quanto aos objetivos

Trata-se de uma pesquisa “Experimental ou Exploratória” que tem por objetivo investigar a pergunta problema que envolve a medição de indicadores que permitam mostrar o nível de acurácia/fiabilidade de dois modelos analíticos utilizados na análise automatizada de sentimentos expressados pelos clientes em seis mídias sociais oficiais de relacionamento de uma instituição financeira: FaleCom, Facebook, Twitter, Whatsapp, Linkedin e Instagram.

Os métodos analíticos NLP (Processamento de Linguagem Natural) utilizados para análise automatizada desta instituição financeira foram o Word2vec (MIKOLOV et al., 2013) que foi criado visando realizar a incorporação de palavras e existe desde 2013 (ALAMMAR, 2019) e BERT (Bidirectional Encoder Representations from Transformers) (DEVLIN et al, 2019), técnica de aprendizado de máquina baseada em *transformer* para pré treinamento de PNL desenvolvida pelo Google.

Como já abordado no capítulo do referencial teórico, os modelos analíticos são criados de forma autônoma e cada profissional deve construir essas ferramentas de análise de sentimento de forma que as adapte às características particulares e nuances de cada negócio adaptadas a atender o negócio.

3.4 Word2Vec

Focando na análise da performance de concordância da máquina com o senso humano,

é possível esclarecer que a técnica Word2vec é uma ferramenta *open-source* (de acesso livre na internet), constituída por Mikolov et al., que visa a representação vetorial de palavras utilizando técnicas de redes neurais artificiais (AGUIAR e PRATI, 2015). A técnica busca identificar a similaridade entre as palavras no texto e a probabilidade condicional de surgir algumas palavras dada a ocorrência de outras palavras. Como exemplo podemos propor que o modelo calcule ou até mesmo prediga a probabilidade, que a palavra “cara” esteja relacionada com a palavra “tarifa” contida na frase obtida das mídias sociais deste banco – “O valor da tarifa da minha conta está cara”.

Foram escolhidos dois avaliadores que trabalham na área comercial da mesma instituição pesquisada. Estes realizaram a classificação das características de sentimento, previamente definidas, expressas em 200 frases dos clientes que interagiram com as mídias digitais da instituição financeira. Esta amostra de frases obtidas junto à área de tecnologia da instituição pesquisa, já teve a análise de sentimentos realizada utilizando o modelo automatizado de análise *Word Embedding*¹ Word2Vec que se propõe comparar com a análise humana (concordância inter-observador) nesta pesquisa.

O modelo analítico que usa Word2Vec, fonte desta pesquisa, foi construído com a constituição de quatro características classificadas e conceituadas conforme abaixo:

- Negativa: Reclamação forte, xingamentos, ofensas, ameaças, etc.
- Reclamação: Reclamação Leve, sem ofensas e palavras de baixo calão.
- Informação: Pedidos de informação, dúvida ou sugestão.
- Elogio: Elogio a produtos, serviços, atendimentos, etc.

Segundo informações obtidas da área técnica de TI que desenvolveu o modelo, essas características foram escolhidas utilizando o critério do sentido inverso de significado de rotulação. Ou seja, inicialmente foram analisados os principais estados de humor obtidos dos textos nas mídias sociais e então para um conjunto de conceitos, foram categorizados os rótulos aqui nominados de características de sentimentos. Não obtivemos da instituição financeira o referencial teórico ou estudos que deram origem à definição de conceitos dessas quatro características.

Com vistas a ampliar a percepção do estado de espírito do interlocutor/cliente e possibilitar um aprofundamento na compreensão da reputação dos serviços oferecidos, os desenvolvedores da ferramenta permitiram que o modelo utilizasse mais de uma característica de sentimento nas avaliações individuais dos textos.

¹ Conforme Carvalho (2018), o termo define um conjunto de técnicas que mapeia a semântica e sintática de uma linguagem natural em um espaço real utilizando estatísticas.

Assim, as características de percepções combinadas foram reclassificadas em subcategorias na seguinte ordem crescente partindo de desfavorável para favorável: 1) Negativa; 2) Negativa e Reclamação; 3) Negativa, Reclamação e Informação; 4) Negativa e Informação; 5) Negativa e Elogio; 6) Reclamação; 7) Reclamação e Informação; 8) Informação; 9) Elogio.

Desta forma, as características de sentimento ou estado de humor, utilizadas com o modelo Word2Vec, classificaram os relatos dos clientes realizando a combinação simultânea de mais de uma característica de sentimentos, inclusive podendo ter três características associadas em uma única classificação, como, por exemplo, perceber ao mesmo tempo em um texto, características como como negativa, reclamação e informação, o que deixou ainda mais complexa a análise e o estudo de concordância inter-observador.

Os dois observadores A e B já relatados receberam as nove características possíveis de classificação, e assim realizaram a tarefa de forma independente julgando o estado de sentimento em cada post/interação obtidos do grupo de dados disponibilizado pela área de tecnologia em planilha Excel conforme modelo da figura 1.

Não obstante a literatura relatar que um único segmento de texto pode expressar mais de um sentimento e referir-se a mais de um objeto, criando incerteza quanto ao sentimento predominante (Liu, 2010), a possibilidade de combinar mais de uma característica sem uma motivação racional ou estudo técnico por parte da instituição, acabou criando algumas situações inusitadas e complexas para a classificação dos textos. Isto foi observado na característica antagônica (5) Negativa e Elogio que causou questionamentos dos avaliadores em relação à possibilidade deste acontecimento ser uma realidade factível. Novamente a instituição não forneceu a motivação técnica para o estabelecimento das características combinadas que proporcionaram o aumento de quatro para nove possibilidades avaliativas.

Encontra-se na literatura sobre métodos de pesquisa que o desenvolvimento de um constructo, está relacionado com a extensão em que diversos pesquisadores concordam com a classificação de dados específicos, ou seja, a taxa de aceitação e concordância entre análises sistemáticas que ajudam a entender algum aspecto de um estudo ou ciência (GWET, 2012).

Basit (2003) defende que os profissionais de TI ou pesquisadores precisam estar ativamente envolvidos na criação de categorias e na decisão de quais dados recuperar e agrupar.

Para compreender porque ocorrem essas situações de aparente anomalia, recorreremos ao que foi abordado na seção *1.1 Abordagem da Escolha do tema*, que alerta que podem existir modelos, metodologias e técnicas que são personalizadas por profissionais de TI e que

terminam por não envolver outras áreas que poderiam contribuir à constituição das características de sentimento, como por exemplo, marketing e psicologia.

Além de estudar a polaridade do sentimento, estes pesquisadores dessas áreas indicadas poderiam analisar estados de humor com maior eficiência, como destacam Laros e Steenkamp (2005) sobre a importância de compreender o tipo específico de emoção vivida pelo consumidor, porque as emoções são altamente diferenciadas em seu impacto (LAROS & STEENKAMP, 2005). Por exemplo, a infelicidade e a raiva são ambas negativas emoções. No entanto, têm consequências diferentes em termos de comportamento do consumidor, em que o primeiro carece de foco, enquanto o último tende a ser direcionado e levará a respostas específicas do contexto (BUSHMAN, BAUMEISTER, & PHILLIPS, 2001).

3.5 BERT (Bidirectional Encoder Representations from Transformers)

BERT é o acrônimo para *Bidirectional Encoder Representations from Transformers*, que pode ser traduzido como representações de codificador bidirecional de transformadores. Possui pouca similaridade com o Word2Vec, a não ser que se identifica como um modelo de PLN baseado em rede neural, mas com uma enorme vantagem: não precisa processar o texto sequencialmente, palavra por palavra, podendo receber todas as palavras de uma só vez, o que diminui o tempo de processamento já que é possível paralelizar as operações computacionais envolvidas (CECCON, 2020).

O que faz do modelo ser tratado como um código revolucionário no ambiente de PLN, é que é treinado exaustivamente para que seja capaz de compreender uma língua, no que é chamado na área de modelagem da linguagem.

Para isso, faz uso de duas estruturas internas na forma de um *encoder* e um *decoder*, os quais têm a tarefa tradicional de um *autoencoder*: ser capaz de representar o *input* da rede (no caso, textos) em um espaço multidimensional complexo (tarefa mais especificamente do *encoder*) que, nesse caso, tem a propriedade de capturar as nuances da língua necessárias à sua compreensão (CECCON, 2020).

Este modelo foi desenvolvido após a utilização da solução Word2Vec, e evoluiu em algumas frentes, como por exemplo, a definição e conceitos de forma bem mais rica em detalhes de compreensão, para uma melhor adequação da classificação das características do estado de sentimentos dos clientes:

Elogio	Considera-se elogio qualquer manifestação do cliente que envolva: elogio a produtos e serviços do BB, sobre a própria instituição, sobre o atendimento do funcionário, como: agilidade, presteza e cortesia, e que utilizem adjetivos claros de contentamento a exemplo de excelente, perfeito, maravilhoso, etc. Também é considerado elogio frases que apresentem gratidão acompanhadas por uma característica daquilo que se é grato, por exemplo, "Obrigado pela atenção de sempre" é elogio, "Obrigado pela atenção" não é elogio
Informação /Solicitação /Dúvida	Considera-se informação, dúvida e solicitação, qualquer pedido do cliente envolvendo dúvidas, solicitação e informação sobre produtos, serviços e canais de atendimento do Banco ou da própria instituição, inclusive a queixa que subentende um pedido e pontos de interrogação. Exclui a prestação de informações (ex: meu CPF é...), perguntas habituais de saudação (ex: como você está?) e expressões comuns de encerramento (ex: Ficarei no aguardo.).
Negativa	Consideram-se negativas frases de cliente/usuário quando subentendido se tratar de produtos, serviços, atendimento do próprio Banco que: (i) envolva alguma AMEAÇA e/ou tomada de AÇÃO junto ao SAC/Ouvidoria ou outras instituições. Essas instituições podem ser órgãos fiscalizadores, defesa do consumidor ou concorrência (ex.: BACEN/PROCON/BRADESCO etc.). (ii) contendo palavras de baixo calão ou palavras ofensivas (ex: banco mesquinho), exceto os casos que explicitamente se refiram a instituição/serviços/produtos alheios ao Banco em estudo. (III) com pedido ou ameaça de cancelamento de produtos e serviços ou abandono da instituição.
Ofensa	Considera-se ofensa a utilização por parte do cliente ou usuários de palavras ou expressões difamatórias, caluniosa, injuriosas, agravantes, ultrajantes, xingamentos que atinja o destinatário em sua honra, dignidade e decoro. Com possibilidade de descrédito perante a opinião pública, podendo ser estas mesmas expressões atribuídas a produtos e serviços do banco ou a instituição, exceto quando a expressão utilizada refletir apenas descontentamento do cliente. Para entendimento: calúnia (crime descrito no artigo 138 do Código Penal) é acusar alguém publicamente de um crime – sabendo que a pessoa não o cometeu Ex: “Você é um ladrão”. Difamação (art. 139) é dizer que a pessoa foi autora de um ato desonroso – ou seja, sair espalhando algo sobre alguém que prejudica sua reputação. Ex: “Fulano é um viciado”. Injúria (art. 140) é basicamente uma difamação que os outros não ouviram: é chegar e dizer para um sujeito algo que esse sujeito considere prejudicial. Ex: “Você é ignorante”.
Reclamação	Considera-se reclamação frases de cliente/usuário quando subentendido se tratar do produtos, serviço, atendimento ou do próprio BB contendo queixa ou expressão de descontentamento, excessivo ou não, exceto quando na reclamação ficar explícito se tratar de outra instituição.

Desta vez o modelo analítico realizou o julgamento de 15 características de estado de sentimento para classificar os *posts* dos clientes da instituição financeira na seguinte ordem de polaridade de negativo para positivo: 1) Ofensa, 2) Ofensa e Negativa, 3) Ofensa, Negativa e Reclamação, 4) Ofensa e Reclamação, 5) Ofensa, Reclamação e Informação, 6) Negativa, 7)

Negativa e Reclamação, 8) Negativa, Reclamação e Informação, 9) Reclamação, 10) Reclamação e Informação, 11) Reclamação e Elogio, 12) Neutro, 13) Informação, 14) Informação e Elogio, 15) Elogio.

A característica “Neutro” foi utilizada para classificar *posts*/interações que não tinham qualquer relação com as outras 14 características e foi ordenada logo anterior à característica informação, tendo em vista as duas estarem muito próximas sob o aspecto da polaridade de sentimento. Foram repassadas todas essas informações aos avaliadores para que realizassem a simulação com os mesmos critérios realizados pelo modelo analítico. Para obter os resultados dos indicadores de confiabilidade e concordância foram utilizadas as ferramentas estatísticas: Alfa de Cronbach, Cohen Kappa, Fleiss Kappa e Kappa ponderado.

3.6 Alfa de Cronbach

Antes de iniciar os testes de concordância entre os avaliadores humanos e a ferramenta, será calculado o indicador de confiabilidade da amostra de *posts*, o coeficiente *Alfa de Cronbach*, que certamente é uma das ferramentas mais difundidas em pesquisas envolvendo a construção de testes de concordância. O coeficiente alfa foi apresentado em 1951 por Lee J. Cronbach (CRONBACH, 1951) e calcula a média das correlações entre os itens, que em nosso caso são os avaliadores humanos e os modelos analíticos que manifestaram em uma escala de percepção as características de sentimentos demonstradas nas interações dos clientes da instituição nas plataformas de mídias sociais. Neste estudo, como resultado, ele dará a medida de confiabilidade ou consistência interna de nossa amostra, em uma escala de 0 a 1 da medida.

Imaginando uma tabela denominada matriz X ($n \times k$) onde cada linha de X corresponde aos julgamentos, que neste caso são os Avaliadores A e B e a Ferramenta, e cada coluna corresponde às respostas quantificadas (1 a 9) das características de sentimentos percebidas. De acordo com Leontitsis e Pagge (2007), o coeficiente alfa de Cronbach é mensurado de acordo com a Equação:

$$\alpha = \frac{k}{k-1} \left[\frac{\sigma_t^2 - \sum_{i=1}^k \sigma_i^2}{\sigma_t^2} \right]$$

onde σ_i^2 é a variância de cada coluna de X , ou seja, é a variância relacionada à cada característica de sentimento percebida de X , e σ_t^2 é a variância da soma de cada linha de X ,

ou seja, é a variância da soma das respostas dos avaliadores. Também deve ser observado que k deve ser maior do que 1 para que não haja zero no denominador e n deve ser maior do que 1 para que não haja zero no denominador no cálculo do σ_i^2 e do σ_t^2 .

3.7 Cohen Kappa, Fleiss kappa – k e Coeficiente de Kappa ponderado – kw

Verifica-se que o termo confiabilidade é cunhado entre avaliadores (inter-observadores) como uma expressão genérica para medir a consistência entre avaliadores. Encontra-se referências bibliográficas que apontam diversas maneiras de medir o nível de acurácia e concordância entre avaliadores, incluindo técnicas como porcentagem, correlação, coeficiente Kappa de Cohen, dentre outras (HANEY et al., 2004). Este último é um procedimento estatístico que foi desenvolvido por Cohen (1960) para aperfeiçoar as técnicas tradicionais que não excluam de sua análise a concordância devido ao acaso, ou seja, vários avaliadores podem casualmente classificar da mesma maneira, sendo isso um acaso e não propriamente uma concordância.

O método alfa de Cronbach, abordado acima, consegue medir o grau de concordância e consistência interna das respostas, mas este método não leva em conta a proporção desse acordo que é devida ao acaso, o que foi proposto por Cohen (1960).

O coeficiente k é o mais utilizado quando é necessário classificar dados em categorias nominais (CHEN e KRAUSS, 2004). Não obstante, o próprio Cohen (1968) desenvolveu o kappa ponderado (kw) indicado a categorias ordinais. Desta forma, atendeu-se à necessidade de alguns pesquisadores que buscavam atribuir diferentes pesos aos desacordos, tornando-se assim uma estatística de concordância preferível quando se quer classificar um conjunto de dados em categorias ordenadas semelhantes às construídas pelos modelos analíticos da instituição financeira, objeto desta pesquisa.

Visando dar mais detalhes ao estudo, e apesar de considerar o kw mais adequado ao presente estudo, será descrita a solução do *kappa* de Cohen com pressupostos, fórmulas e aplicações práticas, partindo com o coeficiente k , destinado a categorias nominais e evoluindo para o kw que passa a considerar cada ordem de peso nas discordâncias, já que se tem estabelecidas características de sentimentos que podem ser ordenadas como, por exemplo, de forma análoga a uma escala de *likert* (1 a 10).

Conforme Fonseca et al. (2007), os dados relativos às classificações dos avaliadores são apresentados em tabelas de frequências cruzadas “Avaliador A *versus* Avaliador B” (semelhantes à Tabela 2). Assim, o coeficiente κ pode ser definido como a proporção de

acordo entre os avaliadores após ser retirada a proporção de concordâncias ou acordos devido ao acaso, conforme equação:

$$\text{Proporção } k = \frac{Po - Pa}{1 - Pa} \text{ ou para frequências } k = \frac{\sum fa - \sum fe}{N - fe}$$

na qual Po (ou $\sum fa$) é a proporção do acordo observado, ou seja a proporção de unidades que os avaliadores classificaram nas mesmas categorias, e Pa (ou $\sum fe$) a proporção do acordo devido ao acaso, ou seja a proporção de unidades classificadas pelos juízes nas mesmas categorias por mera coincidência.

Fonseca et al. (2007) explicam de uma forma mais didática como obter o resultado do coeficiente,

onde a proporção total de acordo observado (Po) é obtida pela soma dos valores das células que se encontram na diagonal traçada na tabela (do canto superior esquerdo para o inferior direito), ao passo que a proporção de acordo devido ao acaso (Pa) é obtida através da soma do produto entre total linha x total coluna e divisão desse valor pelo total de observações também para todas as células da diagonal, conforme tabela x.

Já para obtermos o κw são utilizadas três matrizes $n \times n$: uma para as frequências observadas (equivalente a Po da fórmula do kappa); outra para as frequências esperadas (equivalente a Pa da fórmula do kappa); e uma terceira matriz de dados para os pesos atribuídos aos desacordos. A fórmula do κw é expressa da seguinte forma:

$$\text{Para proporções } \kappa w = 1 - \frac{\sum vij \, poi j}{\sum vij \, pai j} \text{ e para frequências } \kappa w = 1 - \frac{\sum vij \, fai j}{\sum vij \, fei j}$$

na qual vij = peso do desacordo atribuído a cada célula da matriz $n \times n$, $poi j$ =proporção observada em cada célula (ou $fai j$ =frequência observada em cada célula) e $pai j$ =proporção esperada na célula devido ao acaso (ou $fei j$ =frequência esperada na célula devido ao acaso). Na sessão de Resultados são apresentados os exemplos práticos para a realização do cálculo.

Em que pese não existir um valor específico padronizado e definido a partir do qual se deva considerar o valor do *kappa* como adequado, haja visto que para cada teste analisado há de se considerar a complexidade e o tamanho da escala proposta, encontram-se na literatura algumas sugestões que orientam normalmente esta decisão, destacando-se a proposta de Fleiss (1981):

<.40	Pobre
.40 a .75	Satisfatório e Bom
>.75	Excelente

4. Resultados e Discussões

Neste item serão demonstrados nas próximas seções os resultados dos indicadores de confiabilidade e concordância da amostra entre os avaliadores, incluindo os dois modelos analítico utilizados pela instituição financeira. Em relação aos indicadores *kappa*, foi realizada a testagem com todos os três indicadores para mostrar a evolução deles e a adequação da pesquisa ao indicador *kappa* ponderado *kw*, que leva em consideração os pesos de discordâncias.

4.1 Alfa de Cronbach em Word2Vec e BERT

Foi realizado o teste estatístico de confiabilidade da escala utilizada no estudo, conhecido como Coeficiente Alfa de Cronbach, modelo que propõe avaliar a correlação média entre as concordâncias dos *posts* obtidos da base amostral de interações e as três avaliações (funcionário A, Funcionário B e Modelo Analítico), conforme Tabela 1.

Tabela 2 – Estatísticas de confiabilidade

	Cohen Kappa	Fleiss Kappa	Kappa ponderado	N de itens
Word2Vec	,4090	,3890	,5140	3
BERT	,4962	,4920	,5500	3

Fonte: Elaborado pelos Autores

O valor mínimo aceitável para o Alfa é 0,70; abaixo desse valor a consistência interna da escala utilizada é considerada baixa. Em contrapartida, o valor máximo esperado é 0,90; acima deste valor, pode-se considerar que há redundância ou duplicação, ou seja, vários itens estão medindo exatamente o mesmo elemento de um constructo (multicolinearidade); portanto, os itens redundantes devem ser eliminados. Usualmente, são preferidos valores de Alfa entre 0,80 e 0,90 (STREINER, 2003).

Neste estudo, o resultado do coeficiente alfa de Cronbach 0.867 (Word2Vec) e 0,871 (BERT) figurou na faixa que indica uma consistência interna da escala quase perfeita, atendendo à preferência estabelecida por Streiner (2003) o que permite concluir que tal instrumento de medição apresentou alta confiabilidade no espaço onde foi aplicado.

4.2 Cohen *Kappa* – *k*

Medir a porcentagem de casos de concordância de dois observadores ou juízes que classificam um conjunto de itens num dado número de categorias, é talvez o método mais simples para se medir a fiabilidade das variáveis categoriais.

No entanto, este método não leva em conta a proporção desse acordo que é devida ao acaso. Cohen (1960), então desenvolveu o índice de κ como um coeficiente de concordância dos julgamentos de dois observadores visando preencher o hiato de assertividade dos modelos existentes.

O coeficiente de *Kappa* continua a ser amplamente utilizado, sendo que os pressupostos básicos apresentados pelo autor (COHEN, 1960, p. 38) para o seu cálculo são apresentados a seguir:

- (i) as unidades em análise são independentes; (ii) as categorias da escala nominal são independentes, mutuamente exclusivas e exaustivas; e (iii) os observadores atuam de forma independente. Cada juiz pode distribuir as unidades de análise pelas diferentes categorias livremente, partindo-se do princípio que ambos os juízes são considerados igualmente aptos para a realização da tarefa.

4.2.1 Resultados de Cohen *Kappa* - *k* em Word2VEC

Foi iniciada a construção das tabelas (matriz quadrada $n \times n$) de contingência da classificação dos observadores e o modelo analítico, visando identificar as frequências de concordância de avaliações entre os avaliadores com o sentido ordem de combinações de características de sentimentos de 1 a 9, já abordada no item 3 do estudo que mostrou a Metodologia do trabalho. A ordem de percepção e classificação dessas características obedece o sentido de estado de sentimento Negativo para Positivo.

A numeração de 1 a 9 segue a ordem de identificação características: 1) Negativa; 2) Negativa e Reclamação; 3) Negativa, Reclamação e Informação; 4) Negativa e Informação; 5) Negativa e Elogio; 6) Reclamação; 7) Reclamação e Informação; 8) Informação; 9) Elogio. Assim a construção da matriz consiste em colocar todas as classificações cruzadas, sendo que um observador/modelo figura na horizontal e outro na vertical.

Nas Tabelas 2, 3 e 4 a seguir, os valores contidos nas células diagonais, que estão em destaque, indicam o número de observações de concordância entre os avaliadores (por exemplo, o modelo analítico e o profissional A concordam com a característica Negativa em quatro posts na Tabela 2). Fora destas células estão contidos todos os apontamentos discordantes. O somatório de observações das linhas e o somatório de observações das

colunas indicam o total de observações, ou seja, os 200 posts/interações dos clientes com as mídias sociais da instituição financeira.

À medida que as células vão se afastando das células da diagonal (concordância), o grau de discordância vai crescendo. Desta forma quanto maior for o número de discordâncias próximos à diagonal, maior será o índice de Cohen *Kappa*.

Tabela 3 – Matriz quadrada de frequências obtidas entre Ferramenta x Profissional A

	Características	Ferramenta									Total
		1	2	3	4	5	6	7	8	9	
Profissional A	1	2	16	0	0	0	0	0	0	0	18
	2	2	10	0	0	0	1	2	0	0	15
	3	0	0	1	0	0	0	0	0	0	1
	4	0	0	0	0	0	1	0	0	0	1
	5	0	0	0	0	0	0	0	0	0	0
	6	6	23	0	0	0	11	1	3	0	44
	7	3	3	0	0	0	4	1	4	0	15
	8	8	9	1	3	0	6	0	52	0	79
	9	0	6	0	0	1	0	0	0	20	27
Total		21	67	2	3	1	23	4	59	20	200

Fonte: Elaborado pelo Autor

Tabela 4 – Matriz quadrada de frequências obtidas entre Ferramenta x Profissional B

	Características	Ferramenta									Total
		1	2	3	4	5	6	7	8	9	
Profissional B	1	4	29	0	0	0	3	0	0	0	36
	2	0	1	0	0	0	0	0	0	0	1
	3	0	0	0	0	0	0	0	0	0	0
	4	0	0	1	0	0	1	0	0	0	2
	5	0	0	0	0	0	0	0	0	0	0
	6	9	26	1	0	0	8	3	13	0	60
	7	3	3	0	0	0	5	1	2	0	14
	8	4	3	0	3	0	5	0	44	0	59
	9	1	5	0	0	1	1	0	0	20	28
Total		21	67	2	3	1	23	4	59	20	200

Fonte: Elaborado pelo Autor

Tabela 5 – Matriz quadrada de frequências obtidas entre Profissional A x Profissional B

	Características	Profissional B									Total
		1	2	3	4	5	6	7	8	9	
Profissional A	1	15	0	0	0	0	3	0	0	0	18
	2	11	0	0	0	0	4	0	0	0	15
	3	0	0	0	1	0	0	0	0	0	1
	4	1	0	0	0	0	0	0	0	0	1
	5	0	0	0	0	0	0	0	0	0	0
	6	8	1	0	1	0	30	1	2	1	44
	7	0	0	0	0	0	7	7	1	0	15
	8	1	0	0	0	0	15	6	56	1	79
	9	0	0	0	0	0	1	0	0	26	27
Total		36	1	0	2	0	60	14	59	28	200

Fonte: Elaborado pelo Autor

Para obter o total de frequências observadas fa (frequência alcançada) de concordâncias é realizada a soma de todos os valores contidos nas células diagonais. E para obter o total de frequências esperadas fe (concordância devido ao acaso), somado o produto entre o total da linha pelo total da coluna correspondente a cada célula das tabelas acima e dividindo-se esse produto pelo total de observações. Assim obtêm-se o $kappa$ para cada dupla avaliador/modelo e inter-avaliador, conforme orientado no capítulo 3:

Inter Modelo Analítico x Avaliador A

$$\sum fa = 2+10+1+11+1+52+20=97$$

$$\sum fe = ([\frac{18 \times 21}{200}] + [\frac{15 \times 67}{200}] + [\frac{1 \times 2}{200}] + [\frac{1 \times 3}{200}] + [\frac{44 \times 23}{200}] + [\frac{15 \times 4}{200}] + [\frac{79 \times 59}{200}] + [\frac{27 \times 20}{200}])=34,89$$

$$k = \frac{\sum fa - \sum fe}{N - \sum fe} = \frac{97 - 34,89}{200 - 34,89} = \frac{62,11}{165,11} = \mathbf{0,3761}$$

Inter Modelo Analítico x Avaliador B

$$\sum fa = 4+1+8+1+44+20=78$$

$$\sum fe = ([\frac{36 \times 21}{200}] + [\frac{1 \times 67}{200}] + [\frac{0 \times 2}{200}] + [\frac{2 \times 3}{200}] + [\frac{60 \times 23}{200}] + [\frac{14 \times 4}{200}] + [\frac{59 \times 59}{200}] + [\frac{28 \times 20}{200}])=31,53$$

$$k = \frac{\sum fa - \sum fe}{N - \sum fe} = \frac{78 - 31,53}{200 - 31,53} = \frac{46,47}{168,47} = \mathbf{0,2758}$$

Inter Avaliador A x Avaliador B

$$\sum fa = 15+30+7+56+26=134$$

$$\sum fe = ([\frac{18 \times 36}{200}] + [\frac{15 \times 1}{200}] + [\frac{1 \times 0}{200}] + [\frac{1 \times 2}{200}] + [\frac{44 \times 60}{200}] + [\frac{15 \times 14}{200}] + [\frac{79 \times 59}{200}] + [\frac{28 \times 27}{200}])= 44,66$$

$$k = \frac{\sum fa - \sum fe}{N - \sum fe} = \frac{134 - 44,66}{200 - 44,66} = \frac{89,34}{155,34} = \mathbf{0,5751}$$

O indicador k obtido, revela um índice de concordância muito baixo entre os avaliadores humanos e o modelo analítico. Mas cabe ressaltar que neste caso o peso relativo de cada discordância de característica de avaliação de sentimentos não está sendo considerado, e isso acaba influenciando o resultado.

Como era de se esperar, observa-se que o indicador k resulta em uma taxa maior de concordância entre os dois avaliadores se comparado à concordância entre avaliadores e o modelo. Canhoto e Padmanabban (2015) destacam que de uma forma geral, a discrepância

entre o índice de concordância da análise manual e automatizada fica evidente no julgamento de estados de sentimentos, embora isso possa variar para estados de sentimentos positivos.

Destaca-se que a característica de sentimentos obtida pela ferramenta “Negativa e Elogio” não foi utilizada por nenhum dos observadores humanos para classificar as interações, o que faz supor que o antagonismo sintático da combinação dessas duas características julgadas simultaneamente pelo modelo, influenciou o resultado negativamente.

Um dos motivos do baixo desempenho do índice de concordância entre os observadores e o modelo analítico, pode ser atribuído à baixa concordância na característica (2) Negativa e Reclamação, que foi a que obteve a maior quantidade de julgamentos por parte do modelo analítico, totalizando 67 vezes. Já os observadores A e B utilizaram esta apenas 15 e uma vez respectivamente.

4.2.2 Resultados de Cohen Kappa - k em BERT

Novamente foram construídas as matrizes quadradas de concordância entre as duplas de avaliadores/Modelo Analítico, porém agora com 15 características de sentimentos: 1) Ofensa, 2) Ofensa e Negativa, 3) Ofensa, Negativa e Reclamação, 4) Ofensa e Reclamação, 5) Ofensa, Reclamação e Informação, 6) Negativa, 7) Negativa e Reclamação, 8) Negativa, Reclamação e Informação, 9) Reclamação, 10) Reclamação e Informação, 11) Reclamação e Elogio, 12) Neutro, 13) Informação, 14) Informação e Elogio, 15) Elogio.

Tabela 6 – Matriz quadrada de frequências esperadas entre Ferramenta x Profissional A

	Características	Ferramenta															Total
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
Profissional A	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
	2	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	4	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	6	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
	7	0	0	0	1	1	0	0	2	0	1	0	0	1	0	0	6
	8	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
	9	0	0	0	0	0	0	0	1	0	1	0	1	3	0	0	6
	10	0	0	0	1	0	0	1	0	0	5	0	1	8	0	0	16
	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	12	0	0	0	0	0	0	0	0	0	0	0	53	4	0	2	59
	13	0	0	0	0	0	0	0	0	1	5	0	16	70	0	0	92
	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	4
	15	0	0	0	0	0	0	0	0	0	0	1	2	0	1	8	12
Total		0	0	2	2	2	0	1	3	2	13	1	73	86	1	14	200

Fonte: Elaborado pelo autor

Tabela 7 – Matriz quadrada de frequências esperadas entre Ferramenta x Profissional B

	Ferramenta																Total
Características	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15		
Profissional B	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	2	0	0	2	1	1	0	0	1	0	0	0	1	0	0	0	6
	3	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	3
	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	6	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
	7	0	0	0	0	0	0	1	1	1	6	0	3	3	0	0	15
	8	0	0	0	0	0	0	0	0	0	3	0	0	1	0	0	4
	9	0	0	0	0	0	0	0	0	0	2	0	3	5	0	0	10
	10	0	0	0	1	0	0	0	0	0	1	0	4	19	0	0	25
	11	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
	12	0	0	0	0	0	0	0	0	0	0	0	45	3	0	1	49
	13	0	0	0	0	0	0	0	0	0	1	0	15	53	0	0	69
	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	3
	15	0	0	0	0	0	0	0	0	0	0	1	2	0	1	10	14
Total	0	0	2	2	2	0	1	3	2	13	1	73	86	1	14	200	

Fonte: Elaborado pelo autor

Tabela 8 – Matriz quadrada de frequências esperadas entre Ferramenta x Profissional A

	Profissional A																Total
	Características	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
Profissional B	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	2	1	0	0	0	0	1	1	0	1	0	0	1	0	0	0	5
	3	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	2
	4	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	6	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
	7	0	0	0	0	0	0	4	0	3	5	0	0	4	0	0	16
	8	0	0	0	0	0	0	0	1	0	1	0	0	1	0	0	3
	9	0	0	0	0	0	0	0	0	1	2	0	0	7	0	0	10
	10	0	0	0	0	0	0	0	0	0	6	0	3	16	0	0	25
	11	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
	12	0	0	0	0	0	0	0	0	0	0	0	46	2	0	1	49
	13	0	0	0	0	0	0	0	0	1	0	0	7	61	0	0	69
	14	0	0	0	0	0	0	0	0	0	0	0	0	1	3	0	4
	15	0	0	0	0	0	0	0	0	0	0	0	0	2	0	1	11
Total		1	1	1	0	0	1	6	1	7	15	0	59	92	4	12	200

Fonte: Elaborado pelo autor

Adiante será repetido o mesmo cálculo da seção do modelo anterior para ser obtido o índice *Cohen Kappa* entre as duplas de observadores/modelo analítico:

Modelo Analítico x Observador A

$$\sum fa = 5+53+70+8=136$$

$$\sum fe = \left(\left[\frac{1x0}{200} \right] + \frac{1x0}{200} + \frac{0x2}{200} + \frac{1x2}{200} + \frac{0x2}{200} + \frac{1x0}{200} + \frac{6x1}{200} + \frac{1x3}{200} + \frac{6x2}{200} + \frac{16x13}{200} + \frac{0x1}{200} + \frac{59x73}{200} + \frac{92x86}{200} + \frac{4x1}{200} + \frac{12x14}{200} \right) = 63,09$$

$$k = \frac{\sum fa - \sum fe}{N - \sum fe} = \frac{136 - 63,09}{200 - 63,09} = \frac{72,91}{136,91} = \mathbf{0,5325}$$

Modelo Analítico x Observador B

$$\sum fa = 1+1+45+53+10=110$$

$$\sum fe = \left(\left[\frac{0x0}{200} \right] + \frac{6x0}{200} + \frac{3x2}{200} + \frac{0x2}{200} + \frac{1x0}{200} + \frac{15x1}{200} + \frac{4x3}{200} + \frac{1x3}{200} + \frac{10x2}{200} + \frac{25x13}{200} + \frac{1x1}{200} + \frac{49x73}{200} + \frac{69x86}{200} + \frac{3x1}{200} + \frac{14x14}{200} \right) = 50,46$$

$$k = \frac{\sum fa - \sum fe}{N - \sum fe} = \frac{110 - 50,46}{200 - 50,46} = \frac{59,54}{149,54} = \mathbf{0,3981}$$

Observador A x Observador B

$$\sum fa = 4+1+1+6+46+61+3+11=133$$

$$\sum fe = \left(\left[\frac{0x1}{200} \right] + \frac{5x1}{200} + \frac{2x1}{200} + \frac{1x0}{200} + \frac{0x1}{200} + \frac{1x1}{200} + \frac{16x6}{200} + \frac{3x7}{200} + \frac{10x15}{200} + \frac{25x15}{200} + \frac{1x0}{200} + \frac{49x59}{200} + \frac{69x92}{200} + \frac{4x4}{200} + \frac{14x12}{200} \right) = 50,36$$

$$k = \frac{\sum fa - \sum fe}{N - \sum fe} = \frac{134 - 50,36}{200 - 50,36} = \frac{83,54}{149,64} = \mathbf{0,5582}$$

Neste caso, o indicador *Cohen Kappa* evoluiu em concordâncias entre observadores e o modelo BERT em relação ao modelo Word2Vec. Não obstante alcançar limiares de concordância satisfatórios percebe-se nas matrizes cruzadas (tabelas 5, 6 e 7) que as classificações se concentram nas características Neutras e Informações. Nestes casos, Frankenthal (2017), relata que as empresas possuem dificuldade de obter *insights* com itens neutros como "indiferente". Nesse caso, a resposta oferece poucas informações aplicáveis de

maneira prática. Frankenthal ainda afirma que muitas vezes os observadores não realizam a análise de seus sentimentos em relação ao *post*, podendo existir inclusive uma espécie de automatismo e impulsividade para as respostas neutras.

Outro ponto a se destacar é que mesmo aumentando a quantidade de características combinadas neste modelo, passando para 15 alternativas de escolha, o indicador *Cohen Kappa* de concordância melhorou de patamar, o que pode sugerir que tanto a evolução do modelo analítico, como a melhora das definições conceituais de cada característica contribuiu para alcançar este resultado.

4.3 Fleiss Kappa

Apesar do κ de Cohen ser o índice de concordância mais utilizado entre avaliadores e referido na literatura, quando as variáveis são nominais, ou seja, quando não se utiliza nenhuma ponderação, já que não se aplica a variáveis ordinais, a sua eficácia também pode ser considerada restrita às situações em que somente dois avaliadores julgam as características de estado sentimento.

Desta forma, Fleiss (1971) sugere um desenvolvimento do κ para o caso em que, existindo múltiplos avaliadores, diferentes sub-grupos avaliam diferentes unidades de análise, mantendo constante o número de codificações por unidade de análise, como nesta pesquisa.

Seja n = o número de interações (200), k = o número de categorias de avaliação (9) e m = o número de avaliadores para cada interação (3), para cada interação $i=1, 2, 3, \dots, n$ e categorias de avaliação $j=1, 2, 3, \dots, k$, seja x_{ij} o número de avaliadores que atribuem a categoria j à interação i , portanto:

$$0 \leq x_{ij} \leq m \quad \sum_{j=1}^k x_{ij} = m \quad \sum_{i=1}^n \sum_{j=1}^k x_{ij} = mn$$

A proporção de pares de avaliadores que concordam em sua avaliação da interação i é dada por:

$$p_i = \frac{\sum_{j=1}^k C(x_{ij}, 2)}{C(m, 2)} = \frac{\sum_{j=1}^k x_{ij}(x_{ij} - 1)}{m(m - 1)} = \frac{\sum_{j=1}^k x_{ij}^2 - \sum_{j=1}^k x_{ij}}{m(m - 1)} = \frac{\sum_{j=1}^k x_{ij}^2 - m}{m(m - 1)}$$

A média do p_i é obtida por:

$$p_\alpha = \bar{p} = \frac{1}{n} \sum_{i=1}^n \frac{\sum_{j=1}^k x_{ij}^2 - m}{m(m - 1)} = \frac{1}{mn(m - 1)} \left[\sum_{i=1}^n \sum_{j=1}^k x_{ij}^2 - m \right]$$

Usando a seguinte medida para o termo de erro:

$$p_{\varepsilon} = \sum_{j=1}^k x_{ij}$$

Onde

$$q_j = \frac{1}{nm} \sum_{j=1}^k q_j^2$$

Desta forma o *fleiss kappa* é definido:

$$k = \frac{p_a - p_{\varepsilon}}{1 - p_{\varepsilon}}$$

4.3.1 Resultados de Fleiss kappa em Word2Vec

Devido ao fato de as demonstrações individualizadas dos cálculos serem muito extensas para serem visualizadas aqui, e este estudo ter como abordagem principal, a temática de marketing, o cálculo do *Fleiss Kappa* foi efetuado de forma automatizada na ferramenta IBM SPSS Statistics Subscription.

Tabela 9 – Cálculo integral Fleiss Kappa – Word2Vec

Overall Kappa					
	Kappa	Asymptotic Standard Error	Z	P Value	Lower 95% Asymptotic CI Bound
Overall	,389	,020	19,838	,000	,351
					Upper 95% Asymptotic CI Bound
					,427

Fonte: Extraído do cálculo efetuado pela ferramenta IBM SPSS Statistics Subscription

Na tabela 9, foi obtida a concordância simultânea dos três avaliadores (2 humanos e a ferramenta) por meio do *Fleiss Kappa* com Nível de Cobertura de CI em 95%. Observa-se um resultado de concordância ainda insatisfatório (0,389), mesmo considerando que neste modelo ainda não estão sendo ponderados os pesos de discordâncias das características de avaliação de sentimentos.

A tabela 10 demonstra uma forma individualizada de avaliar em quais categorias de avaliação ou características de sentimentos obteve-se a melhor *performance* entre as nove avaliadas, ou quais delas prejudicaram o nível de concordância.

Tabela 10 – Fleiss Kappa por Categorias individuais em Word2Vec

Rating Category	Conditional Probability	Kappa	Asymptotic Standard Error	Z	P Value	Lower 95% Asymptotic CI Bound	Upper 95% Asymptotic CI Bound
1	,280	,177	,041	4,339	,000	,097	,257
2	,133	-,007	,041	-,165	,869	-,087	,073
3	,333	,330	,041	8,083	,000	,250	,410
4	,000	-,010	,041	-,247	,805	-,090	,070
5	,000	-,002	,041	-,041	,967	-,082	,078
6	,386	,221	,041	5,411	,000	,141	,301
7	,273	,230	,041	5,644	,000	,150	,310
8	,772	,660	,041	16,164	,000	,580	,740
9	,880	,863	,041	21,136	,000	,783	,943

Fonte: Extraído do cálculo efetuado pela ferramenta IBM SPSS Statistics Subscription

Nota-se que somente as categorias “informação” e “Elogio”, possuem o k que indicam concordâncias satisfatórias considerando-se uma escala nominal, o que pode sugerir ajustes na calibragem dos vetores com indicação de características de polaridade negativa. As concordâncias entre avaliadores e modelos analíticos tendem a ser maiores entre torno de sentimento positivo, conforme relata Canhoto e Padmanabban (2015), que evidencia que a situação ocorre também entre análise manual x análise automatizada.

4.3.2 Resultados de Fleiss kappa em BERT

Foi obtido o cálculo de Fleiss Kappa para os dois observadores e a ferramenta simultaneamente por meio da ferramenta IBM SPSS com Nível de Cobertura de CI em 95%, conforme tabela 11.

Tabela 11 - Cálculo integral Fleiss Kappa – BERT

Overall Kappa						
	Kappa	Asymptotic Standard Error	Z	P Value	Lower 95% Asymptotic CI Bound	Upper 95% Asymptotic CI Bound
Overall	,491	,022	22,017	,000	,447	,534

Fonte: Extraído do cálculo efetuado pela ferramenta IBM SPSS Statistics Subscription

O resultado indicou uma melhora do nível de concordâncias entre os dois avaliadores com o modelo analítico, de 0,389 em Word2Vec para 0,491 em BERT, mesmo considerando um aumento significativo do número de características, sugerindo que o modelo aperfeiçoou sua acurácia avaliativa em relação ao modelo Word2Vec.

Tabela 12 – Fleiss Kappa por Categorias individuais em BERT

Rating Category	Conditional Probability	Kappa	Asymptotic Standard Error	Z	P Value	Lower 95% Asymptotic CI Bound	Upper 95% Asymptotic CI Bound
1	,000	-,002	,041	-,041	,967	-,082	,078
2	,000	-,010	,041	-,247	,805	-,090	,070
3	,000	-,008	,041	-,206	,837	-,088	,072
4	,000	-,005	,041	-,123	,902	-,085	,075
5	,000	-,003	,041	-,082	,935	-,083	,077
6	,000	-,003	,041	-,082	,935	-,083	,077
7	,217	,186	,041	4,561	,000	,106	,266
8	,143	,133	,041	3,251	,001	,053	,213
9	,053	,022	,041	,530	,596	-,058	,102
10	,226	,151	,041	3,710	,000	,071	,231
11	,000	-,002	,041	-,041	,967	-,082	,078
12	,791	,700	,041	17,154	,000	,620	,780
13	,745	,566	,041	13,876	,000	,486	,646
14	,333	,323	,041	7,916	,000	,243	,403
15	,725	,705	,041	17,278	,000	,625	,785

Fonte: Extraído do cálculo efetuado pela ferramenta IBM SPSS Statistics Subscription

As seis primeiras características não formaram uma quantidade de concordância para que pudesse consolidar o indicador de concordância (Tabela 12).

4.4 Kappa Ponderado – Kw

Os indicadores de concordância entre os avaliadores vistos até aqui levam em consideração a discordância entre os avaliadores, mas não o grau de discordância. Isso é especialmente relevante quando as classificações são ordenadas como em uma escala *likert*, caso análogo ao que foi definido pela área de TI da instituição objeto deste estudo.

Landis et al (1977, p. 159-174), propõem uma abordagem unificada para a avaliação da concordância de observadores para dados categóricos, buscando expressar as quantidades que refletem o grau em que os observadores concordam entre si, de forma que as discordâncias sejam calculadas em proporções. Assim o resultado produz hipóteses relevantes a respeito do viés inter-observador no uso geral da escala de medição. E assim assume-se naturalmente que o peso de uma discordância em que há julgamentos de características opostas, como “Negativo” e “Elogio”, deve ser mais penalizado que, por exemplo, uma discordância de avaliadores entre “Negativa” e “Reclamação”.

4.4.1 Resultados de Kappa Ponderado – kw em Word2Vec

Desta forma o *Kappa* ponderado é calculado usando uma tabela predefinida de pesos que medem o grau de desacordo entre os dois avaliadores: quanto maior o desacordo, maior o peso. A tabela 13 demonstra os pesos de uma matriz simétrica com zeros na diagonal principal (ou seja, onde há concordância entre os avaliadores) e valores positivos fora da diagonal principal. Quanto mais distantes estiverem os julgamentos realizados entre os avaliadores, maiores serão os pesos atribuídos (0 a 8).

Tabela 13 – Matriz de peso linear de discordâncias – Word2Vec

CARACTERÍSTICAS	Negativa	Negativa e Reclamação	Negativa, Reclamação e Informação	Negativa e Informação	Negativa e Elogio	Reclamação	Reclamação e Informação	Informação	Elogio
Negativa	0	1	2	3	4	5	6	7	8
Negativa e Reclamação	1	0	1	2	3	4	5	6	7
Negativa, Reclamação e Informação	2	1	0	1	2	3	4	5	6
Negativa e Informação	3	2	1	0	1	2	3	4	5
Negativa e Elogio	4	3	2	1	0	1	2	3	4
Reclamação	5	4	3	2	1	0	1	2	3
Reclamação e Informação	6	5	4	3	2	1	0	1	2
Reclamação e Elogio	7	6	5	4	3	2	1	1	3
Informação	7	6	5	4	3	2	1	0	1
Elogio	8	7	6	5	4	3	2	1	0

Fonte: Elaborado pelo Autor

Já foi mostrado como isso é feito em *Kappa de Cohen*, onde foram reorganizadas as categorias de classificação das características de sentimento, da menos favorável para a mais favorável, para tornar as coisas um pouco mais claras. Foi utilizada uma ponderação linear, embora penalidades mais altas possam ser atribuídas, por exemplo, às avaliações Negativa $\times$ Elogio.

Um ponto a ressaltar aqui, é que o *kappa* ponderado (kw) só pode ser utilizado para dois avaliadores. Para calcular o kw simultaneamente entre os dois avaliadores e o modelo analítico, é preciso calcular o kw entre: Avaliador A $\times$ Avaliador B, Avaliador A $\times$ Ferramenta e Avaliador B $\times$ Ferramenta e após a obtenção destes índices, são calculadas as médias deles.

Após definir a matriz de peso linear de zero a oito, calcula-se a tabela de frequências esperadas (fe) para cada par de avaliadores, como já abordado na sessão Kappa de Cohen, ou seja, diferente da frequência alcançada (fa), a frequência esperada, prevê a exclusão dos

julgamentos dos avaliadores realizados ao acaso. Para isso foram utilizadas as frequências alcançadas (fa) nas Tabelas 2, 3 e 4 para se obter a frequência esperada (fe) em cada célula através do quociente do produto da soma das linhas (tl) com a soma das colunas (tc) pela soma total da matriz (ti).

$$fe_{n,n} = \frac{tl \times tc}{ti}$$

Usando essa fórmula foram construídas as tabelas 14, 15 e 16 que contém em cada célula a frequência esperada na proporção de avaliação e classificação dos 200 posts.

Tabela 14 – Matriz quadrada de frequências esperadas entre Ferramenta x Profissional A

	Características	Ferramenta									Total
		1	2	3	4	5	6	7	8	9	
Profissional A	1	1,9	6	0,2	0,3	0,1	2,1	0,4	5,3	1,8	18
	2	1,6	5	0,2	0,2	0,1	1,7	0,3	4,4	1,5	15
	3	0,1	0,3	0	0	0	0,1	0	0,3	0,1	1
	4	0,1	0,3	0	0	0	0,1	0	0,3	0,1	1
	5	0	0	0	0	0	0	0	0	0	0
	6	4,6	14,7	0,4	0,7	0,2	5,1	0,9	13	4,4	44
	7	1,6	5	0,2	0,2	0,1	1,7	0,3	4,4	1,5	15
	8	8,3	26,5	0,8	1,2	0,4	9,1	1,6	23,3	7,9	79
	9	2,8	9	0,3	0,4	0,1	3,1	0,5	8	2,7	27
Total		21	67	2	3	1	23	4	59	20	200

Fonte: Elaborado pelo Autor

Tabela 15 – Matriz quadrada de frequências esperadas entre Ferramenta x Profissional B

	Características	Ferramenta									Total
		1	2	3	4	5	6	7	8	9	
Profissional B	1	3,8	12,1	0,4	0,5	0,2	4,1	0,7	10,6	3,6	36
	2	0,1	0,3	0	0	0	0,1	0	0,3	0,1	1
	3	0	0	0	0	0	0	0	0	0	0
	4	0,2	0,7	0	0	0	0,2	0	0,6	0,2	2
	5	0	0	0	0	0	0	0	0	0	0
	6	6,3	20,1	0,6	0,9	0,3	6,9	1,2	17,7	6	60
	7	1,5	4,7	0,1	0,2	0,1	1,6	0,3	4,1	1,4	14
	8	6,2	19,8	0,6	0,9	0,3	6,8	1,2	17,4	5,9	59
	9	2,9	9,4	0,3	0,4	0,1	3,2	0,6	8,3	2,8	28
Total		21	67	2	3	1	23	4	59	20	200

Fonte: Elaborado pelo Autor

Tabela 16 – Matriz quadrada de frequências esperadas entre Profissional A x Profissional B

	Características	Profissional B									Total
		1	2	3	4	5	6	7	8	9	
Profissional A	1	3,2	0,1	0	0,2	0	5,4	1,3	5,3	2,5	18
	2	2,7	0,1	0	0,2	0	4,5	1,1	4,4	2,1	15
	3	0,2	0	0	0	0	0,3	0,1	0,3	0,1	1
	4	0,2	0	0	0	0	0,3	0,1	0,3	0,1	1
	5	0	0	0	0	0	0	0	0	0	0
	6	7,9	0,2	0	0,4	0	13,2	3,1	13	6,2	44
	7	2,7	0,1	0	0,2	0	4,5	1,1	4,4	2,1	15
	8	14,2	0,4	0	0,8	0	23,7	5,5	23,3	11,1	79
	9	4,9	0,1	0	0,3	0	8,1	1,9	8	3,8	27
Total		36	1	0	2	0	60	14	59	28	200

Fonte: Elaborado pelo Autor

Percebe-se nas três matrizes que as características de sentimento 3) Negativa, Reclamação e Informação; 4) Negativa e Informação; 5) Negativa e Elogio não alcançaram frequências esperadas de concordância, sugerindo a dificuldade de percepção de características com sentimentos análogos, como Negativa e Reclamação ou de forte antagonismo entre os sentimentos mútuos Negativa e Elogio.

O valor ponderado de *kappa* aos pares de observadores é calculado somando os produtos de todos os elementos da tabela de observação pelos pesos correspondentes, e dividindo pela soma dos produtos de todos os elementos da tabela de expectativas pelos pesos correspondentes. Como os pesos medem desacordo, o *Kappa* ponderado é então igual a 1 menos esse quociente. Utilizando a notação *Kappa de Cohen* em que p_{ij} são as probabilidades alcançadas, $e_{ij} = p_i q_j$ são as probabilidades esperadas e w_{ij} são os pesos (com $w_{ji} = w_{ij}$).

$$k_w = 1 - \frac{\sum_{i,j} w_{ij} p_{ij}}{\sum_{i,j} w_{ij} e_{ij}}$$

Desta forma os resultados do *kappa* ponderado – *Kw* aos pares de observadores foram obtidos conforme Tabela 17:

Tabela 17 – Cálculo do Kappa Ponderado – *kw*

Pares de Avaliadores	PA e Ferramenta	PB e Ferramenta	PA e P2	$\overline{k_w}$
<i>kw</i>	0,410773	0,421714	0,709801	0,514096032

Fonte: Elaborado pelos Autores

Os resultados entre o modelo analítico e os avaliadores A e B ficaram muito próximos, sendo que a concordância entre a ferramenta automatizada e o Avaliador A alcançou .41 e entre a ferramenta automatizada e o Avaliador B .42. Já entre os observadores humanos o *kw*

ficou acima de .71, demonstrando novamente, um índice de concordância satisfatório entre estes, o que acabou contribuindo por elevar o ***kw*** a .51 entre os três avaliadores simultaneamente (Observadores A e B e a ferramenta automatizada).

Apesar de se observar uma evolução no indicador de concordância entre os avaliadores utilizando o *kw*, ao se comparar com o *Cohen Kappa* e *Fleiss Kappa*, o indicador *kw* ainda se encontra no limiar inferior de resultado satisfatório.

Uma das causas que pode sugerir uma *performance* mediana no resultado de concordância pode ser atribuída à quantidade elevada de categorias de características de percepções. Essa situação naturalmente acarreta a redução da probabilidade de concordância entre os avaliadores.

4.4.2 Resultados de Kappa Ponderado – *kw* em BERT

Aqui será utilizada a mesma ponderação linear de discordância utilizada na seção anterior para construção da Matriz de peso linear de discordância de 0 a 14 do modelo BERT (tabela 18):

Tabela 18 - Matriz de peso linear de discordâncias – BERT

CARACTERÍSTICAS	Ofensa	Ofensa e Negativa	Ofensa, Negativa e Reclamação	Ofensa e Reclamação	Ofensa, Reclamação e Informação	Negativa	Negativa e Reclamação	Negativa, Reclamação e Informação	Reclamação	Reclamação e informação	Reclamação e Elogio	Neutro	Informação	Informação e Elogio	Elogio
Ofensa	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Ofensa e Negativa	1	0	1	2	3	4	5	6	7	8	9	10	11	12	13
Ofensa, Negativa e Reclamação	2	1	0	1	2	3	4	5	6	7	8	9	10	11	12
Ofensa e Reclamação	3	2	1	0	1	2	3	4	5	6	7	8	9	10	11
Ofensa, Reclamação, Informação	4	3	2	1	0	1	2	3	4	5	6	7	8	9	10
Negativa	5	4	3	2	1	0	1	2	3	4	5	6	7	8	9
Negativa e Reclamação	6	5	4	3	2	1	0	1	2		4	5	6	7	8
Negativa, Reclamação e Informação	7	6	5	4	3	2	1	0	1	2	3	4	5	6	7
Reclamação	8	7	6	5	4	3	2	1	0	1	2	3	4	5	6
Reclamação e Informação	9	8	7	6	5	4	3	2	1	0	1	2	3	4	5
Reclamação e Elogio	10	9	8	7	6	5	4	3	2	1	0	1	2	3	4
Neutro	11	10	9	8	7	6	5	4	3	2	1		2	2	3
Informação	12	11	10	9	8	7	6	5	4	3	2	1		1	2
Informação e Elogio	13	12	11	10	9	8	7	6	5	4	3	2	1	0	1
Elogio	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0

Fonte: Elaborada pelo autor

Após definir a matriz de peso linear de zero a 14, será repetida a metodologia de cálculo e a construção das matrizes quadradas de frequências esperadas (*fe*) para cada par de avaliadores, a fim de se obter os indicadores de concordâncias entre os observadores/modelo analítico de *kappa* ponderado em BERT.

Desta forma, obteve-se as frequências esperadas nas matrizes de concordâncias aos pares de observadores/modelo analítico BERT conforme tabelas 19, 20 e 21.

Essas matrizes demonstram células vazias de frequências esperadas (*fe*) nas características de sentimentos de 1 a 6, como observado no modelo anterior, sugerindo novamente que o excesso de características e suas combinações além de confundir os observadores, diminuindo a probabilidade de concordâncias, não proporciona insights precisos e claros, como por exemplo classificar um post como Reclamação e Elogio ao mesmo tempo.

Tabela 19 – Matriz quadrada de frequências esperadas entre Profissional A x Profissional B

	Caract	Ferramenta															Total
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
Profissional A	1	0	0	0	0	0	0	0	0	0	0,1	0	0,4	0,4	0	0,1	1
	2	0	0	0	0	0	0	0	0	0	0,1	0	0,4	0,4	0	0,1	1
	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	4	0	0	0	0	0	0	0	0	0	0,1	0	0,4	0,4	0	0,1	1
	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	6	0	0	0	0	0	0	0	0	0	0,1	0	0,4	0,4	0	0,1	1
	7	0	0	0,1	0,1	0,1	0	0	0,1	0,1	0,4	0	2,2	2,6	0	0,4	6
	8	0	0	0	0	0	0	0	0	0	0,1	0	0,4	0,4	0	0,1	1
	9	0	0	0,1	0,1	0,1	0	0	0,1	0,1	0,4	0	2,2	2,6	0	0,4	6
	10	0	0	0,2	0,2	0,2	0	0,1	0,2	0,2	1	0,1	5,8	6,9	0,1	1,1	16
	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	12	0	0	0,6	0,6	0,6	0	0,3	0,9	0,6	3,8	0,3	22	25	0,3	4,1	59
	13	0	0	0,9	0,9	0,9	0	0,5	1,4	0,9	6	0,5	34	40	0,5	6,4	92
	14	0	0	0	0	0	0	0	0,1	0	0,3	0	1,5	1,7	0	0,3	4
	15	0	0	0,1	0,1	0,1	0	0,1	0,2	0,1	0,8	0,1	4,4	5,2	0,1	0,8	12
Total		0	0	2	2	2	0	1	3	2	13	1	73	86	1	14	200

Fonte: Elaborada pelo autor

Tabela 20 – Matriz quadrada de frequências esperadas entre Profissional A x Profissional B

	Caract	Ferramenta															Total
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
Profissional B	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	2	0	0	0,1	0,1	0,1	0	0	0,1	0,1	0,4	0	2,2	2,6	0	0,4	6
	3	0	0	0	0	0	0	0	0	0	0,2	0	1,1	1,3	0	0,2	3
	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	6	0	0	0	0	0	0	0	0	0	0,1	0	0,4	0,4	0	0,1	1
	7	0	0	0,2	0,2	0,2	0	0,1	0,2	0,2	1	0,1	5,5	6,5	0,1	1,1	15
	8	0	0	0	0	0	0	0	0,1	0	0,3	0	1,5	1,7	0	0,3	4
	9	0	0	0,1	0,1	0,1	0	0,1	0,2	0,1	0,7	0,1	3,7	4,3	0,1	0,7	10
	10	0	0	0,3	0,3	0,3	0	0,1	0,4	0,3	1,6	0,1	9,1	11	0,1	1,8	25
	11	0	0	0	0	0	0	0	0	0	0,1	0	0,4	0,4	0	0,1	1
	12	0	0	0,5	0,5	0,5	0	0,2	0,7	0,5	3,2	0,2	18	21	0,2	3,4	49
	13	0	0	0,7	0,7	0,7	0	0,3	1	0,7	4,5	0,3	25	30	0,3	4,8	69
	14	0	0	0	0	0	0	0	0	0	0,2	0	1,1	1,3	0	0,2	3
	15	0	0	0,1	0,1	0,1	0	0,1	0,2	0,1	0,9	0,1	5,1	6	0,1	1	14
Total		0	0	2	2	2	0	1	3	2	13	1	73	86	1	14	200

Fonte: Elaborada pelo autor

Tabela 21 – Matriz quadrada de frequências esperadas entre Profissional A x Profissional B

	Caract	Profissional A															Total
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
Profissional B	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	0,2	0	0,2	0,4	0	1,5	2,3	0,1	0,3	5
	3	0	0	0	0	0	0	0,1	0	0,1	0,2	0	0,9	1,4	0,1	0,2	3
	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	6	0	0	0	0	0	0	0	0	0	0,1	0	0,3	0,5	0	0,1	1
	7	0,1	0,1	0	0,1	0	0,1	0,5	0,1	0,6	1,2	0	4,7	7,4	0,3	1	16
	8	0	0	0	0	0	0	0,1	0	0,1	0,2	0	0,9	1,4	0,1	0,2	3
	9	0,1	0,1	0	0,1	0	0,1	0,3	0,1	0,4	0,8	0	3	4,6	0,2	0,6	10
	10	0,1	0,1	0	0,1	0	0,1	0,8	0,1	0,9	1,9	0	7,4	12	0,5	1,5	25
	11	0	0	0	0	0	0	0	0	0	0,1	0	0,3	0,5	0	0,1	1
	12	0,2	0,2	0	0,2	0	0,2	1,5	0,2	1,7	3,7	0	14	23	1	2,9	49
	13	0,3	0,3	0	0,3	0	0,3	2,1	0,3	2,4	5,2	0	20	32	1,4	4,1	69
	14	0	0	0	0	0	0	0,1	0	0,1	0,3	0	1,2	1,8	0,1	0,2	4
	15	0,1	0,1	0	0,1	0	0,1	0,4	0,1	0,5	1,1	0	4,1	6,4	0,3	0,8	14
Total		1	1	0	1	0	1	6	1	7	15	0	59	92	4	12	200

Fonte: Elaborada pelo autor

Tabela 22 – Cálculo do Kappa Ponderado – kw

Pares de Avaliadores	Avaliador A Ferramenta	Avaliador B Ferramenta	Avaliadores A e B	$\overline{kw}$
kw	0,589775	0,473373	0,587045	0,550064

Fonte: Elaborada pelo autor

No modelo BERT o resultado de Kappa ponderado obtido (0,55) superou o Word2Vec (0,51). Se no modelo anterior os observadores alcançaram índices muito próximos de concordância com o modelo analítico, neste a concordância entre os avaliadores e ferramenta destoaram entre si, com o Observador A alcançando 0,58 e o Observador B 0,47 (Tabela 22).

Embora tenha ocorrido uma queda significativa do índice *Kappa* ponderado entre o par de observadores A e B em relação ao mesmo par de Word2Vec, não se pode afirmar que o modelo diminuiu sua capacidade preditiva, tendo em vista que a eficiência de concordância a ser buscada, é entre os observadores e o modelo analítico. E neste modelo o resultado de concordância entre observadores e BERT, superaram Word2Vec, mesmo com o aumento significativo de características de sentimentos.

4.5 Investigações das causas de discordâncias

Uma das causas que podem sugerir a falta de concordância entre os observadores e modelos analíticos pode ser atribuída ao tamanho da escala de características de sentimento nos dois modelos analíticos – Word2Vec e BERT, tendo 9 e 15 itens distribuídos entre a polaridade de sentimentos, respectivamente. Essa condição além de diminuir a probabilidade de concordâncias, também podem confundir os avaliadores que julgam as mensagens.

A escolha de uma escala de medida pode sugerir inicialmente que não seja um fator decisivo para o resultado da pesquisa, gerando uma falta de atenção para este item (Cummins; Gullone, 2000; Collings, 2006). No entanto, Dalmoro e Vieira (2013) alerta que quando o pesquisador se abdica de usar o senso comum e passa a se questionar sobre os aspectos da construção de uma escala, descobre que o desenvolvimento de uma escala de medida para um instrumento de pesquisa é uma tarefa complexa.

Devellis (1991), defende que deve-se evitar o uso de escala com formatos diferentes, principalmente com inversão do sentido num mesmo instrumento, visto que isto pode provocar a mudança de posição de alguns avaliadores/observadores. Buscando um constructo mais eficiente, as escalas também devem apresentar claramente um contínuo de direção.

Face as evidências apresentadas pela literatura sobre o tema, e os resultados deste estudo, propõe-se a seguir uma análise considerando apenas a primeira característica de cada uma das combinações, de modo que restem apenas as características individuais.

4.5.1 Análise da concordância utilizando apenas as características individuais em Word2Vec

Se for considerado na classificação dos três avaliadores, dois humanos e um automatizado, apenas a primeira característica de cada uma das nove combinações do modelo Word2Vec, de modo que restem apenas as 4 características primárias, Negativa, Reclamação, Informação e Elogio, têm-se indicadores de concordância mais satisfatórios como se vê nas células diagonais das tabelas 23, 24 e 25.

Tabela 23 – Contingência Observada entre Ferramenta x Profissional A

	Ferramenta					Total
Profissional A	Características	Negativa	Reclamação	Informação	Elogio	
	Negativa	31	4	0	0	35
	Reclamação	35	17	7	0	59
	Informação	21	6	52	0	79
	Elogio	7	0	0	20	27
	Total	94	27	59	20	200

Fonte: Elaborado pelos Autores

Tabela 24 – Contingência Observada entre Ferramenta x Profissional B

	Ferramenta					Total
Profissional B	Características	Negativa	Reclamação	Informaçã	Elogio	
	Negativa	35	4	0	0	39
	Reclamação	42	17	15	0	74
	Informação	10	5	44	0	59
	Elogio	7	1	0	20	28
	Total	94	27	59	20	200

Fonte: Elaborado pelos Autores

Tabela 25 – Contingência Observada entre Profissional A x Profissional B

	Profissional B				Total	
Profissional A	Características	Negativa	Reclamação	Informação		Elogio
Profissional A	Negativa	28	7	0	0	35
	Reclamação	10	45	3	1	59
	Informação	1	21	56	1	79
	Elogio	0	1	0	26	27
Total		39	74	59	28	200

Fonte: Elaborado pelos Autores

Na escala de polaridade de sentimentos foi estabelecida a ordem conforme a figura 2, classificando as características negativa e reclamação tendo uma percepção negativa, informação, uma percepção neutra e elogio, uma percepção positiva. A classificação de concordâncias de polaridade negativa, neutra e positiva alcançaram, 27%, 34% e 39% respectivamente em relação ao total das concordâncias entre os 3 observadores/modelo analítico, conforme gráfico 1.

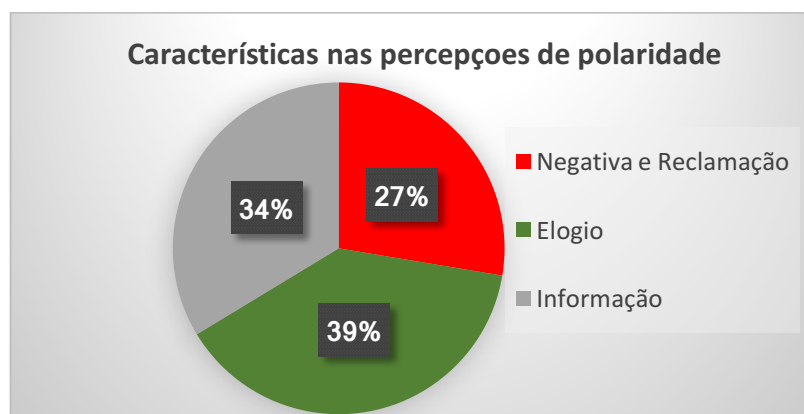
Figura 3 – Polaridade das características primárias em Word2Vec



Fonte: Elaborado pelo Autor

Assim, a percepção que obteve o maior percentual de acertos em relação ao total de concordâncias foi a positiva (elogio), seguida pela percepção neutra (informação) e depois pela percepção negativa (negativa e reclamação).

Gráfico 1 – Concordâncias nas polaridades Negativo – Neutro – Positivo em Word2Vec



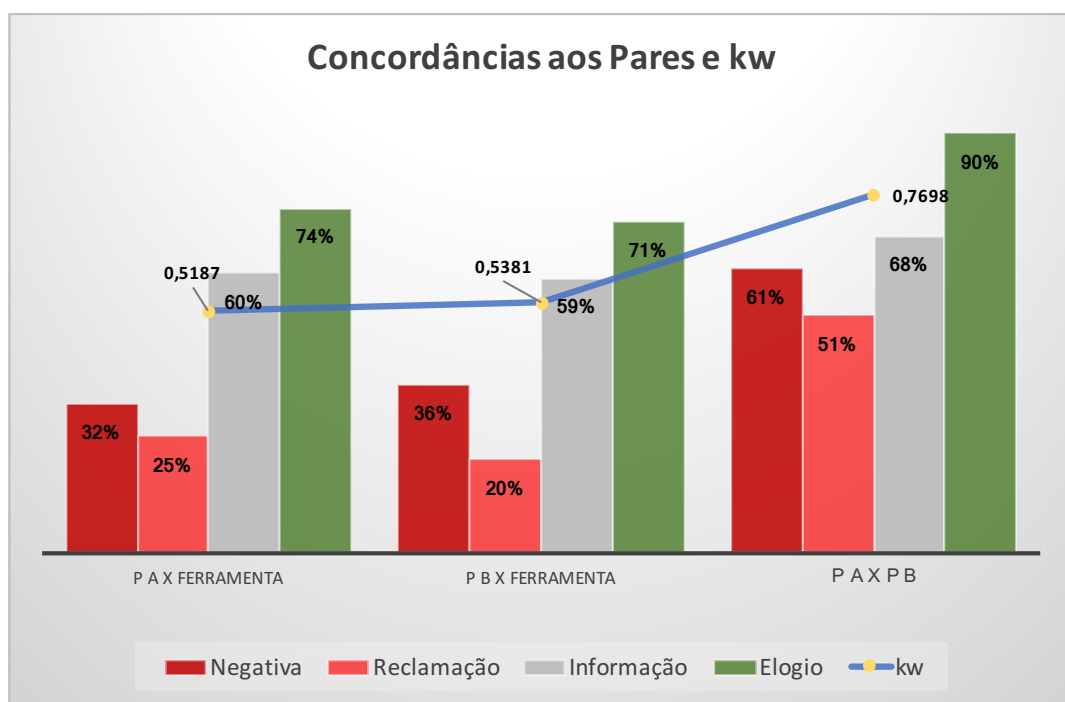
Fonte: Elaborado pelos Autores

Destoam na análise de concordâncias aos pares entre a ferramenta automatizada e os observadores A e B, conforme gráfico 3, as duas características que denotam uma percepção negativa. As características “Negativa” e “Reclamação” obtiveram a pior *performance* de concordâncias, sendo “reclamação” a característica em que os avaliadores menos concordaram entre si. Esta condição sugere que o modelo estatístico utilizado pela instituição

foi desafiado em diferenciar a dicotomia de significados entre as características “Negativa” e “Reclamação”, que pode ser causada pela ambiguidade e aproximação dos seus significados.

A característica “elogio” alcançou o maior percentual de concordância entre os avaliadores, refletindo o que também já foi verificado em outros estudos similares (Vanden Bergh et al., 2011; CANHOTO, 2015), que demonstram que os sentimentos e emoções positivas são mais fáceis de serem identificadas e classificadas.

Gráfico 2 – Distribuição de Concordâncias das Características de Sentimento e *kw* em Word2Vec



Fonte: Elaborado pelo Autor

O aumento do nível de concordância ocorreu adotando a redução do número de características de sentimentos de nove para quatro, onde o *kw* que passou de .51 quando eram nove características, para .61, quando se utilizou apenas as quatro características primárias, perfazendo uma evolução de 17,65%, conforme observa-se na tabela 26.

Tabela 26 – Cálculo do Kappa Ponderado – Características Primárias – k_w

Pares de Avaliadores	P1 e Ferramenta	P2 e Ferramenta	P1 e P2	$\overline{k_w}$
k_w	0,518747908	0,538155163	0,769861437	0,608921503

Fonte: Elaborado pelos Autores

Destaca-se o alto índice de concordância k_w alcançado entre os avaliadores/observadores humanos (.77), sugerindo mais uma vez que o teste de concordância entre estes pode fazer o papel de regra de ouro de calibragem para o modelo estatístico adotado pela instituição, para avaliar o sentimento nas mídias sociais da instituição financeira.

4.5.1.1 Exemplos de posts avaliados

As disparidades de avaliações ocorreram em diversos *posts*, de forma mais recorrente com os dois avaliadores humanos discordando da ferramenta em frases, que fora de um contexto mais amplo, possuem conotação positiva de senso, não obstante a ferramenta classificá-las como percepção negativa, conforme pode se observar na tabela 27.

Nos *posts* 34. “Banco do Povo”, 36, “Banco diferente”, 51, “Nossa, vocês são o melhor banco do mundo”, 189, “Vocês são o melhor Banco”, fica evidente, que apesar de serem frases curtas e relativamente fáceis de se interpretar como uma interação de percepção positiva sob a égide do senso humano, a ferramenta classificou-as como uma percepção negativa, o que pode sugerir que a palavra “banco” está localizada em uma dimensão de vetores de palavras de percepção negativa.

Tabela 27 – Exemplos de disparidade na classificação de sentimentos

<i>Posts</i>	Profissional A	Profissional B	Ferramenta
34. "Banco do Povo"	Elogio	Reclamação	Negativa
36. "Sim, muito obrigado pelo atendimento e pela paciência mais uma vez"	Elogio	Elogio	Negativa
37. "Banco diferente"	Elogio	Reclamação	Negativa
51. "Nossa, vocês são o melhor banco do mundo"	Elogio	Elogio	Negativa
59. "Olá, infelizmente o Correio mentiu. Todos as correspondências chegam aqui e o endereço estava corretíssimo. Alguns funcionários do correio, por preguiça de entregar dizem que o endereço está errado. Mas enfim, obrigado, irei na agência pegar"	Informação	Elogio	Reclamação
60. "Eeeeeeee..."	Reclamação	Elogio	Negativa
189. "Vocês são o melhor banco"	Elogio	Elogio	Negativa

Fonte: Elaborado pelo Autor

O *post* 36, “Sim, muito obrigado pelo atendimento e pela paciência mais uma vez”, foi classificada pelos avaliadores como um elogio. Já a ferramenta automatizada avaliou como uma interação negativa, sugerindo que uma única palavra-chave, provavelmente “atendimento”, pode influenciar a codificação de uma avaliação, que neste exemplo, foi classificada como negativa.

Já o *post* 59, “Olá, infelizmente o Correio mentiu, todas as correspondências chegam aqui e o endereço estava corretíssimo. Alguns funcionários do correio, por preguiça de entregar dizem que o endereço está errado. Mas enfim, obrigado, irei na agência pegar”, apesar de não expressar nenhum sentimento negativo em relação à instituição pesquisada, pode ter sido classificada assim pela ferramenta por se tratar de uma interação longa na qual mais de um sentimento foi expresso nela como as sintaxes negativas “infelizmente”, “mentiu”, “preguiça” e as positivas “corretíssimo”, “obrigado”. Este resultado da ferramenta automatizada realizando uma avaliação negativa (negativa e reclamação) ocorreu em diversas *posts* que foram classificadas de forma oposta pelos avaliadores humanos e estas interações possuíam mais de 10 palavras, mostrando a dificuldade da ferramenta em codificar em seu *data set* frases longas.

Algumas interações são bastante curtas e na ausência do contexto de um diálogo, não permitem uma classificação razoável, inclusive entre os profissionais, como pôde ser observado no *post* 60, “Eeeeeee”, quando um profissional classificou como Reclamação, o outro como Elogio e a ferramenta automatizada como Negativa.

Tabela 28 – Exemplos de consenso na classificação de sentimentos

Posts	Profissional A	Profissional B	Ferramenta
25. “VERGONHA (Nome da Instituição), com grandes empresas devedoras vocês fazem isso? Duvido!!!!!!!!!!!!!!”	Negativa	Negativa	Negativa
53. “Estou há 2 h no (Nome da Instituição) tentando encerrar minha conta e ainda não consegui. Isso por que é a segunda vez que venho. Ineficiência total!”	Negativa	Negativa	Negativa
54. “Pense em um banco com péssimo atendimento aos seus correntistas”	Negativa	Negativa	Negativa
12. “Vou na agência amanhã... mas já não quiseram me ajudar semana passada.. creio que amanhã também não vai”	Reclamação	Reclamação	Reclamação
22. “Não está correta a resposta”	Reclamação	Reclamação	Reclamação
41. “Quanto terem parcelado minha dívida sem minha autorização e cobrarem do meu salário depositado na conta”	Reclamação	Reclamação	Reclamação
121. “E agora o que eu faço? Como faço para cadastrar meu dispositivo (celular) novamente?”	Informação	Informação	Informação
131. “Quero saber se ele já se encontra na agência para ser retirado”	Informação	Informação	Informação
138. “Antecipar meu Pasep que recebo mês que vem”	Informação	Informação	Informação

181. “Muito bom, obrigada equipe (Nome da Instituição)!! Sempre ótimo atendimento.”	Elogio	Elogio	Elogio
187. “Rápido”	Elogio	Elogio	Elogio
192. “E parabéns pela eficiência”	Elogio	Elogio	Elogio

Fonte: Elaborado pelo autor

Nos exemplos de *posts* em que houve consenso (tabela 28) entre os dois avaliadores e a ferramenta automatizada, percebe-se que as características “Elogio” e “Informação” foram obtidas em frases relativamente mais curtas que “Negativa” e “Reclamação”.

Cabe destacar novamente que a ferramenta automatizada não avaliou a característica “Elogio” em frases que continham mais de 10 palavras, sugerindo novamente, o desafio de calibragem do modelo neural em textos longos que não necessariamente, possuem um resultado de interação negativo.

4.5.2 Análise da concordância utilizando apenas as características individuais em BERT

Aqui selecionamos também apenas a primeira característica de cada uma das quinze combinações do modelo BERT, de modo que restem apenas as 6 características primárias, Ofensa, Negativa, Reclamação, Neutro, Informação e Elogio, e construímos as matrizes quadradas de concordância entre os pares de observadores/modelo analítico, conforme as tabelas 29, 30 e 31.

Tabela 29 – Contingência Observada entre Ferramenta x Profissional A

	Características	Ferramenta						TOTAL
		Ofensa	Negativa	Reclamação	Neutro	Informação	Elogio	
Profissional A	Ofensa	2		1				3
	Negativa	3	2	2		1		8
	Reclamação	1	2	6	2	11		22
	Neutro				54	5	2	61
	Informação			6	16	68	4	94
	Elogio			1	2	1	8	12
TOTAL		6	4	16	74	86	14	200

Fonte: Elaborado pelos Autores

Tabela 30 – Contingência Observada entre Ferramenta x Profissional B

	Características	Ferramenta						TOTAL
		Ofensa	Negativa	Reclamação	Neutro	Informação	Elogio	
Profissional B	Ofensa	4	2	1	1			8
	Negativa	1	2	9	3	5		20
	Reclamação	1		3	8	23		35
	Neutro			1	45	3	1	50
	Informação			1	15	54	3	73
	Elogio			1	2	1	10	14
TOTAL		6	4	16	74	86	14	200

Fonte: Elaborado pelo Autor

Tabela 31 – Contingência Observada entre Ferramenta x Profissional A

		Profissional A						TOTAL
		Ofensa	Negativa	Reclamação	Neutro	Informação	Elogio	
Profissional B	Ofensa	3	3	1	1			8
	Negativa		5	10		5		20
	Reclamação			9	3	23		35
	Neutro			1	46	2	1	50
	Informação			1	9	63		73
	Elogio				2	1	11	14
		3	8	22	61	94	12	200

Fonte: Elaborado pelo Autor

Ao reduzir as combinações de características pode-se perceber também neste modelo o aumento de concordâncias nas células diagonais das matrizes quadradas de concordâncias aos pares.

A ordem de polaridade nesta escala do modelo BERT pode ser definida da seguinte forma: as características Ofensa, Negativa e Reclamação como uma interação negativa, Informação e Neutro como sendo uma interação que não denota emoções e Elogio como uma interação positiva, conforme a figura 3.

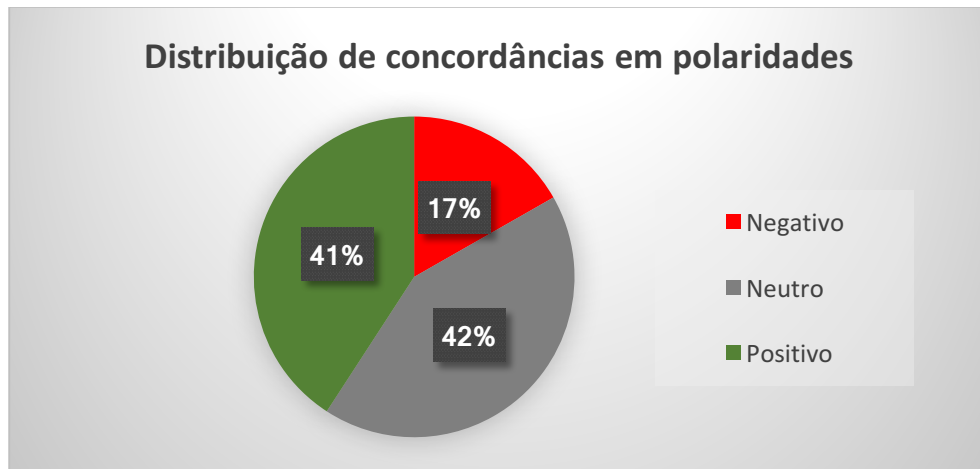
Figura 4 – Polaridade das características primárias em BERT



Fonte: Elaborado pelo Autor

A distribuição em percentuais da totalidade de concordâncias entre os dois observadores e o modelo analítico alcançou 17% nas características de polaridade Negativo (Ofensa, Negativa e Reclamação), 42% nas características neutro (Neutro e Informação) e 41% na única característica positivo (Elogio), conforme gráfico 4.

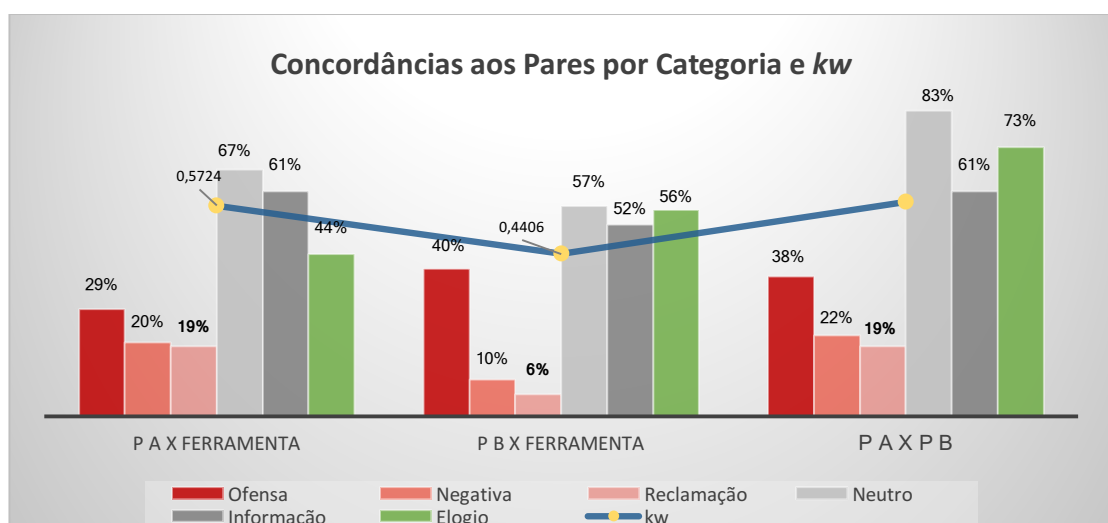
Gráfico 3 – Concordâncias nas polaridades Negativo – Neutro – Positivo em BERT



Fonte: Elaborado pelo autor

Mais uma vez repetiu-se a dificuldade observada em Word2Vec dos classificadores concordarem nas características de cunho negativo. No modelo BERT estas características participaram apenas em 17% do total de concordâncias nos 200 posts analisados. Neste caso o percentual de concordância resultou ainda pior, tendo em vista a desproporção da escolha das seis características por parte dos avaliadores nesta amostra de BERT. O número das classificações das características nas categorias Neutro e Positivo foram bem maiores que das características da categoria Negativo, conforme visto nas tabelas 27, 28 e 29.

Gráfico 4 – Distribuição de Concordâncias das Características de Sentimento e kw em BERT



Fonte: Elaborado pelo autor

Diferente das categorias de Word2Vec, foi incluída no modelo BERT a categoria de percepção Neutro para classificar aqueles posts que não tinham nenhuma relação com as

demais características de sentimento (Ofensa, Negativa, Reclamação, Informação e Elogio). E foi essa categoria que alcançou a melhor performance de concordância dentre as demais, o que contribuiu por elevar o índice de concordância kw , embora não ser esta categoria que diz respeito à percepção de sentimentos que gere insights dos clientes da instituição financeira.

As características Ofensa, Negativa e Reclamação que figuram na categoria negativa de polaridade ficaram com a pior *performance* entre os 3 pares de avaliadores, conforme gráfico 3, pelos motivos já relatados. Mas isso não deixa de serem propostos desafios aos profissionais da área de TI a buscarem a melhora da calibragem das etapas de *machine learning* para a categoria de polaridade Negativo, tendo em vista que mesmo o modelo Word2Vec tendo uma melhor proporcionalidade entre as categorias de polaridade, obteve também uma baixa concordância nesta categoria.

Tabela 32 - Cálculo do Kappa Ponderado – Características Primárias – kw

Pares de Avaliadores	A1 e Ferrament	A2 e Ferramenta	A1 e A2	kw
kw	0,5724	0,4406	0,5829	0,531966667

Fonte: Elabordo pelo autor

Embora as concordâncias tenham crescido entre os avaliadores quando foram reduzidas as características de 15 para 6, quando é considerado o peso entre as discordâncias em BERT o kw acaba sendo impactado e aqui, diferente de Word2Vec, ele ficou menor que o índice quando considerado as 15 características (Tabela 32). Isso ocorre neste modelo porque as características da categoria de polaridade negativo (Ofensa, Negativa e Reclamação) foram poucas utilizadas para classificar os posts, assim a ponderação de uma única discordância nessas características acaba tendo um peso maior (Gráfico 31).

Se compararmos o kw , o resultado de Word2Vec (0.60) acabou resultando em um índice melhor que BERT (0.53), contudo naquele modelo a amostra de *posts* não trouxe a discrepância entre as escolhas de características de sentimento, como neste último modelo. Assim a comparação de eficiência da medida de kw acaba se tornando inconclusiva, tendo em vista não haver a mesma proporcionalidade de características escolhidas pelos dois modelos analítico.

4.5.2.1 Exemplos de posts avaliados

Na tabela 33 estão listados os exemplos de posts em que houveram disparidades de classificação entre os três avaliadores visando identificar quais foram os termos ou palavras que causaram as discordâncias simultâneas.

Nos posts 9 e 115, o modelo classificou as interações como neutro e os observadores divergiram desta escolha. Porém não observamos aqui um erro na classificação do modelo e diante das definições de cada característica, essas interações podem ser avaliadas neutra em uma análise manual. Já no post 35 o modelo classificou a interação como Reclamação, mas pela frase em si, isso não está muito claro, em que pese um dos observadores ter classificado como negativa.

Tabela 33 – Exemplos de disparidade na classificação de sentimentos

Posts	Profissional A	Profissional B	Ferramenta
9. "Só posso migrar o Brasilprev. O restante não é questão judicial."	Informação	Reclamação	Neutro
115. "perfil do Banco do Brasil, tem o selo de verificado lá"	Informação	Reclamação	Neutro
35. "Fiz minhas contas e vai faltar 1600. Deu ruim bem ruim."	Informação	Negativa	Reclamação
81. "Boa tarde, gente estou tentando usar meu cartão no pic pay e não consigo"	Reclamação	Negativa	Informação
59. "Atrasei meu empréstimo e não negociam de maneira nenhuma preciso do meu nome mais não aceitam negociação"	Informação	Negativa	Negativa

Fonte: Elaborado pelo Autor

O post 81 foi classificado como Informação pelo modelo, mas conforme as definições da característica "Informação", e que além disso o cliente estar proferindo uma crítica a um produto que não está conseguindo utilizar, talvez fosse prudente classificar esta interação como uma reclamação, característica julgada por um dos observadores. Na interação 59 a ferramenta julgou um

negativa, divergindo em vários níveis da ordem escala com um dos observadores que a avaliou como informação. Porém neste caso pode-se sugerir que há as duas percepções nesta frase, pois entende-se que o cliente busca também uma solução do seu problema

De uma maneira geral o modelo melhorou muito em relação às disparidades de classificação observadas no modelo Word2Vec, tendo em vista os julgamentos dos *posts* não terem causado nenhuma situação antagônica de classificação de sentimentos.

Tabela 34 – Exemplos de consenso na classificação de sentimentos

Posts	Profissional A	Profissional B	Ferramenta
1. “Atendimento humano horrível, nunca solucionado o problema. Não recomendo a ninguém esse banco e o aplicativo então... Foi desenvolvido por uma criança de 6 anos pelo jeito.”	Ofensa	Ofensa	Ofensa
179. "Quanta incompetência Edilson!"	Ofensa	Ofensa	Ofensa
112. "Mandeí mensagem no WhatsApp, Facebook, Instagram e fiz uma reclamação no Reclame Aqui, e ninguém responde, péssimo atendimento ao cliente.”	Negativa	Negativa	Negativa
177. "Eu ligo para agência e não sou retornada"	Reclamação	Reclamação	Reclamação
183. "Já falei com 3 pessoas, cada um me transfere para outro setor”	Reclamação	Reclamação	Reclamação
8. “Onde a gente acompanha essa premiação e como é a extração dos números”	Neutro	Neutro	Neutro
44. “Já confirmei. Obrigado!”	Neutro	Neutro	Neutro
2. “Bom dia. Consegui ver a questão do carro”	Informação	Informação	Informação
6. “Por favor eu quero saber o que é preciso para aumentar o limite do cartão de crédito?”	Informação	Informação	Informação
15. “Parabéns Banco do Brasil vocês são foda, único banco de verdade que acredita no mundo dos games e acreditou na tribo. #tamojuntosnessegame.”	Elogio	Elogio	Elogio
187. “Olá minha melhor gerente do mundo, estamos bem! Graças a Deus, saudades de vc!”	Elogio	Elogio	Elogio

Fonte: Elaborado pelo autor

Na tabela 34 estão listados alguns dos diversos posts em que houve o consenso entre os observadores e o modelo BERT.

Cabe destacar que o modelo conseguiu identificar mensagens de despedida de atendimento, que contém por exemplo, obrigado, como uma interação neutra sem julgar como uma interação positiva.

5. Conclusão

Captar as emoções é fundamental para explicar e antecipar o comportamento do consumidor. Isto tem sido obtido por meio da análise de sentimentos que oferece aos profissionais de marketing uma maneira de medir e resumir essas emoções. As emoções exibidas nas conversas das mídias sociais, em particular, são muito promissoras para pesquisas, pois essas plataformas oferecem muitas oportunidades para ouvir e compreender as conversas em tempo real, com o mínimo de interrupção e de indivíduos que expressam essas emoções de uma forma muito mais eficiente e econômica que em outros meios de comunicação.

A análise de sentimentos das conversas nas mídias sociais tem se tornado muito popular, com promessas alvissareiras de fiabilidade, não obstante algumas pesquisas indicarem que não se trata de um processo simples e direto.

O problema desta pesquisa foi verificar a acurácia das análises automatizadas de sentimentos de clientes que interagiram com seis mídias sociais de uma grande instituição financeira. Diferente da coleta e análise de dados quantitativos, por exemplo, o estudo de correlação entre duas variáveis – onde existem padrões bem estabelecidos para processar e analisar dados, o estudo das emoções está inserido em nuances de subjetividade.

Durante a discussão dos resultados, foi verificado que em diversos *posts* algumas palavras foram associadas a emoções opostas do que realmente significavam, até mesmo pelos avaliadores humanos. Houve também palavras que não denotavam qualquer tipo de emoção, influenciando e determinando a classificação de característica de sentimentos sem conexão causal.

Estes desafios são acentuados pelo fato de que os segmentos de textos disponíveis nas mídias sociais são muito curtos, ricos em abreviações e gírias, e geralmente com erros de digitação ou erros gramaticais.

Foram utilizadas as ferramentas estatísticas mais difundidas em testes de concordância entre avaliadores para verificar a qualidade de classificação realizada pela análise automatizada: *Alfa de Cronbach*, *Kohen Kappa*, *Fleiss Kappa*, *Weighted Kappa*.

A quantidade elevada de características de sentimentos combinadas sugere um impacto desfavorável nos indicadores de combinação entre avaliadores, mesmo tratando-se do índice escalar de concordância *Kappa* ponderado *kw*, que relativiza os pesos de discordâncias. Contudo ao se reduzir somente às características primárias como passíveis de serem

escolhidas pelos avaliadores, o modelo Word2Vec melhorou sua performance, mas o modelo BERT obteve uma pequena redução do indicador kw , muito embora isso tenha ocorrido pela falta de proporcionalidade das classificações de características de sentimento.

Por estarem muito próximas do ponto de vista lexical, as duas características de percepção de polaridade negativo, “negativa” e “reclamação”, no modelo Word2Vec “ofensa”, “negativa” e “reclamação” foram as que obtiveram as maiores discordâncias entre os avaliadores, inclusive entre os observadores A e B.

Para melhorar o índice de concordância na classificação das interações, os analistas devem além de reduzir as características a serem escolhidas, realizar análise periódica em recortes de amostras do *data set* avaliado pela ferramenta automatizada, visando realizar calibragem do campo vetorial das palavras-chave. Além disso, é importante continuar desenvolvendo dicionários específicos que reflitam a sintaxe e o estilo específicos usados nas conversas de mídia social, na primeira etapa da análise, devem substituir as abreviações comumente usadas com seu equivalente formal - por exemplo, substituir SQN por ‘Só Que Não’.

É preciso reconhecer que, como linguagem e comunicação, os estilos estão em constante evolução. Esses dicionários e ferramentas nunca refletem completamente todas as variações e nuances nas comunicações das mídias sociais. Além disso, a ferramenta precisa buscar formas de capturar o sarcasmo, ironia e também usos altamente contextualizados da linguagem - por exemplo, adolescentes que usam o termo “irado” para se referir a uma experiência muito boa.

Deve-se relevar também que os modelos utilizados pela instituição estão sendo constantemente atualizados e treinados e que algumas das situações destacadas aqui podem ter sido abordadas nas versões da ferramenta posteriores a nossa pesquisa. Verifica-se também que a quantidade de dados captados e tratados pela instituição permite que o modelo analítico possa “aprender” dicionários bem como implantar constantemente novas técnicas lexicais.

A pesquisa demonstrou também, como já observado em outros trabalhos, que as características neutras e positivas possuem um nível de concordância superior às negativas, sugerindo que o modelo analítico deve obter uma calibragem mais acurada para essas últimas.

Por fim, ficou mais evidenciado que as mídias sociais oferecem uma janela para as mentes dos consumidores e prometem diversos desenvolvimentos da pesquisa de comportamento do consumidor. Em particular, a análise das conversas na mídia oferece muitas vantagens sobre métodos alternativos para estudar emoções dos consumidores. No

entanto, como essa pesquisa mostrou, a capacidade do pesquisador de identificar com precisão o sentimento expresso em uma interação com a plataforma ou outro texto curto semelhante, é limitada pela maneira como as emoções são verbalizadas e pela natureza contextual da comunicação das emoções. Além disso, enquanto ferramentas automatizadas podem ser eficazes no processamento de grandes volumes de dados, a falta de sofisticação e contexto de conhecimento dessas ferramentas, além dos vieses introduzidos pelos próprios pesquisadores, podem reduzir a capacidade dessas ferramentas de identificar com precisão a polaridade do sentimento ou o estado emocional. Como em qualquer outra ferramenta automatizada de análise de dados, os pesquisadores precisam avaliar cuidadosamente a adequação do software de análise de sentimentos para seus projetos e entender suas limitações.

Como limitadores da pesquisa pode-se admitir que aferir modelos analíticos de apenas uma instituição financeira pode reduzir a capacidade de descobertas no campo de análise de sentimento apenas ao consumidor bancário.

O estudo não propõe desestimular a análise automatizada do sentimento humano em posts de mídias sociais, mas sugerir que os pesquisadores se familiarizarem com os aspectos técnicos e pragmáticos da comunicação no meio social, e com as características e limitações dos modelos analíticos que eles podem usar para analisar a vastidão dos dados que circulam nas redes sociais.

Futuras pesquisas podem buscar soluções na área de comunicação social para desvendar a razão pela qual as concordâncias em categorias de cunho negativo, principal gerador de insights, não alcançam a mesma performance das categorias “Neutro” e “Positivo”.

6. Referências Bibliográficas

- AARTI, S., Anu, S., NILANJAN, D. (2015). Semantics and Agents Oriented Web Personalization: State of the Art. **International Journal of Service Science, Management, Engineering, and Technology (IJSSMET)**. V. 6, 4. ed.. IGI, Global, 2015, 35-49.
- AGUIAR, R. F.; PRATI, R. C. Incorporação de representação vetorial distribuída de palavras e parágrafos na classificação de sms spam. **ENIAC-Encontro Nacional de Inteligência Artificial e Computacional**. Natal, Brasil, 2015.
- ALAMMAR, J. Intuition & Use-Cases of Embeddings in NLP & beyond. In: **Q Com Plus - InfoQ**. 2019. Disponível em: <<https://www.infoq.com/presentations/nlp-word-embedding/>>. Acessado em 15/02/2021.
- AMARAL, M. B. F., NETO, M. T. R., OLIVEIRA, M. Experiências positivas e negativas no comportamento do consumidor: um estudo no setor bancário envolvendo custos de mudança, comprometimento e emoções. **Revista Ciências Administrativas [en linea]**, 16(2), 544-567. ISSN: 2010, 1414-0896.
- BASIT, T. Changing practice through policy: Trainee teachers and the National Numeracy Strategy. **Research Papers in Education**, 18 (1), 2003, pp. 61-74.
- BENEVENUTO, F.; RIBEIRO, F., ARAUJO, M. **Métodos para Análise de Sentimentos em Mídias Sociais**. Sociedade Brasileira de Computação. Capítulo 2. 2015. pp. 31-59.
- BERGH, B., LEE, M., QUILLIAM, E., HOVE, T. The Multidimensional Nature and Brand Impact of User-Generated Ad Parodies in Social Media. **International Journal of Advertising**. 30. 103-131. 2011. 10.2501/IJA-30-1-103-131.
- BRODY, S.; ELHADAD, N. An Unsupervised Aspect-Sentiment Model for Online Reviews. **Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL**, 2010, p. 804–812.
- BUSHMAN, B., BAUMEISTER, R., PHILLIPS, C. Do people aggress to improve their mood? Catharsis beliefs, affect regulation opportunity, and aggressive responding. **Journal of personality and social psychology**. 81. 17-32. 10.1037//0022-3514.81.1.17.
- CANHOTO, A. I.; PADMANABHAN, Y. ‘We (don’t) know how you feel’– a comparative study of automated vs. manual analysis of social media conversations. **Journal of Marketing Management**, v. 31, 2015, 1.141-1.147.
- CASTELLS, M. A era da informação – economia, sociedade e cultura. Vol 1. **A sociedade em rede**. São Paulo: Paz e Terra, 1999.
- CECCON, D. BERT, o modelo de processamento de linguagem natural que revolucionou a área. **IA Expert Academy**. Disponível em: <<https://iaexpert.academy/2020/04/27/>>, 2020.
- CHAPMAN, P. *et al.* CRISP-DM 1.0. In: **Step-by-step data mining guide**. [S.I.] : [S.E.], 2000.

CHEN P., KRAUSS, A. Interrater agreement. In M. LEWIS-BECK, A., BRYMAN, & T. LIAO. **The sage encyclopedia of social science research methods**. vol. 2, Thousand Oaks, California: Sage Publications, Inc. 2004, pp. 511-513.

CHEN, Lu et al. Extracting diverse sentiment expressions with target-dependent polarity from twitter. In: **International AAAI Conference on Web and Social Media**. 2012

COELHO, L. Machine Learning: O que é, Conceito e Definição. **Cetax**. Disponível em <<https://www.cetax.com.br/blog/machine-learning/>>. Acessado em 24/02/2021.

COHEN, J. B., PHAM, M. T., & Andrade, E. B. The nature and role of affect in consumer behavior. In C. P. Haugtvedt, P. Herr, & F. Kardes (Eds.), **Handbook of consumer psychology**, 2008, 297–348.

CRONBACH, L. J. Coefficient alpha and the internal structure of test. **Psychometrika**. 1951, pp. 297-334.

CUMMINS, R. A., GULLONE, E.. Why we should not use 5-point Likert scales: the case for subjective quality of life measurement. In. **Second International Conference on Quality of Life in Cities**, Singapore, 2000.

DALMORO, M., VIEIRA, K. M. Dilemas na construção de escalas Tipo Likert: o número de itens e a disposição influenciam nos resultados? Revista Gestão Organizacional. 2013. Disponível em: <https://doi.org/10.22277/rgo.v6i3.1386>.

DEVLIN, J., CHANG, M., LEE, K., TOUTANOVA, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Disponível em: arXiv:1810.04805v2. 2019.

DEVELLIS, R. **Scale Development**. Sage Publications, Inc. ISBN: 9781412980449. 2011.

FAYYAD, U., PIATETSKY-SHAPIO, G., SMYTH, P. From data mining to knowledge discovery in databases. In: **AI magazine**, v. 17, n. 3, 1996, pp. 37.

FLEISS, J. **Statistical methods for rates and proportions**. New York: John Wiley & Sons, 1981.

FONSECA, R., SILVA, P., SILVA, R. Acordo inter-juízes: O caso do coeficiente kappa. **Laboratório de Psicologia**. V.5, DO 10.14417/lp.759, 2007, 81-89.

FORGAS, J. P. Affective influences on partner choice: Role of mood in social decisions. **Journal of Personality and Social Psychology**, 61(5), 1991, 708–720. doi:10.1037/0022-3514.61.5.708.

FRANKENTHAL, R. Entenda a Escala de Likert e como aplicá-la em sua pesquisa. **MindMiners**. Disponível em <[http: https://mindminers.com/blog/entenda-o-que-e-escala-likert/](https://mindminers.com/blog/entenda-o-que-e-escala-likert/)> acessado em 27/02/2021. 2017.

GALLIANO, Alfredo Guilherme. **O método científico: teoria e prática**. São Paulo: Harbra, 1986.

GAO, W. *et al.* Joint learning on sentiment and emotion classification. In: **Proceedings of the 22nd ACM international conference on Conference on information & knowledge management**. Nova Iorque: ACM, 2013, 1505-1508.

GARCIA, F. S. Metodologia da Pesquisa Científica: Organização Estrutural e os Desafios para Redigir o Trabalho de Conclusão. **Revista Eletrônica “Diálogos Acadêmicos”** (ISSN: 0486-6266). 2015, pp. 72-87.

GRAHAM, S., *et al.* A Meta-Analysis of Writing Instruction for Students in the Elementary Grades. **Journal of Educational Psychology**. 104. 10.1037/a0029185. 2012.

GWET, K. L. **Handbook of inter-rater reliability**. 4. ed. Gaithersburg, MD: Advanced Analytics, 2014.

HAN, J., KAMBER, M., PEI, J. **Data mining: concepts and techniques**. Amsterdam: Elsevier, 2011.

HANEY, W.; RUSSELL, M.; BEBELL, D. Drawing on education: using drawings to document schooling and support changes. **Harvard Educational Review**, v. 74, n. 3, 2004, 241-271.

KIETZMANN, J. H., HERMKENS, K., MCCARTHY, I. P., & SILVESTRE, B. S. Social media? Get serious! Understanding the functional building blocks of social media. **Business Horizons**, 54 (3), 2011, 241–251. doi:10.1016/j.bushor.2011.01.005.

LAROS, F. STEENKAMP, J.. (2005). Emotions in consumer behavior: A hierarchical approach. **Journal of Business Research**. 2005. 1437-1445. 10.1016/j.jbusres.2003.09.013.

LANDIS, J. R.; GARY G. K. The Measurement of Observer Agreement for Categorical Data. **Biometrics**, vol. 33, no. 1, 1977, pp. 159–174. JSTOR, www.jstor.org/stable/2529310. Acessado em 24 Fev. 2021.

LEONTITSIS, A.; PAGGE, J. A simulation approach on Cronbach’s alpha statistical significance. **Mathematics and Computers in Simulation**, v. 73. 2007, p. 336-340.

LEUNG, K. C.; JOSEPH, K. W. Sports Data Mining: Predicting Results for the College Football Games. **Procedia Computer Science**. Volume 35, 2014, p. 710-719.

LIDDY, E. D., et al. Natural Language Processing. **Encyclopedia of Library and Information Science**, 2003, pages 2126-2136.

LIU, B. Sentiment Analysis and Subjectivity. In: **Handbook of natural language processing**, v. 2. [S.I.] : Chapman & Hall/CRC. 2010, p. 627-666.

LOBATO, FÁBIO, et al. Social CRM: Biggest Challenges to Make it Work in the Real World. **International Conference on Business Information Systems**. Springer, Cham, 2016.

MIKOLOV, T., CORRADO, G., CHEN, K., DEAN, J. Efficient Estimation of Word Representations in Vector Space. Arxiv: 1301.3781 v.3 [cs.CL]. 2013.

MILGRAM, Stanley. **Obedience to authority**. New York : Harper & Row, 1974.

MONTARDO, S.; CARVALHO, C.. Reputação: Monitoramento e Métricas. In: SILVA, Tarcízio (Org.). **Para entender o monitoramento de mídias sociais**. Disponível em: <<http://pt.scribd.com/doc/78400708/Para-Entender-o-Monitoramento-das-Midias-Sociais>>. [S.I.] : [S. E.], 2012, pp. 19-33.

NETO. P. S. R. **Moral e natureza humana em David Hume**. Universidade Federal da Bahia, Faculdade de Filosofia e Ciências Humanas. Salvador. 2019.

PANG, B.; LEE, L. Opinion mining and sentiment analysis. In: **Foundations and trends in information retrieval**, v. 2, n. 1-2, 2008, pp. 1-135.

PASSMORE, D. **Social Network Analysis – Theory and Applications**. Online, 2011.

PLUTCHIK, R. The Nature of Emotions. In: American Scientist Magazine, vol. 89, number 4. **Research Triangle Park**: Sigma Xi. 2001, p. 344-350.

REDHU, S., SRIVASTAVA, S., BANSAL, B., GUPTA, G. Sentiment Analysis Using Text Mining: A Review. **International Journal on Data Science and Technology**. Vol. 4, No. 2. 2018, p. 49-53.

ROSSI, R. H. P. S. **Análise de sentimentos para o auxílio na gestão das cidades inteligentes**. 2019. Tese (Doutorado em Sistemas Digitais) - Escola Politécnica, Universidade de São Paulo, São Paulo, 2019. doi:10.11606/T.3.2019.tde-16092019-110422. Acesso em: 2021-03-15.

SALUSTIANO, S. Análise de sentimento. In: SILVA, T.; STABILE, M. (Orgs.). **Monitoramento e pesquisa em mídias sociais: metodologias, aplicações e inovações**. São Paulo: Uva Limão, 2016. p. 29-52.

SHARMA, A.; DEY, S. Mining marketing intelligence from online reviews using sentiment analysis. In: **International Journal of Intercultural Information Management**, v. 5, n. 1-2, 2015, pp. 57-82.

SILVA, N. F. F. Análise de sentimentos em textos curtos provenientes de redes sociais. São Carlos, 2016, p. 5.

SIRQUEIRA, D. R. Uma arquitetura de pré-processamento para análise de sentimento em mídias sociais em Português brasileiro. Belém, 2018, p. 9.

STREINER, D. L. Being inconsistent about consistency: when coefficient alpha does and doesn't matter. **Journal of Personality Assessment**. v. 80. 2003, p. 217-222.

SUBASIC, P., HUETTNER, A. Affect Analysis of Text Using Fuzzy Semantic Typing. In: **IEEE transactions on fuzzy systems**, Vol. 9, No. 4. 2001.

TAMIS M., JONES, D. F. Algorithmic Improvements to the Method of Martel and Aouni. **Journal of The Operational Research Society**. v. 46, 1994, pp. 254-257.

TITOV, I.; MCDONALD, R. Modeling online reviews with multi-grain topic models. In: **Proceedings of the 17th international conference on World Wide Web**. Nova Iorque : ACM, 2008, pp. 111-120.

TUMELEIRO, N. Metodologia de pesquisa: guia rápido de possibilidades. In: **Mettzer**. Disponível em: < <https://blog.mettzer.com/metodologia-de-pesquisa/>>. 2019.

WIEBE, J. et al. Learning Subjective Language. **Journal Computational Linguistics**. Vol. 30 Issue 3. 2004. pp. 277-308.

WILSON, T. *et al.* Opinion Finder: A system for subjectivity analysis. In: Proceedings of hlt/emnlp on interactive demonstrations. **Stroudsburg: Association for Computational Linguistics**. 2005, p. 34, 35.

WITTEN, I. H. et al. **Data Mining: Practical machine learning tools and techniques**. Morgan Kaufmann, 2016.